ISSN: 2633-4828 Vol. 5 No.4, DECEMBER, 2023

International Journal of Applied Engineering & Technology

A NOVEL INTRUSION DETECTION SYSTEM UTILIZING VIRTUALIZED HONEYPOTS AND MACHINE LEARNING TECHNIQUES

Nidhi Bivalkar[1], Agneya Kolhatkar[2], Ameya Kolhatkar[3], Preeti Jain[4]

[1,2,3,4] Department of Computer Engineering, Pune Institute of Computer Technology, Savitribai Phule Pune University, Pune, Maharashtra, India

Email: [1] 5320nidhi@gmail.com, [2] kolhatkaragneya@gmail.com, [3] ameyajay@gmail.com, [4] pajain@pict.edu

ABSTRACT

The increasing complexity and frequency of cyber threats highlight the critical need for robust Intrusion Detection Systems (IDS). This study presents a new framework that combines Virtualized Honeypots and Machine Learning Techniques to enhance intrusion detection capabilities. By utilizing Suricata as the IDS and Cowrie for virtualized honeypot deployment, along with standard machine learning algorithms, the proposed system provides a comprehensive defence against various cyber threats. Suricata's effective detection engine and Cowrie's emulation of vulnerable services allow for real-time monitoring and analysis of network traffic and attacker behaviours, facilitating prompt threat detection. Machine learning algorithms further enhance the system's ability to identify known and unknown threats, including zero-day attacks. Additionally, a user-friendly web dashboard has been created to offer administrators an intuitive interface for viewing real-time alerts, intrusion events, and network traffic. This dashboard acts as a central platform for monitoring and responding to potential security breaches, enabling organizations to implement timely and informed security measures. By integrating virtualised honeypots, machine learning techniques, and a user-centric dashboard, this research emphasizes the importance of proactive defence strategies in combatting cyber threats and protecting critical assets.

Index Terms – Intrusion Detection System, Honeypots, Machine Learning, Suricata, Cowrie, Network Traffic Analysis

INTRODUCTION

In the face of continuous cyber threats and ever-changing attack methods, the necessity for robust and adaptable Intrusion Detection Systems (IDS) has become increasingly critical. Traditional IDS solutions, while somewhat effective, often struggle to keep up with the dynamic nature of modern cyber threats. With cybercriminals using more complex tactics like polymorphic malware and zero-day exploits, organizations are confronted with a constantly expanding threat landscape. To tackle this challenge, researchers and industry professionals are exploring novel strategies that utilize emerging technologies like Virtualized Honeypots and Machine Learning Techniques [1]. These advancements can potentially transform intrusion detection capabilities by providing a proactive and adaptive defence mechanism against a highly sophisticated adversary. By integrating the controlled environments of Virtualized Honeypots with the analytical capabilities of machine learning algorithms, organizations can gain a deeper understanding of attacker behaviour and detect subtle signs of compromise that may evade conventional detection methods. This combined approach enables organizations to proactively address emerging threats and promptly respond to potential security breaches, thereby bolstering their overall security posture in an era characterized by persistent cyber threats and evolving attack methods.

The primary objective of this research is to merge data obtained from Suricata, functioning as the primary IDS engine, and Cowrie, utilized for virtualized honeypots, within an Intrusion Detection System (IDS) framework [10], [11]. Through the strategic combination of these two data sources, a robust infrastructure is established to monitor network traffic and attacker behaviours in real time, thereby improving threat detection capabilities. The proposed system architecture, as depicted in Fig. 1, demonstrates the integration of Suricata and Cowrie to facilitate comprehensive threat monitoring. Both Suricata and Cowrie are capable of capturing a wide spectrum of network traffic, including both malicious and non-malicious activities. Each tool produces logs that are customized to their unique capabilities in capturing distinct aspects of network behaviour. Suricata is particularly proficient in identifying and logging intrusion attempts, whereas Cowrie focuses on simulating and logging engagements with possible attackers [10], [11]. This data is aggregated into a CSV dataset. After executing various data preprocessing techniques on the data, it becomes ready for threat identification with the help of ML algorithms. By utilizing both tools, we can gather a varied and intricate dataset, facilitating a comprehensive comprehension of network behaviour and bolstering our capacity to identify and counteract threats efficiently. Through the use of a variety of industry-standard machine learning algorithms, our study explores advanced techniques to enhance the IDS's ability to detect threats by identifying subtle patterns and anomalies associated with malicious activities. This integration of data from Suricata and Cowrie empowers the IDS to more accurately and efficiently identify both known and unknown threats. Furthermore, our goal is to create a sophisticated yet user-friendly web dashboard that acts as a central hub for administrators to visualize real-time alerts, intrusion events, and network traffic patterns. Through systematic experimentation, rigorous evaluation, and practical deployment scenarios, our research seeks to showcase the transformative impact of this integrated IDS framework. By equipping organizations with proactive defence mechanisms against evolving cyber threats, we aim to advance cybersecurity practices and fortify digital infrastructures against emerging security challenges. The following segments of this paper will explore an extensive review of the literature concerning progress in intrusion detection and honeypots, along with an explanation of our methodology for data collection. Subsequently, we will delve into the supervised classification algorithms utilized in our research and evaluate their effectiveness, supported by visual aids. To conclude, we will derive insights from our results, address the constraints of our study, and propose potential directions for future investigations.

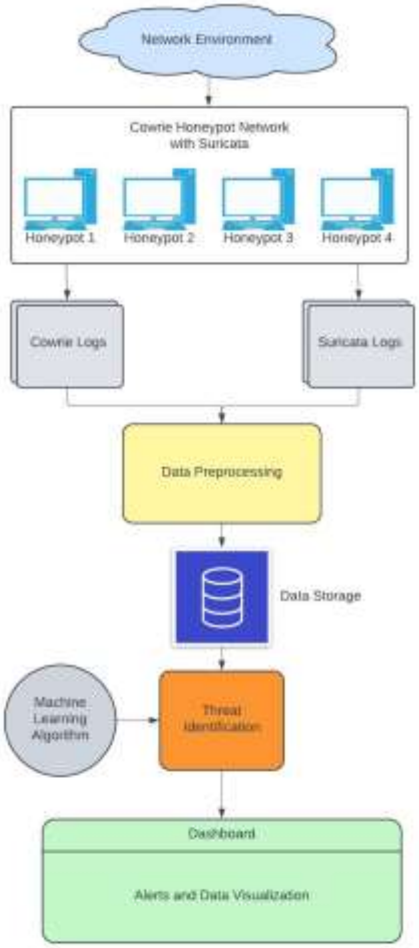

Fig. 1. Architecture of Proposed IDS

LITERATURE REVIEW

In the dynamic realm of cybersecurity, network intrusion detection has emerged as a crucial focus area for both researchers and practitioners. Given the continuously evolving nature of cyber threats in terms of complexity and frequency, it is essential to explore novel approaches and technologies to bolster network security. This literature review offers a thorough examination of recent research endeavours aimed at enhancing network security through a variety of methods and resources. It specifically explores the significance of honeypot-based network security frameworks, the incorporation of machine learning algorithms in intrusion detection, and the importance of utilizing tools such as Suricata and Cowrie to replicate vulnerable systems and identify network intrusions. By amalgamating findings from these studies, this review seeks to present useful techniques and perspectives for strengthening network security against emerging cyber threats.

Bivalkar et al. (2024) conducted a thorough survey that delved into the intricacies of honeypots and intrusion detection methods, emphasizing their importance and efficiency in modern cybersecurity strategies [1]. The study offered valuable insight into the realm of intrusion detection approaches and how honeypots contribute to enhancing network security, serving as a significant resource for gaining a deeper understanding of these concepts. The study by Sethi et al. provides an in-depth exploration of advancements in honeypot-based network security models, addressing the evolving landscape of cyber threats and the critical role honeypots play in detecting and deterring malicious activities [16]. Mehta et al. stress the importance of updating security measures for home networks in response to rising cyber threats, particularly during the COVID-19 crisis, suggesting the integration of Raspberry Pi 4 with IDS Suricata and honeypots to enhance protection [12]. Similarly, Hariawan et al. supports Raspberry Pi 4-based solutions enhanced with multiple honeypots and packet analyzers to strengthen home network security, conducting simulations to evaluate system performance under different attack scenarios [15]. Subhan et al. analyze patterns of brute force attacks on SSH ports using honeypots, providing insights into hacker techniques and proposing mitigation strategies based on identified attack patterns [14]. Hancock et al. introduce a simple method for detecting SSH and FTP brute force attacks using Decision Tree models, demonstrating effective classification performance on the CSE-CIC-IDS2018 Big Data dataset [13]. Additionally, Kathiresan et al. and Kiran et al. investigate the effectiveness of machine learning algorithms in intrusion detection, underscoring the importance of classification techniques like logistic regression, KNN, Decision Trees, and SVM in identifying network irregularities [17], [18]. Furthermore, Li and Dsouza et al. stress the significance of intrusion detection systems (IDS) in safeguarding computer networks, with Li focusing on theoretical analyses and practical implementations of intrusion detection technology, while Dsouza et al. propose a real-time network intrusion detection system utilizing machine learning for anomaly detection [19], [20]. Lastly, Liang introduces a cloud-based monitoring system for processing security event logs generated by IDS in real time, utilizing parallel algorithms for offline correlation analysis and real-time clustering of security logs to improve threat detection capabilities [21]. Together, these studies provide valuable insights and methodologies for enhancing network security in the midst of evolving cyber threats.

Various innovative strategies have been investigated in the domain of network intrusion detection to address the ever-changing landscape of cybersecurity threats. Husain et al. conducted a study on a robust intrusion detection model that utilized Extreme Gradient Boosting (XGBoost) on the UNSW-NB15 dataset, underscoring the importance of utilizing comprehensive datasets for evaluating detection systems effectively [5]. In a similar vein, Negi et al. recommended the use of Honeypots as a proactive approach to intrusion detection and prevention in cloud environments, demonstrating the effectiveness of such tactics [6]. Additionally, A. Halimaa et al. focused on the role of machine learning algorithms in intrusion detection systems, highlighting the significance of accuracy and suggesting methods like Support Vector Machine (SVM) and Naïve Bayes for improvement [7]. Moreover, a study on network intrusion detection using ensemble machine learning techniques introduced a novel approach through voting ensemble methods, outperforming traditional detection methodologies in terms of detection rates [8]. Furthermore, research on modelling and visualizing attack behaviours offered valuable insights into leveraging data analytics algorithms such as Self-Organizing Map (SOM) and Association Rule Mining (ARM) to differentiate between normal and malicious traffic patterns [9]. Together, these efforts constitute a thorough exploration of advanced techniques and methodologies essential for building resilient intrusion detection systems and proactive security measures against evolving cyber threats.

Modern research endeavours in cybersecurity heavily rely on the utilization of tools like Cowrie for simulating vulnerable systems and Suricata for detecting network intrusions. These tools play a vital role in gaining a thorough comprehension of cybersecurity threats and implementing effective mitigation strategies.

A. Suricata

Suricata is an open-source Intrusion Detection System (IDS), Intrusion Prevention System (IPS), and Network Security Monitoring (NSM) engine [10]. It was created by the Open Information Security Foundation (OISF) to oversee network traffic and pinpoint potentially harmful activities such as intrusion attempts, malware infiltration, and denial-of-service attacks. The system employs a combination of signature-based detection, protocol examination, and anomaly detection methods to recognize suspicious patterns and behaviours in network data. Suricata is distinguished by its scalability and ability to handle high-speed network traffic in real-time, making it a suitable choice for deployment in various network sizes, from small-scale setups to large corporate environments. Users can tailor the detection capabilities of Suricata to meet their specific security needs through the creation of custom rules and plugins. Moreover, the system's multi-threaded design and efficient packet processing ensure minimal impact on network performance [10]. In our study, we used the network data-collecting ability of Suricata to aid us in creating our dataset.

B. Cowrie

Cowrie is an open-source honeypot software that is utilized to replicate vulnerable systems and services to attract and ensnare malicious individuals attempting to gain unauthorized entry [11]. Programmed in Python, Cowrie creates a lifelike setting by emulating various services such as SSH and Telnet to entice and interact with intruders. It gathers a wealth of data on attacker actions, including executed commands, accessed files, and login trials, offering crucial insights into their strategies and procedures. The software's adaptable structure allows for effortless personalization and expansion, empowering users to integrate additional functionalities and plugins to boost its performance. By documenting and evaluating attacker conduct, Cowrie proves to be a valuable resource for us during our study. When operating in medium interaction mode (shell), Cowrie replicates a UNIX system using Python, while in high interaction mode (proxy), it acts as an SSH and telnet proxy to monitor attacker actions towards another system [11]. We used Cowrie in high interaction mode to acquire detailed logs of any intrusion attempts.

DATA COLLECTION AND METHODOLOGY

Creating a robust dataset is a crucial initial step in enhancing an Intrusion Detection System (IDS) through the integration of machine learning capabilities. This dataset plays a dual role: firstly, it enables the assessment and selection of the most appropriate machine learning algorithms for intrusion detection, and secondly, it serves as a repository of authentic data obtained from Virtual Machines experiencing exploits from penetration testing distributions like Kali Linux. Our research initially concentrated on assembling a varied dataset tailored to validate different machine-learning algorithms. After this, data collection activities were launched by deploying a set of computers equipped with Cowrie honeypot and Suricata IDS in a coordinated manner. This coordinated effort aimed to replicate genuine network settings and activities, capturing a broad range of network traffic and potential intrusion efforts for a thorough examination. Through meticulous dataset creation and collection of real-world data, our study aims to establish a solid groundwork for developing and assessing an integrated IDS framework enhanced with machine learning methods.

A. Synthetic Dataset Generation for Algorithm Testing

The research findings of Pravin et al. have significantly informed our decision-making process regarding the selection of machine learning algorithms and the design of the data collection environment [2]. By rigorously evaluating the 12 different machine learning algorithms in simulated attack scenarios, we gained valuable insights into their performance metrics such as accuracy, precision, recall, and F1 score. These insights were instrumental in guiding our choice of supervised learning models for intrusion detection, ensuring that the selected algorithms exhibit robustness and adaptability to dynamic cyber threats. Additionally, the work by Pravin et al. - a simulated OpenStack cloud environment, comprising a network of virtual machines deployed across two laptops, and usage of tools like Metasploit for orchestrating attacks and Netdata for real-time data collection, resulted in the generation of a system logs dataset [2]. Combining multiple studies alongside algorithm testing on a similar environment built by us we were able to successfully choose the optimal machine learning algorithms as XGBoost, AdaBoost, and Random Forest [2], [3], [4], [5], [8].

B. Real-world data Collection using Suricata and Cowrie

For comprehensive network data analysis, we coordinated the setup of a cluster of computers all running the Ubuntu operating system, version 22.04 [2], [6], [15]. This initiative aimed to create a strong infrastructure for deploying Cowrie honeypots and Suricata intrusion detection systems (IDS) throughout the network [6]. By connecting these computers, we established an extensive surveillance network capable of monitoring and analyzing traffic from multiple perspectives within the system. Each Ubuntu machine was carefully configured to accommodate both Cowrie and Suricata, enhancing the network's security measures and improving its ability to identify and address potential threats. Additionally, a dedicated Kali Linux virtual machine (VM) served as a centralized platform for orchestrating simulated cyber-attacks, executing meticulously crafted bash scripts that mimicked real-world threat scenarios. Seven kinds of exploits were performed - SSH Brute Force, SSH Remote Command Execution, Directory Traversal, Webshell, FTP_Exploit, HTTP_Exploit and HTTP_Slowloris [9], [12], [13], [14]. While not engaged in simulated attacks, the IDS and honeypots diligently captured and analyzed regular network traffic, including activities like ping messages, TCP requests, and non-malicious remote command executions. This continuous monitoring and data collection effort generated a comprehensive dataset covering both malicious and benign network behaviours. Data from Cowrie had to be first converted from JSON format to CSV format. By consolidating the data collected from Suricata IDS and Cowrie honeypots, we compiled a detailed aggregated dataset containing 100000 entries in total, out of which 21174 are malicious and 78826 are non-malicious. Fig. 2 gives an overview of the attack distribution. SSH_Brute_Force has the largest number of entries as it is one of the most common attacks, especially on an SSH application [9], [13], [14].

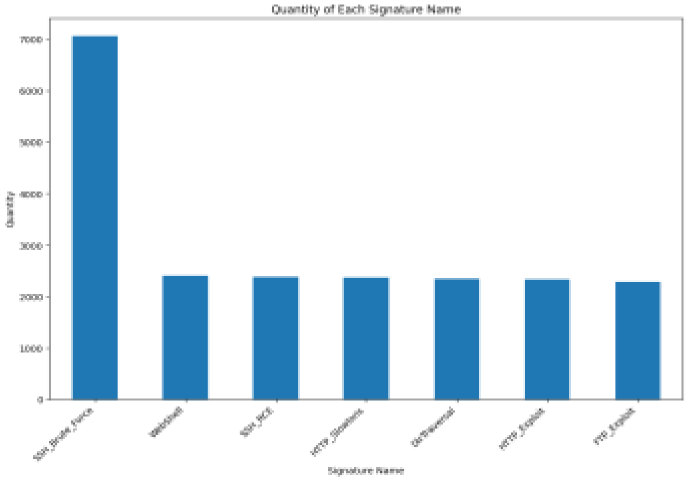

Fig.2. Attack Distribution

C. Dataset Features

The data collected from the above setup contains 58 essential characteristics crucial for in-depth network analysis. The features incorporated into the proposed intrusion detection system are outlined in Table 1, offering a systematic breakdown of the various information sources that form part of the analysis framework. Key features such as src_port and dst_port offer detailed insights into the source and destination ports of network traffic, while signature_name reveals identified signatures that may indicate suspicious activities. The protocol attribute specifies the communication protocol used, playing a fundamental role in understanding network operations [2]. Additionally, the commands attribute records executed commands, aiding in the identification of potentially harmful actions within the network [9], [12], [13]. In addition to these attributes, flow_bytes measures the amount of data transferred during network flows, enhancing the dataset with valuable traffic details. This meticulously gathered dataset serves as a robust tool for examining both legitimate and malicious network behaviours, enabling researchers to identify emerging cybersecurity threats and recognise patterns of normal activity are crucial for strengthening defence mechanisms against potential cyber-attacks.

Table 1. Features of the dataset

Category	**Features**
Basic Information	timestamp, src_ip, src_port, dst_ip, dst_port, protocol, packet_length, signature_id, signature_name, action, alert_level, alert_message, session
Authentication	username, password
Network Information	country, flow_id, flow_bytes, flow_packets, flow_duration, flow_start_time, flow_end_time, flow_direction, flow_state, flow_flag, flow_ttl, flow_window, flow_initiator, flow_responder
File information	directory, file_path, file_action, file_size, file_type, result, version
Packet Information	packet_data, packet_header, packet_payload, packet_protocol, packet_flags, packet_fragmentation, packet_icmp_type, packet_icmp_code, packet_icmp_checksum, packet_udp_length, packet_tcp_flag, packet_tcp_sequence, packet_tcp_acknowledgment, packet_tcp_window, packet_tcp_checksum, packet_tcp_urgent_pointer, packet_http_method, packet_http_uri, packet_http_response_code

D. Data Preprocessing

Initially, the dataset is loaded into a Pandas DataFrame to allow for efficient data manipulation and exploration. To ensure data completeness, missing values in the 'signature_name' column are filled with the value "Normal". Redundant features such as 'timestamp', 'src_ip', and 'dst_ip' are identified and subsequently removed to streamline the analysis process [18], [20]. Categorical variables undergo encoding using LabelEncoder to convert textual data into numerical forms, while OneHotEncoder further transforms categorical features into binary vectors to enhance their compatibility with machine learning algorithms. The preprocessed dataset is then divided into training and testing sets to facilitate model evaluation and validation. Moreover, real-time data is processed similarly, ensuring consistency and alignment with the trained model for predicting signature names related to network activities [20], [21]. This meticulous data preprocessing approach establishes a strong groundwork for accurate and reliable machine learning analysis, enabling researchers to extract valuable insights and predictions from intricate network datasets.

SUPERVISED CLASSIFICATION ALGORITHMS

Classification algorithms play a pivotal role in our IDS for the precise identification and categorization of network intrusions. Our focus is on three robust supervised classification algorithms: Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), and Random Forest Classifier. These algorithms are instrumental in enhancing the efficiency of IDS by facilitating prompt and accurate detection of anomalous network activities that could signify potential cyber hazards. Each algorithm boasts distinct strengths and functionalities, rendering them indispensable resources for our research endeavour [3], [4], [5].

A. Extreme Gradient Boosting (XGBoost)

The study by Husain et al. highlights XGBoost as a crucial component in network intrusion detection systems (NIDS) for effectively managing complex datasets such as UNSW-NB15 [5]. Its significance lies in its capacity to handle numerical features with precision and conduct optimal feature selection, allowing for the identification of key network attributes that differentiate various types of attacks. Through the use of gradient-boosting techniques within a decision tree ensemble, XGBoost stands out in extracting valuable insights from intricate network traffic data, enabling accurate predictions of network attack types [5]. Its performance surpasses that of traditional algorithms, as evidenced by its high accuracy levels and robust analysis of feature importance. Therefore, within the realm of IDS, XGBoost emerges as a powerful tool for bolstering network security by facilitating the swift and accurate detection of abnormal network behaviours indicative of potential cyber threats.

B. Adaptive Boosting (AdaBoost)

In the study by Sidharth V. et al. AdaBoost was highlighted for its ability to enhance the performance of a weak learner, specifically a Decision Tree algorithm, through iterative adjustments of misclassified data points' weights [3]. The findings of the research revealed that the model that utilized AdaBoost, in conjunction with XGBoost as another base classifier, demonstrated superior accuracy and detection capabilities when compared to the Stacked Generalization ensemble method. The AdaBoost algorithm significantly contributed to the overall improvement of the NIDS performance by effectively managing misclassified instances and leading to enhanced classification results [3].

C. Random Forest Classifier

Random Forest is chosen as one of the classifiers in this study due to its inherent advantages in handling high-dimensional data, dealing with large datasets, and robustness to noise and overfitting. Its ensemble learning approach, as described in the study by Kumar et al. combines multiple decision trees and aggregates their predictions, tends to yield more accurate and stable results compared to individual decision trees [4]. Additionally, Random Forest can handle both numerical and categorical data, making it suitable for intrusion detection tasks where diverse types of features are involved. Moreover, the ability of Random Forest to provide feature importance scores aids in understanding the significance of different features in the detection process, contributing to the interpretability of the model [4].

RESULTS

In this section, we present the results of our study on intrusion detection using supervised classification algorithms. Our main goal is to assess the effectiveness of the proposed intrusion detection system in accurately identifying and categorizing various types of network intrusions by examining performance metrics and model selection thoroughly. Additionally, we explore the performance of the XGBoost classifier and analyze the system's ability to conduct real-time analysis to highlight its capabilities and advantages. These results demonstrate the potential of our approach in strengthening cybersecurity measures and addressing emerging cyber risks.

A. Performance Metrics

Intrusion Detection relies on multiple features to correctly identify intrusion attempts. This makes it a multiclass classification problem. To evaluate our proposed system's efficiency, we utilise various performance metrics specifically designed for multiclass classification tasks [17]. These metrics encompass accuracy, precision, recall, F1-score, and an in-depth analysis of the confusion matrix. Accuracy assesses the overall correctness of the classifier's predictions across all classes, while precision and recall offer insights into the classifier's capability to correctly identify instances of a specific class and its ability to retrieve all relevant instances, respectively. The F1-score, a combined measure of precision and recall, provides a balanced evaluation of the classifier's performance. Furthermore, examining the confusion matrix enables us to pinpoint areas where the classifier may struggle, such as misclassifying closely related types of intrusions. By utilizing this comprehensive set of performance metrics, we seek to conduct a thorough assessment of our intrusion detection system's effectiveness in precisely detecting and categorizing various network intrusions, showcasing its contribution to bolstering cybersecurity defences [7], [17], [18].

B. Model Selection

Fig. 3 shows the F1 score and accuracy comparison of AdaBoost, XGBoost, and Random Forest. XGBoost showcases outstanding results with an accuracy score of 0.9087 and an F1 score of 0.9088. It slightly outperforms Random Forest and significantly outperforms AdaBoost in the context of our dataset. These scores have been obtained after averaging the scores of three training and testing attempts. Additionally, its resilience to overfitting, scalability, and adaptability in managing intricate datasets position it as a favoured option for diverse machine-learning assignments [5]. Consequently, owing to its exceptional performance metrics and versatility, XGBoost emerges as the prime classifier for the present intrusion detection task.

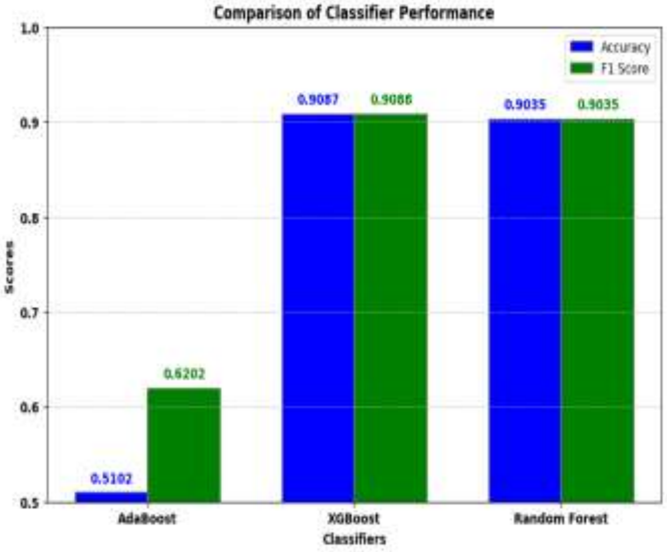

Fig. 3. Comparison of AdaBoost, XGBoost and Random Forest

C. Performance Analysis of XGBoost

The efficacy of XGBoost in intrusion detection is evident from the performance evaluation presented in Table 2. Achieving an accuracy rate of 0.9087, XGBoost showcases a significant level of overall correctness in categorizing instances across various intrusion types. Moreover, its precision score of 0.9095 signifies the model's capability to precisely recognize specific intrusion instances, thereby reducing false positives. The recall score of 0.9081 further emphasizes XGBoost's effectiveness in retrieving all relevant intrusion instances, indicating its comprehensive detection abilities. Additionally, the F1 score of 0.9088, amalgamating precision and recall reflects a well-balanced performance in terms of both accurately identifying relevant instances and minimizing misclassifications. In conclusion, these findings highlight XGBoost's strong performance in intrusion detection, positioning it as a promising tool for enhancing cybersecurity measures.

Table 3 shows the performance metrics when the XGBoost classifier is used on individual logs of Suricata and Cowrie, respectively. Upon examination of these individual metrics in comparison to those obtained from the combined dataset, it is evident that their performance metrics show a slight but objectively significant decrease when evaluated independently. Suricata, renowned for its strong intrusion detection features, showcases commendable accuracy, precision, recall, and F1 scores; however, these metrics exhibit a minor decline when contrasted with the combined dataset [10]. Similarly, Cowrie, which capitalizes on its honeypot capabilities, demonstrates respectable but lower metrics across various parameters when assessed in isolation [11]. As a result, the integration of insights gleaned from both Suricata and Cowrie logs remain essential for acquiring a holistic comprehension of network operations, thereby bolstering organizations' capacity to effectively identify and address cybersecurity threats.

Table 2. Performance Metrics of XGBoost Classifier on Aggregation of Suricata and Cowrie data

Metric	Score
Accuracy	0.9087
Precision	0.9095
Recall	0.9081
F1	0.9088

Table 3. Performance Metrics of XGBoost Classifier on Cowrie and Suricata data separately

Metric	Cowrie	Suricata
Accuracy	0.8813	0.9021
Precision	0.8782	0.9032
Recall	0.8795	0.9016
F1	0.8789	0.9024

D. Visualization using Plotly Dash

Plotly Dash is a powerful framework for building interactive web-based dashboards using Python [22]. With Plotly Dash, researchers can seamlessly integrate various data visualization components such as graphs, charts, and maps into a single dashboard interface. We leveraged the capabilities of Dash to develop a modern-looking dashboard which provides information regarding various network metrics and helps cybersecurity admins determine the correlation between various features, allowing them to improve the Machine Learning model through manual tuning. By utilizing its extensive library of interactive visualization tools, we could update and customize the content of its dashboards based on user interactions or real-time data feeds. Additionally, its integration with popular Python libraries like Pandas and NumPy enables researchers to easily manipulate and preprocess data before visualization [22]. Overall, Plotly Dash offers researchers a versatile platform to create interactive dashboards that facilitate data exploration, analysis, and communication in diverse research domains.

E. Real-time Intrusion Detection

The importance of analyzing network traffic in real-time for promptly detecting and addressing potential threats is crucial in the realm of modern intrusion detection systems (IDS) [20], [21]. This research focuses on the utilization of the XGBoost classifier within the IDS framework, which was chosen for its proven reliability and effectiveness in handling complex datasets. Incorporating real-time data streams from Suricata and Cowrie honeypots presents a distinct challenge that demands smooth processing and alignment with the training data format [20]. As the trained data employs one-hot encoding for feature representation, integrating real-time data is complicated by inconsistencies in the number of features. To address this obstacle, Python code has been developed to establish a mapping mechanism that ensures the accurate processing and alignment of incoming data streams. This mechanism is essential for synchronizing real-time data with the trained model, enabling the XGBoost classifier to provide accurate and timely predictions within the dynamic network security landscape. Furthermore, the IDS framework has been enriched with an integrated alerting system powered by Plotly Dash, which triggers alerts upon detecting suspicious activities, facilitating proactive response measures to safeguard the network infrastructure [22].

LIMITATIONS AND FUTURE SCOPE

A. Limitations of proposed IDS

Our research has shown promising outcomes concerning the integration of virtualized honeypots, machine learning methods, and real-time analysis for intrusion detection. However, it is essential to recognize certain constraints. One significant limitation is the potential for certain cyber-attacks to display similar feature values, making it challenging to classify them accurately and potentially leading to misclassifications or false alarms. This emphasizes the continuous need for refining and adjusting the intrusion detection system to tackle evolving threats and enhance its resilience. Additionally, the system's efficiency could be impacted by the quality and variety of the training data, as well as its scalability to more extensive and intricate network environments. Moreover, the use of machine learning algorithms introduces the issue of interpretability, where comprehending the models' decision-making process is crucial for ensuring trust and transparency in the system's functionality. These limitations point towards future research and advancement opportunities, such as exploring advanced feature engineering techniques, ensemble learning approaches, and interpretability tools to boost the overall effectiveness and dependability of the intrusion detection system.

B. Future research and development

The field of cybersecurity is constantly evolving, and the incorporation of virtualized honeypots and machine learning into intrusion detection systems (IDS) offers promising pathways for future research. Firstly, the development of machine learning algorithms specifically designed for intrusion detection has the potential to enhance the accuracy and efficiency of cyber threat detection, minimizing false alarms and enhancing system dependability. Secondly, research efforts focusing on dynamic feature engineering techniques optimized for real-time analysis could improve the agility and responsiveness of IDS frameworks, enabling them to quickly adapt to new threats. Thirdly, integrating external threat intelligence feeds could broaden the contextual understanding of network traffic, empowering IDS frameworks to proactively detect and address evolving threats. Furthermore, addressing scalability issues and improving performance through distributed processing methods could facilitate the implementation of IDS frameworks in large network environments. Additionally, ongoing enhancements to user-friendly dashboards could equip security administrators with intuitive tools for real-time monitoring and response, streamlining security operations and increasing overall situational awareness. Lastly, exploring adversarial machine learning defences tailored for intrusion detection systems could strengthen resilience against sophisticated evasion tactics employed by attackers, ultimately bolstering overall security defences.

CONCLUSION

In conclusion, our study presents a novel method for intrusion detection that combines virtualized honeypots and machine learning within an Intrusion Detection System (IDS) framework. Utilizing Suricata as the main IDS engine and Cowrie for virtualized honeypot emulation, along with strong machine learning algorithms, our system demonstrates promising abilities in detecting and categorizing threats in real time. Our research highlights the effectiveness of Extreme Gradient Boosting (XGBoost) in accurately identifying network intrusions, confirming its trustworthiness and flexibility in handling complex datasets. Additionally, the creation of a user-friendly web dashboard provides security administrators with an easy-to-use interface to visualize real-time alerts and network traffic patterns, enabling proactive response strategies to strengthen digital infrastructures. While acknowledging certain limitations, such as the difficulty of distinguishing attacks with similar feature values and concerns regarding interpretability and scalability, our investigation reveals potential areas for future research. By exploring advanced feature engineering methods, ensemble learning techniques, and scalability enhancements specifically designed for large network environments, we aim to push the boundaries of cybersecurity practices. Ultimately, the integrated IDS framework outlined in this study emphasizes the importance of proactive defence strategies in addressing evolving cyber threats and protecting critical assets in today's dynamic threat landscape.

REFERENCES

[1] Nidhi Bivalkar, Agneya Kolhatkar, Ameya Kolhatkar, Preeti Jain. (2024). Honeypots and Intrusion Detection: A Comprehensive Survey doi: 10.37896/HTL30.1/9971

[2] Patil, Pravin & Kale, Geetanjali & Bivalkar, Nidhi & Kolhatkar, Agneya. (2023). Comparative Analysis of Weighted Ensemble and Majority Voting Algorithms for Intrusion Detection in OpenStack Cloud Environments. International Journal of Advanced Computer Science and Applications. 14. 10.14569/IJACSA.2023.0141276.

[3] Sidharth, V. & C.R, Kavitha. (2021). Network Intrusion Detection System Using Stacking and Boosting Ensemble Methods. 357-363. 10.1109/ICIRCA51532.2021.9545022.

[4] M. R. Kumar and K. Malathi, "An Innovative Method in Improving the accuracy in Intrusion detection by comparing Random Forest over Support Vector Machine," 2022 International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, 2022, pp. 1-6, doi: 10.1109/ICBATS54253.2022.9759062.

[5] A. Husain, A. Salem, C. Jim and G. Dimitoglou, "Development of an Efficient Network Intrusion Detection Model Using Extreme Gradient Boosting (XGBoost) on the UNSW-NB15 Dataset," 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Ajman, United Arab Emirates, 2019, pp. 1-7, doi: 10.1109/ISSPIT47144.2019.9001867.

[6] P. S. Negi, A. Garg and R. Lal, "Intrusion Detection and Prevention using Honeypot Network for Cloud Security," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 129-132, doi: 10.1109/Confluence47617.2020.9057961.

[7] A. Halimaa A. and K. Sundarakantham, "Machine Learning Based Intrusion Detection System," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 916-920, doi: 10.1109/ICOEI.2019.8862784.

[8] M. Raihan-Al-Masud and H. A. Mustafa, "Network Intrusion Detection System Using Voting Ensemble Machine Learning," 2019 IEEE International Conference on Telecommunications and Photonics (ICTP), Dhaka, Bangladesh, 2019, pp. 1-4, doi: 10.1109/ICTP48844.2019.9041736.

[9] C. Yao, X. Luo and A. N. Zincir-Heywood, "Data analytics for modeling and visualizing attack behaviors: A case study on SSH brute force attacks," 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, USA, 2017, pp. 1-8, doi: 10.1109/SSCI.2017.8280913.

[10] https://docs.suricata.io/en/latest/

[11] https://cowrie.readthedocs.io/en/latest/README.html

[12] S. Mehta, D. Pawade, Y. Nayyar, I. Siddavatam, A. Tiwart and A. Dalvi, "Cowrie Honeypot Data Analysis and Predicting the Directory Traverser Pattern during the Attack," 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2021, pp. 1-4, doi: 10.1109/ICSES52305.2021.9633881.

[13] J. Hancock, T. M. Khoshgoftaar and J. L. Leevy, "Detecting SSH and FTP Brute Force Attacks in Big Data," 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Pasadena, CA, USA, 2021, pp. 760-765, doi: 10.1109/ICMLA52953.2021.00126.

[14] A. Subhan, Y. N. Kunang and I. Z. Yadi, "Analyzing the Attack Pattern of Brute Force Attack on SSH Port," 2023 International Conference on Information Technology and Computing (ICITCOM), Yogyakarta, Indonesia, 2023, pp. 67-72, doi: 10.1109/ICITCOM60176.2023.10441929.

[15] F. R. Hariawan and S. U. Sunaringtyas, "Design an Intrusion Detection System, Multiple Honeypot and Packet Analyzer Using Raspberry Pi 4 for Home Network," 2021 17th International Conference on Quality in Research (QIR): International Symposium on Electrical and Computer Engineering, Depok, Indonesia, 2021, pp. 43-48, doi: 10.1109/QIR54354.2021.9716189.

[16] T. Sethi and R. Mathew, "A Study on Advancement in Honeypot based Network Security Model," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 94-97, doi: 10.1109/ICICV50876.2021.9388412.

[17] V. Kathiresan, S. Karthik, P. Divya and D. P. Rajan, "A Comparative Study of Diverse Intrusion Detection Methods using Machine Learning Techniques," 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2022, pp. 1-6, doi: 10.1109/ICCCI54379.2022.9740744.

[18] A. Kiran, S. W. Prakash, B. A. Kumar, Likhitha, T. Sameeratmaja and U. S. S. R. Charan, "Intrusion Detection System Using Machine Learning," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-4, doi: 10.1109/ICCCI56745.2023.10128363.

[19] X. Li, "Research and Design of Network Intrusion Detection System," 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), Shenyang, China, 2022, pp. 1069-1072, doi: 10.1109/ICPECA53709.2022.9718920.

[20] A. Dsouza, V. Lanjewar, A. Mahakal and S. Khachane, "Real Time Network Intrusion Detection using Machine Learning Technique," 2022 IEEE Pune Section International Conference (PuneCon), Pune, India, 2022, pp. 1-5, doi: 10.1109/PuneCon55413.2022.10014863.

[21] F. Liang, "Design of a Real-time Cloud-based Monitoring System for Network Security Events," 2023 IEEE 6th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE), Shenyang, China, 2023, pp. 753-757, doi: 10.1109/AUTEEE60196.2023.10407638.

[22] https://dash.plotly.com/