JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

EXPLAINABLE AI-DRIVEN INTRUSION DETECTION SYSTEM FOR REAL-TIME CYBER THREAT MITIGATION IN DYNAMIC NETWORK ENVIRONMENTS

[1]Dr. Madhavi Dave, [2]Dr. Harshal Arolkar [3]Dhruv Panchal

[1]Project Manager, [2]Head, [3]AI/ML Team Lead
[1]IoT Security, [2] FCAIT, [3]Industry
[1] DRDO Industry Academia SVP CoE Gujarat University, [2]GLS University, [3]Data Vidwan
[1]Ahmedabad, Gujarat, India [2] Ahmedabad, Gujarat, India [3] Ahmedabad, Gujarat, India

Abstract
Cyber-attacks are becoming more complex, frequent, and sophisticated every day, so intrusion detection systems are now an extremely important part of any organization's cybersecurity strategy. The use of Artificial Intelligence and Machine Learning have shown to enhance the capacity to detect network intrusions. The users though are not able to be interpret or explain many Machine Learning models, thus limiting their usefulness in terms of practical applications for detecting and/or preventing cyber threats. The authors of this paper propose an explainable AI-based intrusion detection system that provides real-time mitigation against cyber threats in a dynamic network environment through the use of both Machine Learning to classify intrusions and SHAP (SHapley Additive exPlanations) to explain model decision making processes. The authors have also established and explained how their suggested framework works by developing two separate implementation levels: a baseline sample-level implementation and a large-scale implementation that employs a multi-class approach for classifying intrusions. The framework also includes the capacity to describe the contribution of certain network flow properties in identifying various attack types (e.g., brute force, Port Scanning, Denial of Service). This can provide further insight into attacks and enhance trust in the system. It combines scalability of Intrusion Detection Systems (IDS) with a detailed understanding (feature-level interpretations) of how features relate to various attack types to create a solution that has both performance and transparency within a cybersecurity system. The results indicate that utilizing Explainable Artificial Intelligence (XAI) with machine-learning-based IDS produces an effective, interpretable and scalable solution for real-time cyber defense within rapidly changing, continuous network environments.

Keywords: Intrusion Detection System, Explainable AI, Multi-Class Classification, Network Traffic Analysis, SHAP, Cybersecurity.

1. INTRODUCTION

Organizations face tremendous difficulty in maintaining continuity of their networks, as well as protecting their network resources and data from unauthorized access or loss. With exponential growth of digital technologies and increasing reliance of organizations on connected systems, the scope of an organization's vulnerability has increased. The traditional intrusion detection systems (IDS) that only use pre-existing rules or signatures to provide detection thus does not fit well for new or evolving threats occurring in rapidly evolving environments. This leads to an immediate need for an innovative and adaptive IDS that is be able to respond in real-time to the complex disparate patterns of attack that occur.

Artificial Intelligence and Machine Learning have been introduced into the field of intrusion detection, leading to improved systems as compared to legacy systems. In terms of the accuracy of anomaly detection, AI-based systems outperform traditional systems. Overall, there are many different AI-based tools and technologies available today that have been proven to provide enhanced detection and prediction capabilities as well as automation of security events. Finally, AI also becomes an important component when integrating with new technologies like IoT and 6G networks in order to provide smarter and more scalable security architectures. Despite the advances made to date, there are still many solutions out there that are considered "black boxes" as they do not explain how they determine a decision and do not build a relationship of trust between end-users of those tools or technologies and the "black box" security expert who provided them.

Explainable Artificial Intelligence (XAI) has emerged as a powerful approach to address the limitations of black-box models by providing transparency and interpretability in Machine Learning systems. Techniques such as SHAP level insights into model decisions, enabling analysts to understand how different input features influence predictions [1], [7]. Several studies have explored the application of XAI in cybersecurity domains including insider threat detection and anomaly detection, demonstrating improved interpretability and decision support [7], [18]. However, most existing approaches focus on limited

datasets or specific attack types, lacking comprehensive evaluation across multiple attack categories and realistic network traffic conditions.

This work represents a new way to create a complete Explainable AI-based intrusion detection system that will use both a baseline and a large-scale multi-class analysis approach. The architecture uses a two-level implementation strategy to first establish an explainable foundation at the individual sample-level with an initial model and then to build a complete multi-class intrusion detection system on top of that using the aggregated network traffic. In addition, class-wise SHAP analysis was used to identify the distinct contributions of each feature to the overall outputs of all classes of attacks in order to provide a better understanding of how the network behaves as a whole and to increase the transparency of decisions. Overall, this integrated approach to developing explainable AI-based models allows for increased scalability and interpretability, and thus supports the ability to mitigate cyber threats in real-time by providing a robust solution to support dynamic networks.

2. Literature Review

Cybersecurity experts are looking at Artificial Intelligence technology to help detect complex patterns of attack and automate the analysis of security events so it can improve the decision-making process associated with defense. There is currently a trend toward the use of Machine Learning and deep learning methods for threat detection, anomaly detection, and intelligent responses to security incidents within modern digital ecosystems [3],[4],[6]. There have also been special discussions about the use of Artificial Intelligence in cybersecurity; these discussions have helped identify many potentials uses of these technologies in protecting critical infrastructure, cyber-physical systems, and smart environments, all of which require dynamically scalable security mechanisms to address the demands of the environments in which they operate [2],[14],[16].

The second key requirement for Explainability is the need for a security analyst to comprehend why the system identified a certain event as malicious. The studies conducted on XAI in the realm of cybersecurity have confirmed that while black box models can make very accurate predictions, they suffer from lack of trust, accountability, and deployability in real-world operations [1]. Research into insider threat detection and adaptive security has demonstrated that XAI can provide meaningful feature-level explanations to the security analyst enhancing the analyst's confidence when making decisions [7], [20]. Therefore, the results also validate the importance of Explainability to operational cybersecurity systems in practice as well as in theory [1],[7].

The field of intrusion detection is considered to be the leading field for the use of Artificial Intelligence (AI) within cybersecurity. Numerous works have been published that focus on developing predictive models for identifying threats to computer networks, defense systems employing AI against attacks using the Distributed Denial of Service (DDoS) strategy and prevention systems using AI technology within cyber-physical environments as well as industrial environments [8], [9] and [14]. Additionally, a number of publications provide surveys of the primary techniques, datasets and evaluation methods associated with the development of AI-based Intrusion Detection Systems (IDSs) which indicate that this research is moving towards more data-driven and adaptive approaches to implementing IDS solutions [13] and [22]. Further research examining the Internet of Things (IoT), sixth generation (6G) technologies and autonomous environments indicates that the development of future Intrusion Detection Frameworks will need to provide support for disparate traffic patterns and continually changing attack surfaces [5], [11] and [17].

The analysis of intrusions based on class and interpretable security methods is gaining a significant amount of activity. An investigation into the use of explainable AI in detecting DoS attacks has shown that incorporating AI methods and tools with explanation capabilities increases our ability to comprehend the behavior of the attacks, as well as features contributing toward the existence of the attacks [18]. There are numerous findings regarding intelligent cognitive intrusion detection systems (ICS) in the industrial 4.0 setting indicating that intrusion detection systems are evolving toward more context aware and AI powered platforms [19]. There are also studies focusing on proactive cyber defense and AI based threat detection systems, and they have emphasized the requirement to move beyond simple detection of threats toward actionable security intelligence and transparent security systems [15, 21, 23].

Table 1: Literature Review Summary

Author(s) with Citation	Focus Area	Key Contribution	Limitation
Srivastava et al. [1]	XAI in cybersecurity	Comprehensive study on explainable AI concepts, challenges, and future scope	No practical IDS implementation
Mylrea et al. [2]	AI in critical infrastructure security	AI-driven cybersecurity for industrial systems and digital twin environments	Limited focus on IDS models
Sarker et al. [3]	AI-driven cybersecurity	Overview of AI-based threat detection and intelligent security systems	Lacks explainability perspective
Tabassum et al. [4]	AI-powered cybersecurity review	Systematic review of AI techniques in cybersecurity	General survey, lacks practical framework
Nguyen et al. [5]	6G security and AI	Security challenges and AI-based solutions for 6G networks	Conceptual, lacks experimental validation
Sarker [6]	Deep learning in cybersecurity	Detailed analysis of deep learning-based security models	Limited interpretability focus
Rahman & Shahdat [7]	Explainable threat detection	XAI-based behavioral analytics for insider threat detection	Limited to specific threat scenarios
Kishore et al. [8]	Network intrusion	AI-based predictive modeling for intrusion de-	Minimal explainability analy-

	detection	tection systems	sis
Alhindi [9]	DDoS attack mitigation	AI-based techniques for DDoS detection and prevention	Focus restricted to one attack type
La et al. [11]	IoT security in 6G networks	AI-based IDS, penetration testing, and blockchain-based trust management for IoT environments	Work-in-progress, lacks full experimental validation
Kalpani & Rodrigo [13]	IDS survey	Review of AI-based intrusion detection systems and trends	No explainable framework provided
Roohani et al. [14]	AI in cyber-physical systems	AI-enabled threat detection in cyber-physical environments	Limited interpretability discussion
Ali et al. [15]	Intelligent cyber defense	AI-based proactive threat detection and automated response systems	Lacks feature-level explainability
Razavi et al. [16]	AI-driven cybersecurity	Comprehensive insights into AI applications in modern cybersecurity	Broad scope, lacks implementation details
Ogenyi et al. [17]	IoT security	AI-based intrusion detection in autonomous IoT environments	Limited evaluation on large datasets
Ghosh et al. [18]	Explainable IDS (DoS)	SHAP-based explainable model for DoS attack classification	Limited to single attack category
Mahavaishnavi et al. [19]	Cognitive IDS (Industry 4.0)	Intelligent AI-based intrusion detection for industrial environments	Context-specific, limited generalization
Alsubaei [20]	XAI in IoT security	Explainable and adaptive AI security framework for IoT systems	Not generalized for all IDS scenarios
Nagpal et al. [21]	AI-based threat intelligence	Predictive cyber defense using AI-driven intelligence systems	Limited interpretability focus
Fatima et al. [22]	IDS techniques & datasets	Survey of IDS techniques, datasets, and evaluation strategies	Lacks feature-level explainability
Telrandhe et al. [23]	AI threat detection systems	AI-based automated threat detection and response systems	Limited scalability analysis

3. Methodology

In this paper, we have developed a detailed Explainability AI-based intrusion detection framework. It uses both a baseline and large-scale multi-class approach. A two-level implementation strategy has been used. Initially, an experiment has been performed at the sample level to develop a system based on aggregated network traffic data. A Class-wise SHAP analysis was performed to offer unique insight into each attack category's respective features and provide a deeper understanding of how the network operates, while also increasing the degree of interpretability. This framework will combine scalability, multi-class intrusion prevention and Explainable AI into one unified framework, making it appropriate for use against real-time cyber threats.

The explainable and scalable intrusion detection framework developed analyzes network traffic and classifies it into a wide range of attacks while maintaining the interpretability of the model's decisions. The architecture consists of a modular design with four components: data input abstraction, feature representation, multi-class detection using a supervised model, and explainable AI integration. Various representations of network traffic can be generated to capture the underlying patterns of cyber-attacks based on observed behavior. The modified supervised learning is applied at two different points to perform multi-class classification. After classification we utilize the explainability (SHAP) of features that contributed to the classification of each class. A dual-stage evaluation strategy was employed to validate the overall performance and scalability of the framework. This design also ensures that the intrusion detection framework provides timely and accurate results, while at the same time offers transparency in the form of actionable information and allows cybersecurity personnel to make effective decisions regarding future cybersecurity strategies.

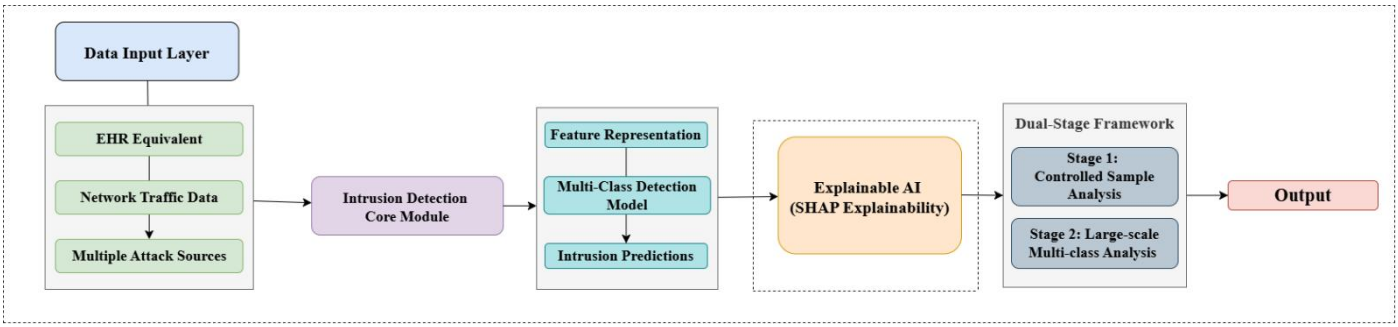

Fig. 1: Proposed Explainable AI-based Multi-Class Intrusion Detection Architecture with Dual-Stage Evaluation Framework

As can be seen in Fig 1, the architecture proposes a modular approach for multi-class intrusion detection, along with Explainable AI technologies. Network traffic from multiple origins is converted into a common representation, then processed by a multi-class intrusion detection model that is capable of identifying different types of attacks. An explainability model based on SHAP is then used to aid in interpreting the model predictions by examining individual features. A dual-stage

evaluation framework provides verification and validation of the model in addition to determining overall robustness when deployed at large scale. The outputs of the proposed architecture will be both intrusion classification results as well as interpretable information useful for making cybersecurity-related decisions.

Data Input and Unified Representation

The unified data representation layer allows collection of heterogeneous network traffic from different types of attacks and represent it in a consistent manner. The first layer of data input layer takes raw network flows and convert them to a standardized representation that retains the most critical statistical and behavioral attributes of all network traffic. Thus, using this abstraction enables the system to generalize many different types of attacks and also produce data that can be accepted by all processing modules downstream. The final step in the conversion of raw data into a structured format is of high importance because it allows for proper learning and interpretation.

Feature Representation and Pattern Encoding

The goal of feature representation is to represent the behavior of network traffic in a high-dimensional feature space that retains the discriminating features associated with the various types of attacks. This feature space includes flow-level characteristics such as packet statistics, behavior over time (temporal) and the directions in which the packets flow. This structured feature space enables the model to learn the complex relationships between different network features and the types of attacks. The distributed distribution of features will also support ease of interpretation and classification.

Multi-Class Intrusion Detection Model

The supervised learning model used in this study is the Random Forest classifier, selected due to its robustness in handling high-dimensional data, ability to manage class imbalance and strong performance in multi-class classification tasks. The model learns decision boundaries by analyzing the underlying feature representation of structured network traffic data, enabling effective separation of different attack classes. This detection approach accommodates diverse attack behaviors and overlapping patterns across multiple categories. Overall, the model demonstrates strong classification capability, making it suitable for real-world intrusion detection systems.

Explainable AI Integration using SHAP

SHAP-based feature attribution is the method used to provide explainability based on a quantitative evaluation of how much each feature contributes to a prediction made by the model. SHAP values can be computed to look at both global and local behavior; therefore, SHAP can provide an overall picture of which features are most impactful for the entire data set (global) or try to explain individual predictions (local). The explainability module allows for transparency in the way that classification results are affected by different attributes of the underlying network, creating confidence in the model's ability to classify accurately.

Dual-Stage Evaluation Framework

Two distinct evaluations will be done for the dual-stage evaluation strategy. Stage one includes analyses that are controlled with limited sample data. This evaluation will determine whether or not the proposed models and explainability pipelines are correct. The second stage will take the framework into an open multi-class environment at a much larger scale to evaluate robustness and real-world conditions. The consistency of behavior observed throughout both evaluation stages supports the dependability of the system and its ability to generalize results.

4. Result Analysis

This research utilizes the publicly available Network Intrusion Dataset, obtained from the Kaggle platform, which consists of network traffic flow records representing both benign and malicious activities. The dataset includes multiple attack categories such as Denial-of-Service (DoS), Port Scanning, brute-force attacks, web-based attacks and botnet traffic. It is inherently imbalanced, where certain classes such as benign traffic and DoS attacks dominate, while other attack categories contain comparatively fewer instances. The diversity of attack patterns and the large-scale nature of the dataset make it suitable for evaluating the effectiveness, robustness and scalability of Machine Learning-based intrusion detection systems in realistic network environments.

Dataset Link: LINK

A dual-stage implementation strategy was adopted to assess the effectiveness and scalability of the proposed framework. The initial stage involved a controlled experiment using a sample dataset extracted from a single file, namely Friday-WorkingHours-Afternoon-DDoS.pcap_ISCX.csv, where binary classification was performed to distinguish between benign and malicious traffic. This phase focused on validating baseline model performance and analyzing feature-level interpretability through SHAP.

The subsequent stage extended the evaluation by combining multiple dataset files into a unified large-scale dataset, enabling multi-class classification across various attack categories. Increased complexity arising from class imbalance and diverse attack patterns allowed simulation of realistic network conditions. The outcomes from both stages provide a comprehensive evaluation of the model's performance, demonstrating its capability to scale from controlled environments to real-world intrusion de-

tection scenarios while maintaining interpretability. The evaluation results demonstrate consistent performance across both binary and multi-class scenarios.

Table 2: Performance Evaluation of Random Forest Model

Implementation Stage	Algorithm Used	Accuracy	Precision	Recall	F1-Score
Sample Dataset (Binary)	Random Forest	0.99	0.99	0.99	0.99
Multi-Class Dataset	Random Forest	0.98	0.97	0.96	0.96

Sample Dataset Result Analysis

This study sets a starting point for evaluating how well the suggested Random Forest-based Identification System would operate. The experiment shows very good results with the ability to accurately separate benign network traffic from malicious network traffic. The dataset is simple and has lower variances, allowing the model to learn patterns and classify with almost perfect accuracy. Although classifying types of network traffic correctly demonstrates over-fitting, the accuracy level is an indicator that the model may have learned very few but very specific patterns, which it will never find in the future, as opposed to being general representations. Nevertheless, the controlled exploratory experimental setup creates a platform for assessing the effectiveness and interpretability of the model prior to expanding the assessment into more sophisticated multi-type testing.

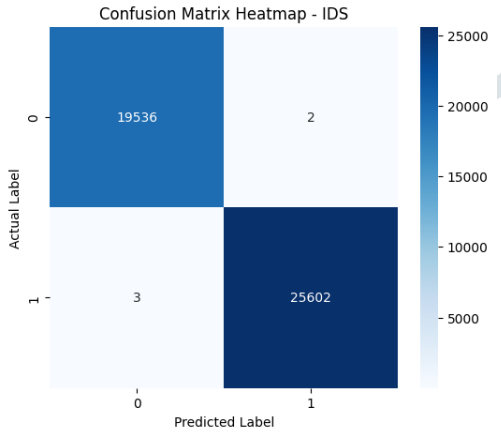

Fig. 2: Confusion Matrix for Sample Dataset

Fig. 2 shows that the model achieves near-perfect classification performance with minimal misclassifications between benign and malicious classes. A very small number of false negatives can be observed, indicating prediction reliability. The diagonal dominance in the matrix confirms that the model correctly classifies the majority of instances, such performance highlights the effectiveness of the model in a simplified environment, while also suggesting the possibility of slight over-fitting due to the limited complexity of the dataset.

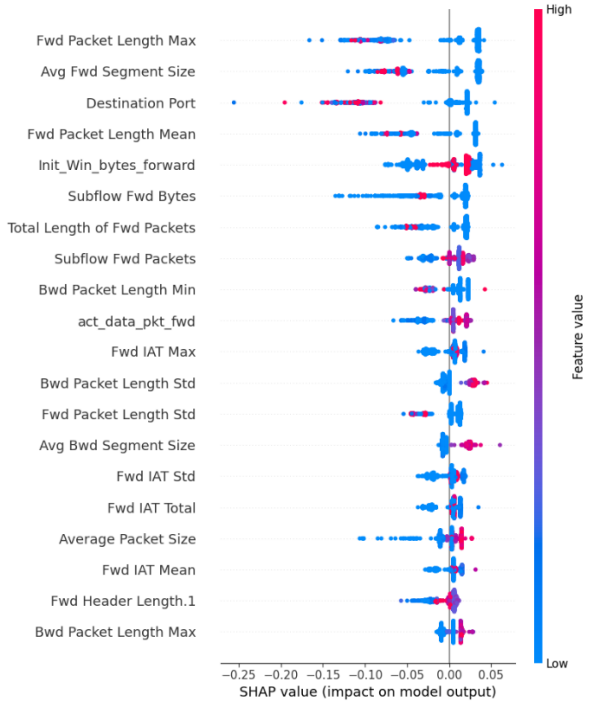

Fig. 3: SHAP Summary Plot for Feature Importance

Fig. 3 shows that the SHAP-based feature importance analysis provides clear insights into the contribution of individual features toward model predictions. The characteristics such as the length of the packet that is sent forward, how large the packet segment is when it is being sent forward on average, and which port the forward packet is going to are features that have a significant im-

pact on whether or not a particular activity has been detected as an intrusion. Feature Value Impact is represented by the Colour gradient of the plot, with colour red representing features with high values, and colour blue representing features with low values. High values (red) tend to skew predictions toward the attack class, while low values (blue) tend to skew predictions toward benign usage patterns. The range of SHAP values represent the degree to which features have an influence on the model's output. The high degree of clustering in the distribution of feature impacts lends further evidence to the conclusion that this model has acquired strong patterns during training and therefore results in a very accurate model; however, it also suggests some small degree of over-fitting by virtue of the fact that this was a controlled environment in which it was trained.

Multi-Class Dataset Result Analysis

The evaluation on the multi-class dataset demonstrates the robustness and scalability of the proposed Random Forest-based intrusion detection model across multiple attack categories. The model demonstrates strong overall performance, even though it was trained on a more complex data set with different numbers of instances of some classes (i.e., class imbalances). Minor differences in the model's performance for those minority classes indicate this model will exhibit realistic operational capabilities in real-world production environments with large data sets. This phase demonstrated the model's ability to generalize from a controlled environment and still provide highly interpretable outputs.

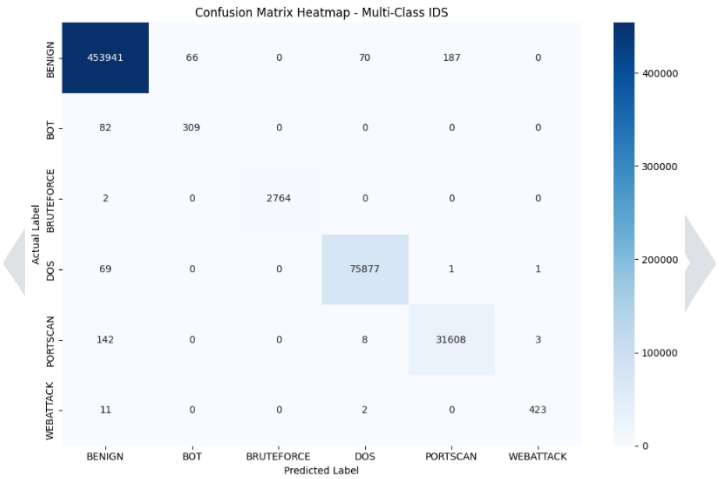

Fig. 4: Confusion Matrix for Multi-Class Intrusion Detection

Fig. 4. shows that the different classes of the model, predictions are predominantly found along the diagonal of the classification, indicating strong classification accuracy (i.e., predicting the correct class). While some minor misclassifications exist in classes with fewer samples (e.g., BOT and WEBATTACK), the majority of high frequency classes have virtually perfect classification accuracy. The distribution of error predictions shows evidence of a class imbalance between those classes. However, the model's reliability remains high for multi-class scenarios.

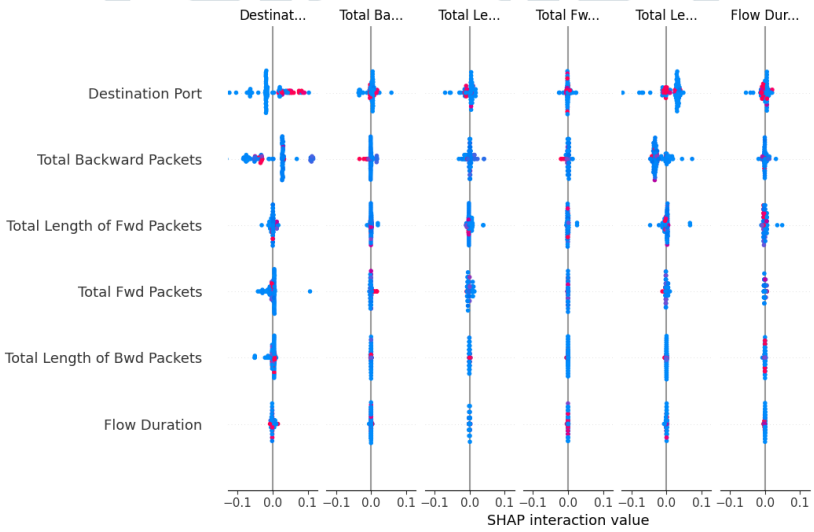

Fig. 5: SHAP Interaction Plot for Multi-Class Feature Analysis

Fig. 5 shows that the interaction of multiple features will have a combined effect on every model prediction when utilizing SHAP interaction analysis with multi-class settings. For example, the destination port, length of packets, and time of flow have significant interaction effects which differentiate between the class types. The color gradient represents the intensity of each feature value with red representing the highest level of feature value and blue the lowest. Differences in the SHAP interaction value show how interactions among the features in question provide evidence that complex relationships are being captured by the model within these datasets, thereby improving large-scale intrusion detection interpretability. The implementation at the sample level as opposed to the entire dataset is an illustration of the strength and generalizability of the proposed model, as well as its value in defining real-world intrusion detection scenarios.

5. Conclusion

An intrusion detection framework based on the Random Forest classifier that utilizes Explainable AI through SHAP algorithms to provide high-performance predictions with an easily understandable explanation for the predictions has been presented. The framework has undergone two stages of evaluation, using a small controlled data set followed by a larger multi-class data set. The evaluated models demonstrated excellent classification accuracy, but there were minor differences when considering the classification accuracy for the multi-class scenario, which reflects many of the challenges present in real-world intrusion detection, such as class imbalance and a vast array of possible attack patterns. The SHAP analysis contributed meaningful insight at the feature level, which can enhance the transparency of the decision-making process within the model and increase overall confidence in automated cybersecurity solutions. The primary contribution of this research is the dual-stage evaluation framework combined with Explainable AI, which have been developed to close the gap between having accurate models and interpreted models within an intrusion detection system. The work provides evidence that automated Machine Learning techniques demonstrate strong potential as reliable and explainable models for real-world cybersecurity applications.

6. References

1. Srivastava, Gautam, Rutvij H. Jhaveri, Sweta Bhattacharya, Sharnil Pandya, Praveen Kumar Reddy Maddikunta, Gokul Yenduri, Jon G. Hall, Mamoun Alazab, and Thippa Reddy Gadekallu. "XAI for cybersecurity: state of the art, challenges, open issues and future directions." arXiv preprint arXiv:2206.03585 (2022).

2. Mylrea, M., Nielsen, M., John, J., & Abbaszadeh, M. (2021). Digital twin industrial immune system: AI-driven cybersecurity for critical infrastructures. In Systems Engineering and Artificial Intelligence (pp. 197-212). Cham: Springer International Publishing.

3. Sarker, I.H., Furhad, M.H. and Nowrozy, R., 2021. Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. SN Computer Science, 2(3), p.173.

4. Tabassum I, Bazai SU, Zaland Z, Marjan S, Khan MZ, Ghafoor MI. Cyber Security's Silver Bullet-A Systematic Literature Review of AI-Powered Security. In2022 3rd International Informatics and Software Engineering Conference (IISEC) 2022 Dec 15 (pp. 1-7). IEEE.

5. Nguyen, Van-Linh, et al. "Security and privacy for 6G: A survey on prospective technologies and challenges." IEEE Communications Surveys & Tutorials 23.4 (2021): 2384-2428.

6. Sarker, Iqbal H. "Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective." SN Computer Science 2, no. 3 (2021): 154.

7. Md Whahidur, Rahman, and Hossain Md Shahdat. "An Explainable AI Framework for Insider Threat Detection Using Behavioral Business Analytics." An Explainable AI Framework for Insider Threat Detection Using Behavioral Business Analytics 1, no. 8 (2024): 70-97.

8. Kishore, M., PJ Sathish Kumar, and Srinath Doss. "Predictive Modelling for Network Threat Detection Using Artificial Intelligence Techniques." International Conference on Advances in Artificial Intelligence and Machine Learning in Big Data Processinging. Cham: Springer Nature Switzerland, 2024.

9. Alhindi, A., 2024, March. Exploring Artificial Intelligence's Potential in Developing Advanced Distributed Denial of Service Defense Strategies. In International Conference on Computing and Machine Learning (pp. 251-264). Singapore: Springer Nature Singapore.

10. Talib, Saja A. "Machine Learning to Enhance Security Systems." Machine Learning 9 (2024): 1-2024.

11. La, Vinh Hoa, Wissam Mallouli, Manh Dung Nguyen, Edgardo Montes de Oca, Ana Cavalli, Péter Vörös, Károly Kecskeméti et al. "Enhancing IoT security in 6G networks: AI-Based intrusion detection, penetration testing, and Blockchain-based trust management (work-in-progress paper)." In IFIP International Internet of Things Conference, pp. 53-67. Cham: Springer Nature Switzerland, 2024.

12. Sarcea, O.A., 2024, July. AI & Cybersecurity–connection, impacts, way ahead. In International Conference on Machine Intelligence & Security for Smart Cities (TRUST) Proceedings (Vol. 1, pp. 17-26).

13. Kalpani, Nethma, and Nureka Rodrigo. "Securing industry 4.0: a systematic review of AI-driven intrusion detection approaches and emerging trends." Journal of Reliable Intelligent Environments 12, no. 1 (2026): 1.

14. Roohani, Basudeo Singh, Ramesh Kumar Verma, Nripendra Dwivedi, Prabhat Kumar Srivastava, and Aditi Sharma. "AI-Enabled Threat Detection and Prevention in Cyber Physical Systems." In AI and Cyber Security in Cyber-Physical Systems, pp. 81-109. Cham: Springer Nature Switzerland, 2026.

15. Ali, Bisma, Syed Imad Shah, Laiba Sajid, Mir Rahib Hussain Talpur, Muhammad Umar Javed, and Muhammad Umair Warsi. "Design of Intelligent Cyber Defense Frameworks Using Artificial Intelligence for Proactive Threat Detection, Prediction, and Automated Response." Global Research Journal of Natural Science and Technology (2026).

16. Razavi H, Ouaissa M, Ouaissa M, Nakouri H, Abdelgawad A, editors. AI-driven Cybersecurity: Revolutionizing Threat Detection and Defence Systems. CRC Press; 2025 Sep 26.

17. Ogenyi, Fabian Chukwudi, Chinyere Nneoma Ugwu, and Okechukwu Paul-Chima Ugwu. "Securing the future: AI-driven cybersecurity in the age of autonomous IoT." Frontiers in the Internet of Things 4 (2025): 1658273.

18. Ghosh, Sneha, Raman Kumar Goyal, and Kuntal Chowdhury. "Explainable AI-Driven Intrusion Detection System for DoS Attack Classification Using Deep Learning and Optimization Techniques." IEEE Access 14 (2026): 5618-5642.

19. Mahavaishnavi, V., Saminathan, R., & Ramachandran, G. (2026). Intelligent Cognitive Cyber-Physical System–Based Intrusion Detection for AI-Enabled Security in Industry 4.0. Securing Cyber-Physical Systems: Fundamentals, Applications and Challenges, 45-64.

20. Alsubaei, Faisal S. "Luminous defense: XAI-driven adaptive security for critical IoT infrastructures." Journal of Cloud Computing (2026).

21. Nagpal, M., Singh, R.P., Shubneet, Yadav, A.R., Siwach, P. and Talwandi, N.S., 2025, November. AI-Driven Threat Intelligence for Predictive Cyber Defense in Smart Cities. In International Conference on Optimization and Data Science in Industrial Engineering (pp. 362-372). Cham: Springer Nature Switzerland.

22. Fatima, Ruksar, et al. "AI-Driven Intrusion Detection System: A Survey of Techniques, Datasets, and Evaluation Frameworks." Journal of Systems Engineering and Electronics (ISSN NO: 1671-1793) 35.12 (2025).

23. Telrandhe, A.V., Nishane, D., Puri, C. and Gayaki, U., 2025, October. AI-Powered Threat Detection and Response System for Next-Gen Cyber Defense. In 2025 2nd International Conference on Electronic Circuits and Signaling Technologies (ICECST) (pp. 1132-1137). IEEE.

24. Elatab SA, Almaktoof A. Artificial Intelligence in Cyber Security: A Comprehensive Overview. InInternational Conference on AI: Current Research, Industry Trends, and Innovations 2025 Jul 9 (pp. 108-125). Cham: Springer Nature Switzerland.

25. Mahto, M.K., 2026. Dynamic Threat Intelligence: Leveraging Generative AI for Real-Time Security Response. Generative Artificial Intelligence for Next-Generation Security Paradigms, pp.107-136.