Journals Pub

International Journal of
Telecommunications &
Emerging Technologies

ISSN: 2455-0345
Volume 10, Issue 1, 2024
DOI (Journal): 10.37628/IJTET

https://journalspub.com/journal/IJTET/

Research IJTET

Network Intrusion Detection Using Wireshark and Machine Learning

D. Chandravathi[1], P.V.S.S. Praneeth[2,*], Abida Sakina[3], Kotti Anjanaa[4], Botcha Priyanka[5]

Abstract

The growth of networked devices has highlighted the desire for advanced intrusion detection (IDS) tools to protect digital systems from evolving cyber threats. Traditional IDS systems are often difficult to adapt to the threat environment because they rely on predefined signature lists. This study presents a new approach that combines Wireshark, a widely used network packet analysis tool, with advanced machine learning for intrusion detection. Our system leverages Wireshark's data ingestion and analysis capabilities and algorithms such as gradient boosting, Naive–Bayes, and random forests, providing greater accuracy in detecting defects and potential intrusions in network traffic data throughput. It provides effective protection against a variety of cyber threats, including DDoS attacks, and complies with regulatory standard. This research represents a significant advance in cybersecurity reform, enabling organizations to mitigate threats in real-time and support collaborative defenses in a persistent digital environment. A system called an intrusion detection system (IDS) observes network traffic for malicious transactions and sends immediate alerts when it is observed. It is software that checks a network or system for malicious activities or policy violations. Each illegal activity or violation is often recorded either centrally using an SIEM system or notified to an administration. IDS monitors a network or system for malicious activity and protects a computer network from unauthorized access from users, including perhaps insiders. The intrusion detector learning task is to build a predictive model (i.e. a classifier) capable of distinguishing between "bad connections" (intrusion/attacks) and "good (normal) connections."

Keywords: DDos attack, IDS, intrusion detection, machine learning, malicious attacks, Naïve–Bayes, random forest

***Author for Correspondence**
P.V.S.S. Praneeth
E-mail: praneethprayaga1432@gmail.com

[1]Associate Professor, Department of Computer Science Engineering, Gayatri Vidya Parishad College, Srinivasa Nagar, Madhura Wada, Visakhapatnam, Andhra Pradesh, India
[2-5]Student, Department of Computer Science Engineering, Gayatri Vidya Parishad College, Gayatri Vidya Parishad College, Srinivasa Nagar, Madhura Wada, Visakhapatnam, Andhra Pradesh, India

Received Date: May 23, 2024
Accepted Date: June 11, 2024
Published Date: June 24, 2024

Citation: D. Chandravathi, P.V.S.S. Praneeth, Abida Sakina, Kotti Anjanaa, Botcha Priyan. Network Intrusion Detection Using Wireshark and Machine Learning, International Journal of Telecommunications and Emerging Technologies. 2024; 10(1): 23–31p.

INTRODUCTION

In today's interconnected digital environment, the rise of networked devices has transformed the way we communicate, work, and do business. But as the network continues to rise, so does the danger of cyber threats, from the spread of malware to sophisticated Denial of service (DDoS) attacks. Traditional Intrusion Detection Systems (IDS) have long been used to identify and mitigate these threats [1]. But these systems often struggle to keep up with the rapid evolution of cyber-attacks because they often rely on signatures or rules, limiting their ability to adapt to emerging threats. Therefore, there is an immediate need for newer methods that can efficiently detect and respond to threats that evolve over time. To solve this challenge, our project offers a new approach that combines Wireshark, a widely used network

packet analysis tool, with advanced machine learning algorithms. Wireshark's powerful capabilities to capture and analyze network communications data provide rich data to identify vulnerabilities and potential intrusions. By using machine learning classification models like Random-Forest, K-Nearest-Neighbors, Naïve–Bayes, Support vector machine and Gradient Boosting, our model is designed to improve the performance and efficiency of intrusion detection, enabling participants to proactively identify and mitigate cyber threats in real-time.

The motive of the system is to detect network interference without decrypting the contents of packets, thus improving confidentiality and integrity between sender and receiver. Here, packets are defined, and different attributes are considered for analysis. The machine is trained to utilize various information to distinguish bad or suspicious unencrypted data or plaintext from regular unencrypted packets. Any intrusion or anomaly in packet properties can be easily detected and alerted to promptly address interference. Therefore, data can be classified based on packet characteristics without compromising confidentiality and security. Broadcast intrusion detection is employed to identify vulnerabilities in the network during transmission from host to client or across the network. These systems operate by verifying the digital signatures and certificates of senders and recipients. Any breach is monitored and reported to network administrators, alerting them to potential threats. Wireshark is an open-source tool used to analyze packet data transmitted over a network, providing comprehensive details about captured packets. It is extensively utilized by network administrators to identify errors, investigate security issues, debug protocol implementations, and verify network applications. Wireshark can capture data packets in real-time transmission over the network and filter them according to various criteria. The system utilizes Wireshark to capture packets sent over the network, analyze their contents, organize them into datasets, extract features using machine learning, and classify packets into categories such as encrypted, malicious encrypted, unencrypted, and malicious unencrypted packets [2].

A Distributed-Denial-of-Service (DDoS) attack is a malicious attempt to disrupt a target server, service, or network by inundating it with multiple requests or traffic from various sources across the internet [3]. Unlike traditional denial-of-service attacks, which usually originate from a single source, DDoS attacks utilize multiple distributed sources, making them more challenging to mitigate. DoS attacks exploit vulnerabilities in network protocols or server configurations to generate huge volumes of traffic [4–6] These attacks look to prevent legal users from using the targeted service or website, thereby denying its service. The distributed nature of DDoS attacks complicates their prevention efforts, as the attacking traffic may originate from numerous compromised devices, including computers, servers, IoT devices, and even cloud services [7, 8].

This project is focused on several important goals:

1. Evaluate the effectiveness of multiple machine learning algorithms in detecting suspicious and potentially intrusive network traffic data extracted by Wireshark.
2. Explore the possibility of combining Wireshark with machine learning models to improve intrusion detection.
3. Examine the performance of our proposed system in real-world environments, including parameters such as capacity development, resource utilization, and threat adaptation.

LITERATURE SURVEY

The growth of connected systems and the emergence of IOT have changed the way we use technology, enabling easy communication and universal connectivity across multiple devices and platforms. But this interaction also creates an unprecedented cybersecurity challenge as criminals exploit vulnerabilities in the network to launch attacks and impact sensitive information. Traditional anomaly detection systems rely on signature-based detection, which allows organizations to respond to emerging threats and zero-day attacks, often struggling to keep up with evolving threats [9].

To solve these problems, scientists and practitioners are turning to machine learning techniques to

make discoveries and detect patterns indicative of potential intrusions. Concurrently, tools such as Wireshark have become indispensable resources for network administrators, providing detailed information about network data and facilitating real-time traffic monitoring.

In this front, this article presents a novel predictive search method leveraging a combination of Wireshark and machine learning algorithms. By harnessing Wireshark's robust data collection and analysis capabilities alongside advanced machine learning models, our research aims to enhance the accuracy, efficiency, and adaptability of intrusion detection [10]. Through comprehensive experiments and comparative analysis, we have identified the gradient boosting algorithm as particularly effective for detecting anomalies in network traffic data extracted by Wireshark [11].

This research contributes to the expanding body of knowledge in cybersecurity. Our method combines the advantages of Wireshark and machine learning, providing an effective means to develop intelligent cybersecurity systems capable of thwarting cyber-attacks in real-time. By disseminating our research findings, we aim to support further exploration and innovation in the cybersecurity domain, ultimately bolstering the resilience of digital defense systems against cyber threats.

METHODOLOGY

Machine Learning Models

Random Forest, Logistic Regression, Naïve–Bayes, Support vector machine, and Gradient boosting machine are the five algorithms implemented in this system.

Naive–Bayes is the classification algorithm that shows high accuracy when our data has more independent attributes, which is best for classifying categorical data.

Random forest has the advantage of greater versatility compared to other models. It overcomes overfitting and variance issues, thereby improving accuracy. In the prediction process, attributes are selected based on each set of features present in the dataset, and nodes are included using the optimal separation method. This splitting carries on until all attributes are split into nodes, resulting in multiple trees. The trees are then combined to create a random forest model.

Support vector machine (SVM) is a model that is used for classification of data based on prior and posterior information [12]. It selects important factors for a particular attribute and assigns them values corresponding to relevant attributes, creating boundaries to separate attributes in the selection process. Classification accuracy improves when a boundary separates several features. SVM is effective when features in the data are interrelated rather than independent, making it not linearly separable [13].

Logistic regression is an algorithm used primarily for binary classification. It estimates the probability of an event based on one or greater than one predictive variable. Because of its simplicity, understandability, and efficiency, Logistic regression is widely used in fields such as medicine, finance, and business. Despite the name, logistic regression is a classification algorithm, making it a crucial tool in the machine learning toolbox.

Gradient boosting machine (GBM) is a powerful model for classification tasks. It combines many weak learners, normally decision trees, to create robust predictive models [14]. GBM builds trees one after the other, with every new tree correcting errors from the previous one. This iterative process results in a more accurate and predictable model. GBM maintains data relationships, checks for missing values, and prevents overfitting through methods like regularization. It finds applications in predictive modeling, error detection, and optimization [15].

Existing System

IoT applications are used in many places to make work easier and save time. People need to manage and operate equipment to understand accurate information. Integration as a development

reality. All the ml models are described, and the accuracies are calculated. Intrusion detection can realize network and host-based intrusion detection. ML algorithms can be classified into supervised models and unsupervised models. The advantages and disadvantages of various types of machine learning are compared. It is necessary to perform a general reconnaissance to determine whether the network data is normal data or attack data. The motive of accessing the network is to protect the network from various types of interference. Preprocessing is the main step in the search process and must be completed before the input can be discovered. Preprocessing helps find missing values. The primary purpose is to increase accuracy and reduce time difficulty.

The author reviews various algorithms and classifies them into two parts using data. One for testing and the other one for training data [16]. The feature selection process is basically about selecting the desired features. Different features are selected by different selection algorithms. Benchmarking is used to understand the performance of an algorithm. Analyzing the effects often reduces the cost of FA. This will help in establishing good communication. Various algorithms have been learned. Machine learning helps solve many simple problems. Traditional algorithms are subject to many intrusions. Newer machine learning models are best for analyzing the impact. An examination fee is calculated for the entrance examination. Combinatorial machine learning such as RF, SVM, and KNN showed good results. The data can be better calculated based on actual data to be provided. Nowadays, cybersecurity is very important when it comes to sharing confidential information. Use different machine learning models to get higher accuracy. Wire Shark could be used to obtain network information that could be used to detect intrusions. Shows package details such as package sending time and received time. Improve visibility of access to select desired names. Feature selection is done to increase accuracy. The difference between an IDS and a firewall is that firewall only protects against attacks from outside the network. It does not protect the network from the internal network. IDS is a system that is supposed to inform users about intrusions. This information can then be used to analyze access, or the Wireshark tool can also be used to obtain real-time data packets from the network. The tools and techniques used by attackers vary. The main concern is false positives and negatives. RF is the most accurate. Today's communication is free, confidential and honest.

Proposed System

Data Packet Generation and Collection

Initially, HTTP, HTTPS, TLS, etc., generate various types of network packets, each containing different content types such as text, audio, video, images, and data links. Additionally, packets with weak content are designed for distribution purposes. Both encrypted and unencrypted bad packets are created to accurately simulate real-world scenarios. This comprehensive database comprises traditional, encrypted, and unencrypted packets, serving as a rich resource for training and testing our search engines.

Packet Capture using Wireshark

The generated packets are transmitted over the network, and their contents are captured using Wireshark [17]. Wireshark captures and analyzes packet-level information, enabling us to scrutinize the characteristics and attributes of each packet as it traverses from source to destination [5].

Preprocessing and Feature Selection

After capturing the packet data, we identify the various attributes and properties that are in the packets. We initially eliminate irrelevant features that have minimal impact on the packet distribution, thus optimizing the given data for more analysis [18].

Feature Selection using One-Hot Encoding

One-hot encoding is a largely used technique in machine learning for feature engineering, especially while dealing with categorical data. It is employed to convert categorical variables into an understandable format for machine learning algorithms, thereby enhancing predictive performance.

Machine Learning Implementation

Following data preprocessing, we use different machine learning models, including Random-Forest, Support-Vector-Machine, and K-Neighbors, to efficiently classify the intrusive packets. The classification achieved by each algorithm serves as an indicator of its effectiveness in classifying between normal packets, encrypted packets, and malicious packets [19].

Report Generation

If any intrusive packets are detected, a report is generated and displays the intrusive packets.

Architecture Of Proposed System

Block Diagram of proposed system is shown in Figure 1.

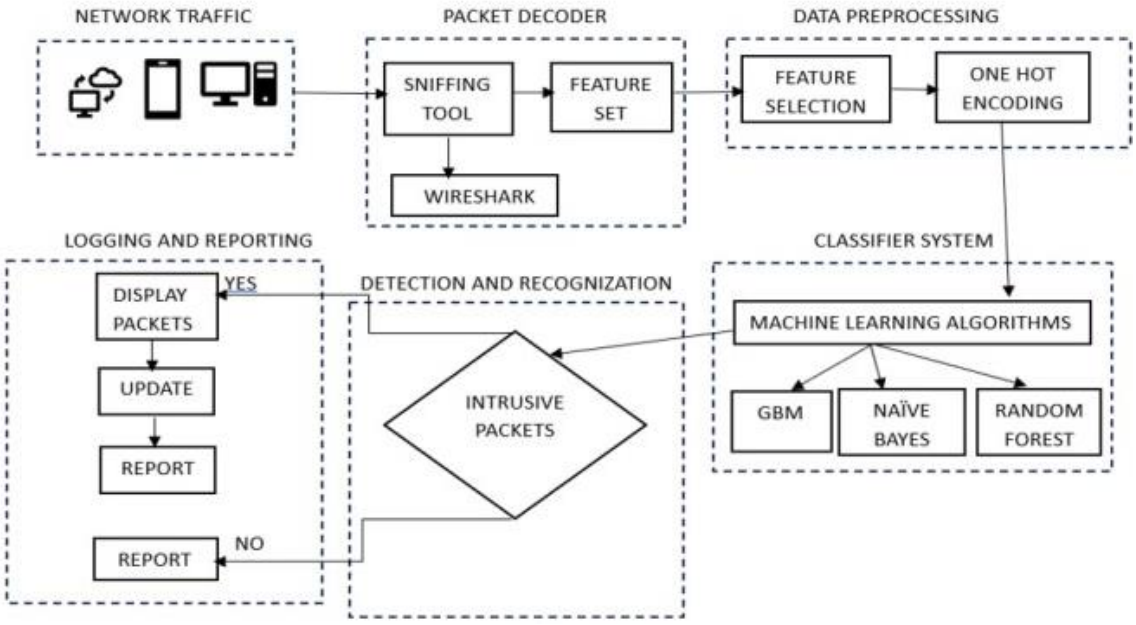

Figure 1. Block diagram of proposed system.

Performance Evaluation Metrics

The evaluation metrics used in this system are accuracy, precision, recall, and F1 Score.

Accuracy is a metric utilized to gauge the of machine learning model's effectiveness. It represents the ratio of correctly predicted cases out of the complete number of cases assessed. While accuracy provides a straightforward measure that is easy to interpret, it may not suffice, particularly in the cases where one class is dominating over the other.

Accuracy is a crucial measure for evaluating the efficacy of classification models, especially in binary classification processes. It quantifies the model's capability to make the right predictions by indicating the proportion of accurate predictions for each event where the prediction was accurate. Essentially, accuracy assesses the model's proficiency in minimizing errors. A high accuracy score implies a lower false positive rate, rendering the model reliable in tasks where minimizing false positives holds significance, such as medical diagnosis or fraud detection.

Precision, also known as sensitivity or true positive rate, is another metric employed to evaluate classification models, especially in contexts where determining the correct status is paramount. It signifies the percentage of true positive cases identified by a sample out of all positive results in the given dataset.

F1 score is a commonly employed metric in machine learning and statistics for assessing the performance of classification models [20]. It amalgamates precision, and recall into a single score, ensuring a balanced evaluation of sample accuracy. Precision measures the ratio of correctly predicted results among all the positive predictions made by the model, whereas recall gauges the percentage of correctly predicted results accurately identified in each case. The F1 score computes the harmonic mean of precision and recall, giving equal weight to both measure

RESULTS

Step-by-step procedure of execution is shown in Figures 2 and 4 while results after evaluation through the proposed system is explained below through window pages in Figures 5 and 6.

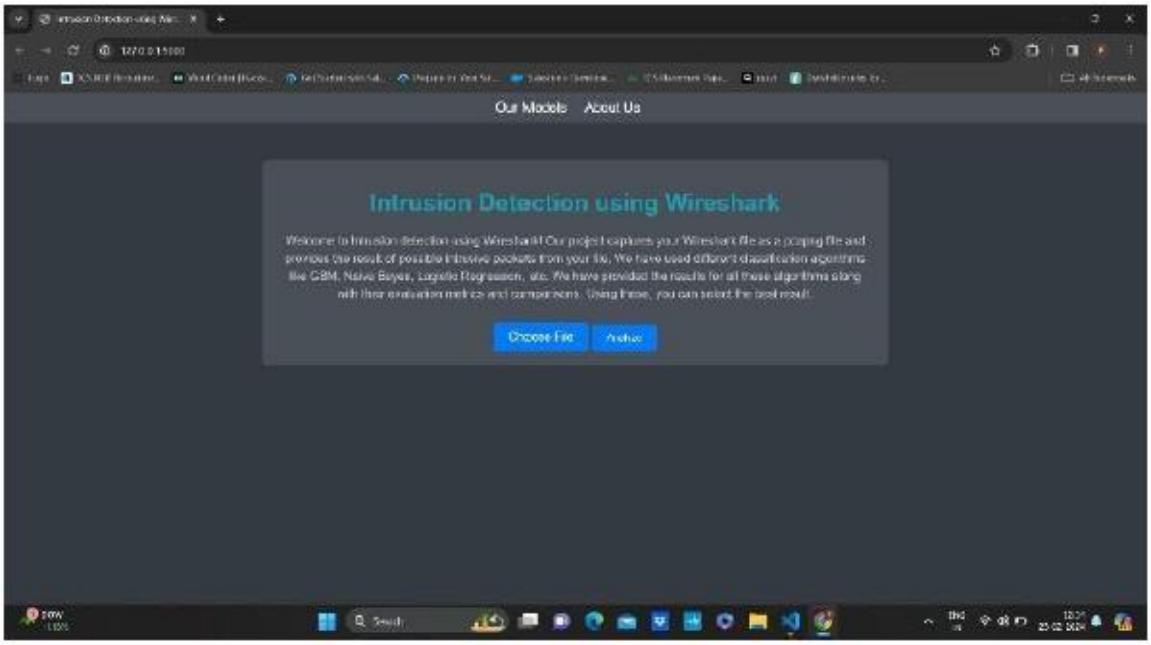

Figure 2. Caption-Homepage viewed after visiting the website. Select choose file.

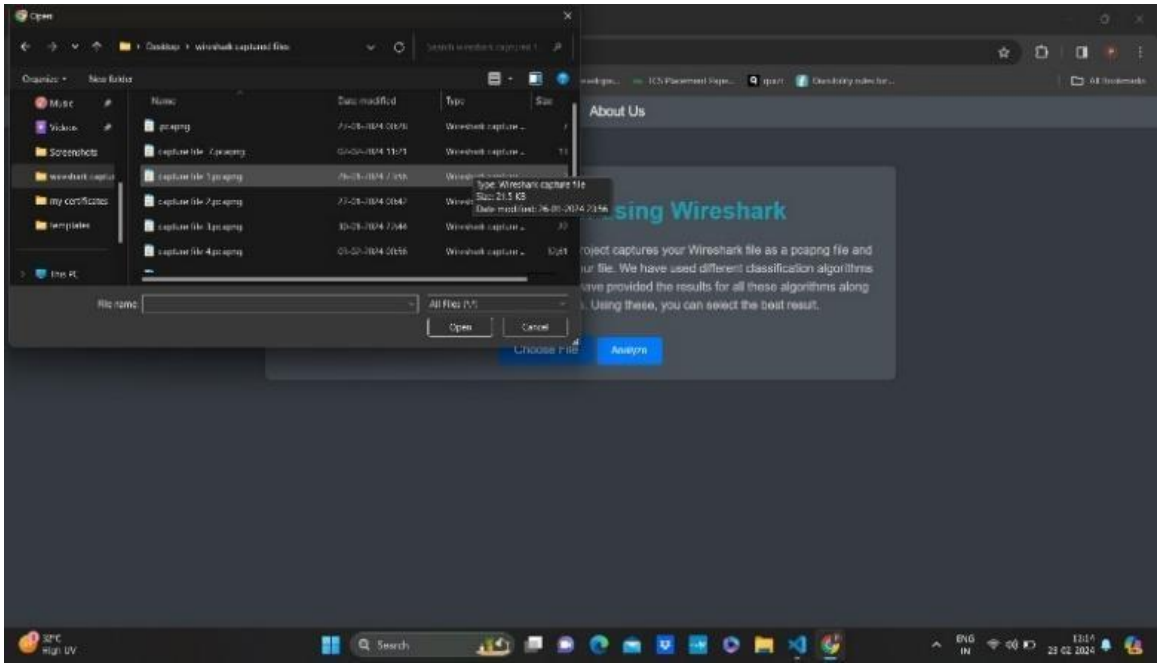

Figure 3. Uploading the file to the website.

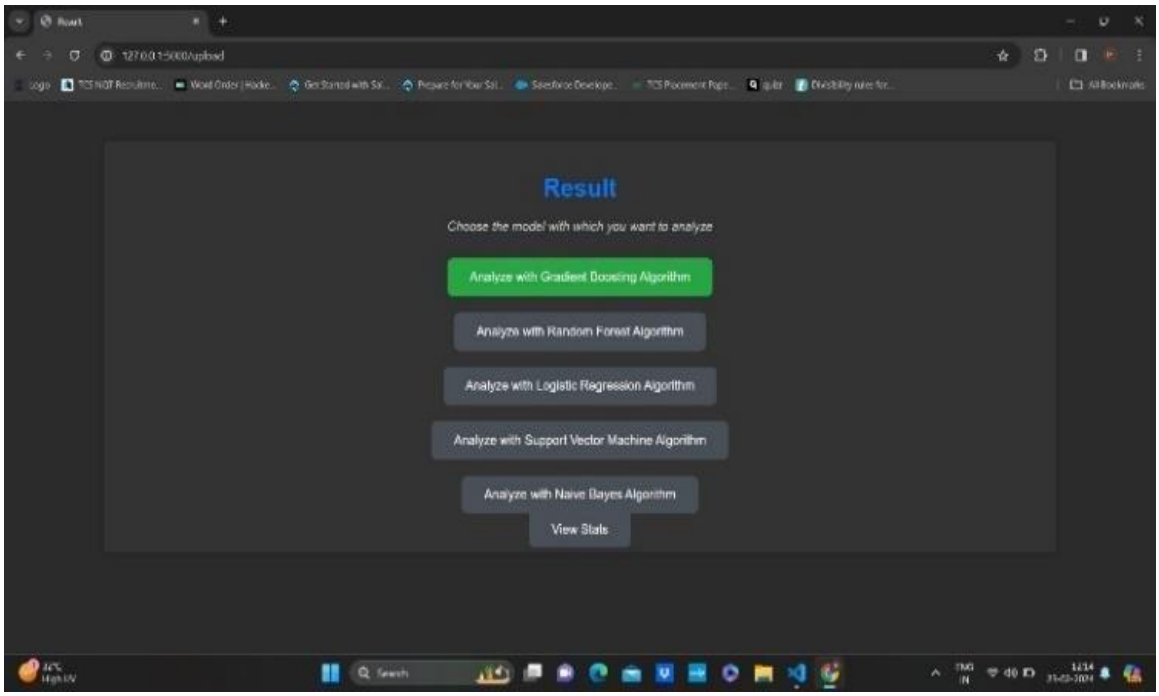

Figure 4. Page for algorithm selection.

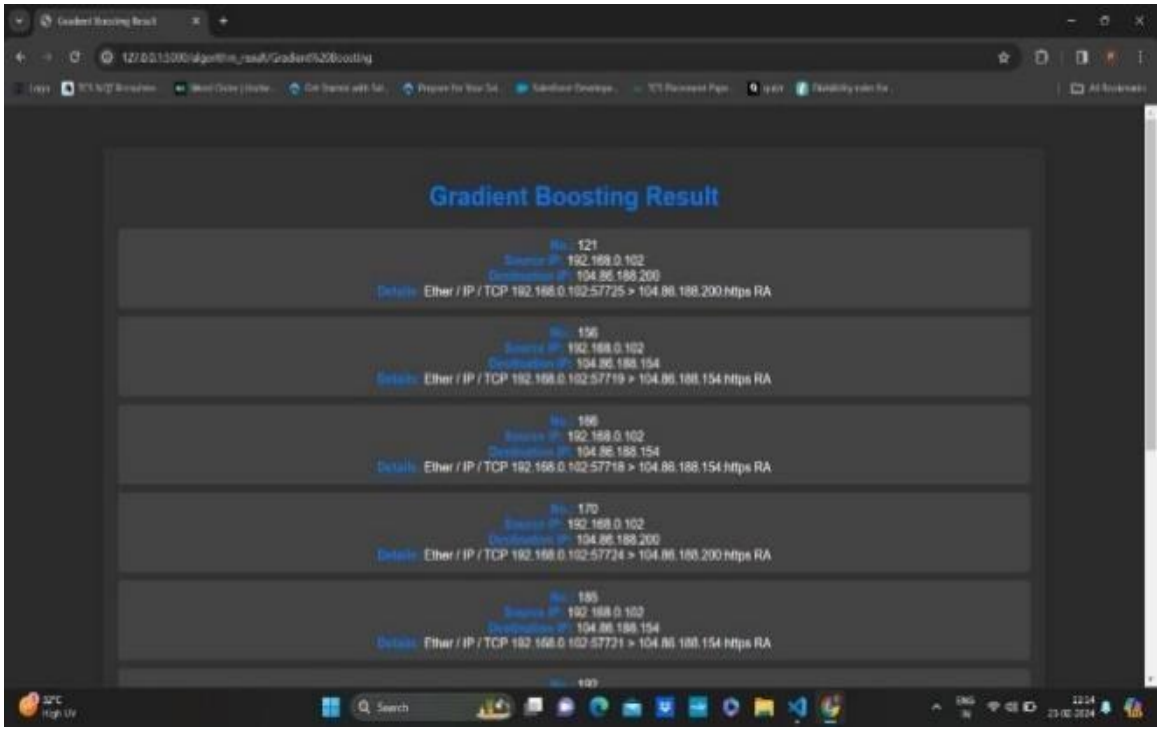

Figure 5. Displaying results for GBM algorithm.

CONCLUSION

The comparative study of machine learning models like Naïve–Bayes, Support Vector Machine(SVM), Random-forest, Logistic Regression and Gradient Boosting machine with the other classifiers and also all the classifiers at once, so that the users can themselves determine which algorithm to be more useful when compared with the other algorithms used in the system for the prediction and classification of various types of data packets as whether it is normal packet or encrypted packet or normal packets.

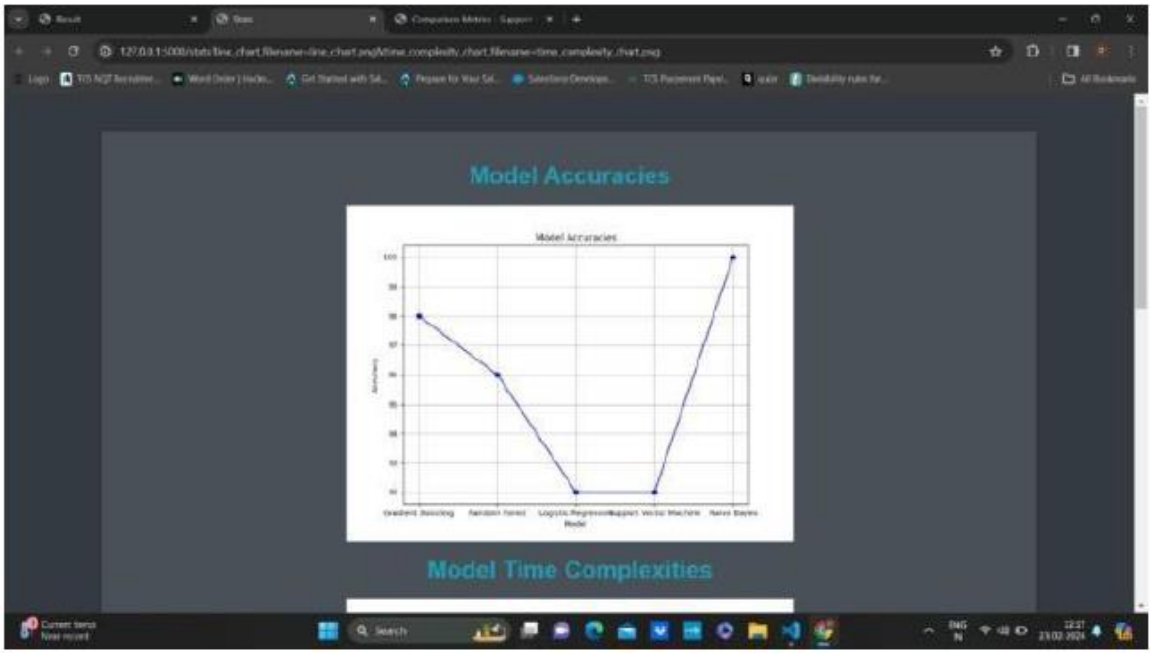

Figure 6. Caption-accuracies of different algorithms for a given real-time dataset.

REFERENCES

1. Alia Y, Eric A. Evaluation of capabilities of Wireshark as intrusion detection system. J Glob Res Comput Sci. 2018; 9 (8).
2. Kumar S. Detect/analyze scanning traffic using Wireshark. PenTest Magazine, 2013, June.
3. Pavithirakini S, Bandra DDMM, Gunawardhana CN. Improve the capabilities of Wireshark as a tool for intrusion detection in DOS attacks. Int J Sci Res Publicat. 2016; 6 (4): 378–384.
4. Naaz S, Badroo FA. Investigating DHCP and DNS protocols using Wireshark. IOSR J Comput Eng. 2016; 18 (3):1–8p.
5. Pottner W-B, Wolf L. IEEE 802.15.4 packet analysis with Wireshark and off-the-shelf hardware. Institute of Operating System and Computer Networks.
6. Khan M, Alshomrani S, Qamar S. Investigation of DHCP packets using Wireshark. Int J Comput Appl. 2013; 63 (4): 1–9p.
7. Choudhary S, Singh N. Safety measures and auto detection against SQL injection attacks. Int J Eng Adv Technol. 2019; 9 (2): 2827–2833p.
8. Sinha K, Choudhary S, Paul S, Paul P. Security of multimedia in cloud using secret shared key. International Conference on Computing, Power and Communication Technologies. IEEE, Greater Noida, India. 2018. pp. 908–912.
9. Iqbal H, Naaz S. Wireshark as a tool for detection of various LAN attacks. Int J Comput Sci Eng. 2019; 7 (5): 833–837.
10. Banerjee U, Vashishtha A, Saxena S. Evaluation of Capabilities of Wireshark as a tool for intrusion detection system. Int J Comput Appl. 2010; 6 (7).
11. Chiu M-H, Yang K-P, Meyer R, Kidder T. Analysis of a man-in-the-middle experiment with Wireshark.
12. Bejtlich R. The Tao of network security monitoring: beyond intrusion detection. Pearson Education; 2004.
13. Stolze M, Pawlitzek R, Wespi A. Visual problem-solving support for new event triage in centralized network security monitoring: challenges, tools and benefits. GI-SIDAR Conference IT-Incident Management and IT-Forensics (IMF); 2003.
14. Roesch M. Snort-lightweight intrusion detection for networks. Proceedings of Thirteenth Systems Administration Conference (LISA). 1999. pp. 229–238.
15. Pinkas B, Sander T, Securing passwords against dictionary attacks. Proceedings of the 9th ACM

Conference on Computer and Communications Security. Association for Computing Machinery, Washington, DC, USA. 2002. pp. 161–170.
16. Lee W, Stolfo SJ, Mok KW. Adaptive intrusion detection: a data mining approach. Artif Intell Rev. 2000; 14 (6): 533–567p.
17. Ndatinya V, Xiao Z, Meng K. Network forensic analysis using Wireshark. Int J Sens Netw. 2015; 10 (2): 91p.
18. Hebbar R, Mohan K. Packet analysis with network intrusion detection system. Int J Sci Res. 2015; 4 (2): 2246–2249p.
19. Tsai C-F, Lin C-Y. 'A triangle area based nearest neighbors' approach to intrusion detection. Pattern Recognit. 2010; 43 (1): 222–229p.
20. Mishra P, Varadharajan V, Tupakula U, Pilli ES. A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Commun Surv Tutor. 2018; 21 (1): 686–728p.

