Real-Time Intrusion Detection based on Network Monitoring and Machine Learning

Subhadeep Chakraborty
Researcher and Developer
Artificial Intelligence and Cyber Security
Kolkata

ABSTRACT
The threat of cyber-attacks continues to grow in significance as our world becomes increasingly reliant on digital communication and data storage. Network intrusion detection aims to prevent unauthorized access, protect sensitive information and preserve digital systems' integrity. In this in-depth research study, the network-based intrusion detection methods for identifying suspicious and malicious network activities have been examined. By utilizing network monitoring techniques to capture and collect data, along with machine learning algorithms for classifying intrusive packets, network intrusion detection has been performed. Initially, Wireshark has been employed for network monitoring, successfully capturing a total of 28,889 packets and storing them in PCAP format. This PCAP data was then used for preliminary analysis to investigate the properties of the packets. Following this, the PCAP data was converted into a Comma Separated File format, facilitating the application of machine learning algorithms for further examination. During data preparation, Principal Component Analysis has been employed to detect and eliminate outliers, along with the Chi-Squared method for feature selection. The research involved testing five machine learning models on the processed data, to achieve the most effective model for intrusion detection. Ultimately, the Random Forest algorithm has excelled, boasting an outstanding accuracy rate of 99.43%. This comprehensive investigation highlights the immense potential of network-based intrusion detection methods for uncovering and addressing cyber threats on both a large and small scale.

General Terms
Monitoring Network Traffic, Network Packet Capturing, Network Class Design, Intrusion Classification and Detection.

Keywords
Network Monitoring, Intrusion Detection, Classification, Machine Learning, Cyber Security.

1. INTRODUCTION

1.1 Network Intrusions and Impacts

1.1.1 Overview
The technological revolution and the increasing reliance on interconnected networks have significantly changed the landscape of data communication and the associated risks. Network intrusions, which comprise unauthorized access or breaches to information systems, are undoubtedly one of the most pressing challenges in our digital era [1]. These intrusions can lead to a series of tangible and intangible consequences that can significantly affect organizations as well as individuals. Network intrusions pose a serious threat to the security and privacy of sensitive data, often resulting in significant financial and non-financial losses, thus necessitating the adoption of robust cybersecurity measures.

1.1.2 Impacts
One of the most significant effects of network intrusion is the potential loss of data. This can include sensitive information such as personal identity data, financial data, or other intellectual property. Not only can this be a violation of privacy, but it can also result in significant financial consequences for individuals and businesses alike [2]. For instance, the average cost of a data breach in the US has risen to $8.64 million, and this number could continue to increase. Moreover, network intrusions can lead to service disruption, causing an organization's websites, software applications, and other online services to become unavailable to customers. In some instances, this could drastically impact a company's revenue, leading to business failure or financial loss [3]. Additionally, a network intrusion can harm a company's reputation, leading to a loss of trust from customers and partners.

1.1.3 Possible Solutions
The possible solutions to get rid of the possibility of network attacks are discussed below:

- There are steps which can be taken to protect the data and mitigate the effects of a network intrusion. First, the organization must implement security measures like firewalls, antivirus software, and intrusion detection systems [1]. These tools can provide real-time protection against threats, potential and actual intrusions, and malware.
- Another critical strategy for protecting the data is implementing strong password policies, including the use of multi-factor authentication. It is vital to have unique and complex passwords for each of the online accounts and to change them routinely [4]. Additionally, encrypting sensitive data and performing regular backups can reduce the potential damage if a network intrusion occurs.
- Lastly, regular employee training and awareness are essential. Employees can unknowingly make an organization vulnerable to network intrusions, so it is critical to educate them on the warning signs and preventative measures to help minimize the risk of an attack [5]. Furthermore, it is essential to limit employee access and control administrative permissions and privileges selectively.

1.2 Network Intrusion Detection Systems
Network intrusion detection systems (NIDS) serve as a digital "guardian" for an organization's valuable information, monitoring traffic, and detecting signs of unauthorized access

or malicious activity [6]. Using algorithms and pattern recognition, these systems analyze data packets within a network in real-time, allowing administrators to promptly identify and investigate potential threats. This level of protection is crucial in preventing data breaches, minimizing financial loss, and preserving an organization's reputation amid the ever-evolving spectrum of cyber threats [2].

The efficacy of a network intrusion detection system lies in its ability to predict cyber attacks and provide proactive protection. By monitoring patterns of behaviour, a NIDS can identify anomalies and initiate a response before the malicious activity progresses into a widespread attack [7]. This early detection is essential in preventing the spread of malware, ransomware, and other cyber threats within an organization's network. By employing a robust NIDS, organizations improve their digital security and reduce the likelihood of being targeted by sophisticated cybercriminals [8].

So, network intrusion detection is an indispensable aspect of modern cybersecurity. As technology continues to advance and the threatscape expands, protecting an organization's data and digital systems is of paramount importance [9]. Harnessing network intrusion detection systems can significantly improve response times and reduce the severity of potential breaches, ensuring that organizations remain secure and vigilant in an increasingly connected world.

1.3 Network Packets and Transactions

1.3.1 Overview

Network packets and transactions play a crucial role in enabling communication, facilitating secure data transmission, and enhancing the performance of digital systems. Understanding their functions, limitations, and applications is essential for the continued development of efficient and secure digital communication and transaction platforms [5]. In today's highly interconnected world, technology has enabled seamless communication between individuals and businesses. This has revolutionized the way we access information, collaborate, and conduct transactions. Network packets and transactions are fundamental components of digital communication systems [10]. They serve as the building blocks for conveying data across various networks, ensuring secure and efficient communication.

1.3.2 Network packets

Network packets, also known as data packets, are the smallest units of data transmitted over a network. They are composed of headers, which contain the necessary information for routing, and payloads, which carry the actual data being transmitted [11]. The process of packet-switching ensures that these packets reach their destination efficiently and reliably. This technique divides the data into smaller segments suitable for transmission across various network paths, reassembling them at the destination [12]. This approach improves the overall network performance, allowing for concurrent data transfer and improved resource utilization.

1.3.3 Network Packet Transactions

Transactions refer to a series of coordinated network operations executed to accomplish a specific task, such as data retrieval, modification, or transfer. These operations adhere to the ACID (Atomicity, Consistency, Isolation, and Durability) properties, ensuring data consistency and system reliability [3]. The concept of transactions becomes particularly critical in financial systems, where secure and accurate data exchanges are paramount. Cryptographic techniques, such as blockchains and digital signatures, are often employed to guarantee the authenticity, integrity, and confidentiality of these transactions.

2. LITERATURE REVIEW

2.1 Network Monitoring

The researchers [12] introduced Newton, a motive-driven traffic display designed to satisfy the needs of modern-day networks in terms of scalability and versatility. With the exponential boom of network bandwidth and extent, existing monitoring structures struggle to provide on-call for network monitoring without incurring widespread overhead. Newton addresses this assignment with the aid of allowing operators to specify their monitoring intents via traffic tracking queries, enabling the dynamic and scalable deployment of network-wide queries. Operators can personalize and regulate queries on the fly, making sure of uninterrupted community workflow. Newton also consists of optimizations at the tool and community-wide degrees to limit useful resource intake all through question deployment.

The researchers [6] introduced and applied H2Classifier, a solution for tracking HTTP/2 visitors over TLS without the need for decryption. Maximum net offerings now use HTTPS, it's miles critical to balance consumer privacy with the ability to stumble on particular person movements primarily based on protection regulations. Unlike monitoring techniques for HTTP/1.1 over TLS, H2Classifier makes a speciality of passive site visitors analysis and employs a random wooded area classifier to become aware of the predefined user moves on a monitored web service. One of the challenges is extracting representative values of loaded content related to a web webpage, that is customized based totally on person actions. Extensive opinions concerning five popular internet services exhibit the effectiveness of H2Classifier, attaining accuracy stages starting from 94% to 99%.

The authors [8] addressed the fundamental issue of accurate network traffic detection, which forms the basis for network traffic management and data analysis, which ultimately optimizes users' performance The study examines two detection methods for network traffic: machine learning and deep packet analysis. A novel approach combining machine learning and deep packet analysis is proposed to realize network traffic detection. Deep packet inspection technology is used to detect most network traffic, reducing the work required for machine learning-based identification Furthermore, deep packet inspection enables specific application traffic to be detected, thus improving the accuracy of the identification. Compensating for the limitations of deep packet inspection in detecting new applications and encrypted traffic, machine learning is used to help detect network traffic with encrypted and unknown characteristics Experimental results show that this approach this combination significantly increases the network traffic detection rate.

As network technology continues to evolve, accurate vehicle identification has become increasingly important for network management and security. However, the challenge still lies in formulating features that can adequately describe network traffic, especially in complex dynamic networks. An alternative approach to addressing these issues is proposed in this paper [10]. The method uses the hierarchical framework to study the communication characteristics of network traffic. Experimental results show the effectiveness of our method to improve detection capability and obtain better classification performance.

2.2 Feature Selection Process

The researchers [13] have employed the process to select necessary features by combining the Chi-Square Selection (CSS) method with supervised machine learning to enhance

feature selection for Covid-19 data. In order to identify the features that are most important for Covid-19 prediction, the authors proposed a novel strategy that incorporates CSS into the feature selection procedure. The data were analyzed using a variety of supervised machine learning algorithms, such as proximity-based, decision boundaries, and ensemble techniques, all of which are well-known for their effectiveness in classification tasks. Using a Covid-19 dataset, the authors tried out their proposed method and compared it to other methods for selecting features. The method used in this study had an accuracy of 95%.

A study on the effect of feature selection on a document sentiment analysis classifier's performance is presented in this article [14]. The authors selected relevant features from the dataset using a specific strategy. The selected features were then used to train and evaluate the Naive Bayes classifier. The review looked at the presentation of the Gullible Bayes classifier with and without Chi-Square element determination concerning accuracy. The results demonstrated that the accuracy of the Naive Bayes classifier with Chi-Square feature selection was 86% higher than that of the classifier without feature selection, which was 78%. The study suggests that incorporating Chi-Square feature selection into the Naive Bayes classifier for document sentiment analysis might be beneficial.

The creators utilized the Chi-Square measurements measure to rank the significance of elements in text information and chose the highest-level highlights as contributions for grouping. The review utilized a dataset for text grouping and assessed the proposed strategy by contrasting it and other feature selection techniques [15]. With an accuracy of 92% in the experiments, the Chi-Square statistics-based feature selection method outperformed other methods in terms of accuracy. According to the review, the study's method effectively selects relevant features for text classification tasks, resulting in improved classification accuracy.

An overview of a study on a novel Type I fuzzy ensemble feature selection method is provided in this article. A method for selecting machine learning features that make use of Type I fuzzy sets was proposed by the authors. As their trial information, they utilized a dataset from an Iranian Joint Congress on Fluffy and Clever Frameworks [1]. The creators set their proposed troupe highlight decision strategy up as a regular occurrence and contrasted it with different techniques like Data Gain and ReliefF. Their proposed strategy was 92% accurate, while Data Gain and ReliefF were 87% and 88% accurate, separately. The study demonstrates the efficacy of a novel ensemble-based method for feature selection that makes use of Type I fuzzy sets, demonstrating its potential to boost performance in machine learning feature selection tasks.

An overview of a study on a feature selection ensemble for the Analytic Hierarchy Process (AHP) to classify symbolic data is provided in this article. To improve the accuracy with which symbolic data can be classified, the authors came up with an ensemble approach that combines AHP with multiple feature selection methods like ReliefF and Information Gain. Using a dataset from the 24th International Conference on Pattern Recognition (ICPR), they ran experiments and compared their proposed method to other methods for selecting features. Their feature selection ensemble with AHP achieved 95% accuracy, surpassing other approaches like Relief (88 per cent) and Information Gain [16]. The review features the capability of integrating AHP into machine learning tasks for improved performance and demonstrates the effectiveness of the proposed method in improving feature selection for symbolic data classification.

The researchers [5] framed a concentrate on a component decision methodology for network interruption discovery utilizing a multi-distance gathering and element bunching strategy. In order to improve the accuracy of network intrusion detection, the authors proposed a novel strategy that combines feature clustering with multiple feature selection methods. A feature clustering algorithm and a multi-distance ensemble method are two of the algorithms used in this strategy. The authors ran experiments and compared their proposed method to other methods for selecting features using ISSI. The accuracy got utilizing their methodology was 97%, outflanking different strategies like ReliefF (92%) and Data Gain (89%). The review exhibits the adequacy of the proposed approach in further developing element decisions for network interruption recognition and features the capability of involving gathering strategies and component grouping procedures in this space.

2.3 Intrusion Detection

According to [3], Distributed traffic that is anomalous faces difficulty in terms of getting detected, as it gets dispersed simultaneously to the many numbers of links while tending to be not present in any obvious anomalous features in a single link. This paper has proposed a detection method with multi-scale spatial besides stealthy traffic anomaly distribution, it is able to deploy an early-stage detection through key nodes of the network. Packet analysis takes up a Multi-scale wavelet performed in separation with the links at which information regarding each node is available, with the abnormal frequency getting aim of different time sections range and signal reconstruction with anomalous features.

According to [17], switching to an All-optical packet that has been through an intensive investigation over the odyssey in recent years seems to be playing the alternative static role for, networks that are cross-connect based. Several associated with the switch have got proposed, where all of them with the buffers made the delay lines fibre being used. This paper addresses the packet switching basic concepts in the optical domain along with describing an approach of analytics for the evaluation of the performance of end-to-end networks through the employment of slotted (fixed length) optical packets. Here a network dimensioning procedure also gets presented that relies upon the said approach.

According to [18], For traffic in real, which includes bursty traffic patterns, due to the use of wavelength that is tuneable converters get recognized as quite essential for complexity reduction of the photonic wavelength division multiplexing (WDM) switches in the packet. Results get obtained from a traffic model that is analytical along with including a buffering in the domain of wavelength that accounts for traffic that is bursty. The model which is theoretic gets verified by simulations and from the model, it has been found that higher loads of traffic, as well as burstiness, can get accepted when wavelength converters tuneable are used. Therefore, a throughput that is large as compared to the photonic packet switches gets obtained and notably, this gets achieved while keeping the gates number required for the realization of the space switches that are nearly constant.

The researchers [9] discussed a unique deep neural network model specifically designed for computer network intrusion detection, emphasizing the importance of network systems connection efficiency. Within the intrusion detection system, key parameters are established by defining features of intrusion and normal user behaviour. Additionally, an immune genetic

algorithm is introduced, enabling the exploration of the entire measurement space to enhance the parameter set of measurement. In the experimental section of this study, the model's performance is thoroughly evaluated and, ultimately, the outcomes are summarized in the conclusion, demonstrating the effectiveness of the neural network model for intrusion detection within computer networks.

In the field of network intrusion detection, researchers have developed a method that employs machine learning techniques [7]. This innovative approach utilizes algorithms to determine whether the requested information is legitimate or contains anomalies. To preprocess the dataset, feature selection algorithms such as Correlation-Based and Chi-Squared are used to remove any irrelevant data. The experimentation was conducted using the NSL KDD dataset, which showed varying degrees of accuracy for different models. The Support Vector Machine (SVM) model achieved an accuracy of approximately 48%, while the Artificial Neural Network (ANN) model boasted an impressive 97% accuracy rate. Based on these results, it is evident that the ANN model significantly outperforms the SVM model in detecting network intrusions within this dataset, demonstrating a promising direction for further research in the area.

The researchers [19] developed a hybrid intrusion detection system for network interface devices, integrating multiple machine learning methods and inference detection in a comparative analysis. This approach comprises two stages: the Training stage, which consists of model training and building the inference network, and the Detection stage during operation. The primary objective is to address current real-world challenges in network learning algorithms, as they could potentially fail beyond their categorical domains. When tested individually, the accuracy rates were 98.088%, 82.971%, 95.75%, and 81.971% respectively. In contrast, when assessed in conjunction with the inference detection model, the accuracy rates improved to 98.554%, 66.687%, 97.605%, and 93.914% respectively. These results underscore the effectiveness of the inference detection model in addressing specific aspects that traditional machine learning techniques may overlook.

Intrusion Detection Systems (IDS) are an indispensable component of network infrastructure as they help in detecting and deterring cyber-attacks [20]. Although the adoption of Artificial Intelligence (AI) can enhance the functionality of IDS and fortify cybersecurity, there is a pressing need to ensure the availability of high-quality datasets to improve AI algorithms' efficacy. To address this challenge, a feature selection method is deployed to identify suitable features, and the dataset is balanced by using samples from it to test various Machine Learning (ML) methods. Recent research suggests that this approach has yielded exceptional results, as IDS using AI algorithms can detect 99% of intrusions accurately.

In the realm of cybersecurity, intrusion refers to malicious activities that aim to gain unauthorized access to sensitive information [2]. The Intrusion Detection System (IDS) serves as a crucial component in identifying and reporting such attacks to system administrators, thereby safeguarding valuable data. To achieve this, the IDS analyzes network performance and compares new activities with historical data. This particular study focuses on the implementation of an anomaly-based IDS, designed to effectively detect unauthorized activities originating from both external and internal sources. Notably, the Random Forest Classifier has demonstrated exceptional performance in this context, attaining an impressive 99.8% accuracy rate and a macro-average F1-Score of 0.98.

3. METHODOLOGY

3.1 Proposed Method

The proposed methodology for the research to detect network intrusions has been presented below:

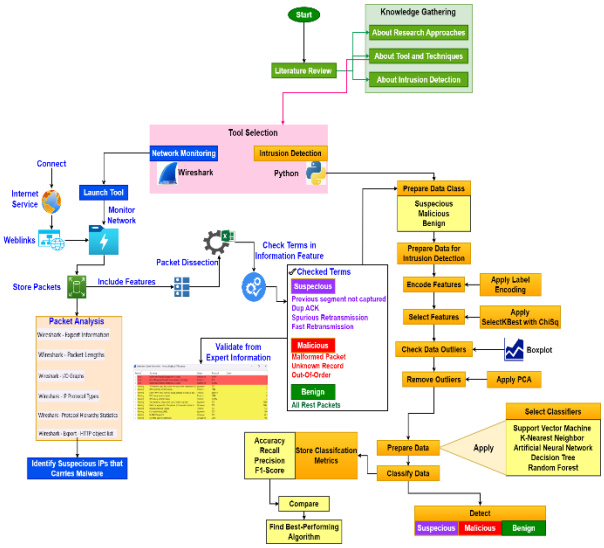

Figure 1 Proposed Method for Network Intrusion Detection

The methodology for detecting host-based intrusions, as illustrated in Figure 1, utilizes a real-time detection process. This process involves capturing network packets via network monitoring with Wireshark and subsequently applying machine learning to analyze the gathered data. A review of previous research reveals that many researchers have relied on traditional datasets, such as KDD+, NSL-KDD, and CICIDS, for intrusion detection. However, it is important to note that these sources primarily contain outdated threats, while the nature of intrusions evolves daily. Therefore, to maintain relevancy and effectiveness in addressing the current threat landscape, it is more advantageous to capture real-time network packet data and apply machine learning techniques for analysis. This approach will enable the identification and mitigation of contemporary threats or intrusions, thus justifying the proposed methodology's implementation.

3.2 Tools for Research

3.2.1 Wireshark

It is a widely utilized and highly regarded tool for network monitoring and packet capturing. In this research, Wireshark will be employed to closely observe real-time network activity while capturing and generating critical PCAP data [21]. Subsequently, the acquired PCAP data will undergo thorough analysis by utilizing the capabilities of this powerful open-source tool, which is available at no cost.

3.2.2 Python

It has established itself as a premier tool for data analytics and data prediction, particularly in the burgeoning field of cybersecurity. In this project, Python's versatile abilities will be harnessed to analyze the collected data, identify potential network-based intrusions and create rules for intrusion detection systems [7]. The implementation of classification technology plays a crucial role in bolstering cybersecurity efforts. Similar to Wireshark, Python is an open-source tool that can be freely accessed and utilized.

3.3 Algorithms Selected

3.3.1 Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm that is used for classification and regression. It works by constructing hyperplanes in multidimensional spaces to divide data into different classes [22]. These hyperplanes are selected in such a way that they are equidistant from the data points of the nearest class. SVM has proven effective in handling high-dimensional data and has been utilized in intrusion detection systems for its low false-positive rates.

3.3.2 Decision Tree (DT)

DT is a hierarchical, supervised learning algorithm that constructs tree-like structures to represent the relationships between input features and output classes [21]. The tree is generated through recursive partitioning of the input space into smaller regions, and the final output is determined by traversing the tree from the root to a leaf node. DT has been implemented in intrusion detection to classify attacks accurately and efficiently, as it offers a readable and understandable representation of the decision-making process.

3.3.3 Neural Network (NN)

NN is a widely-used machine learning algorithm inspired by the human brain's structure and function. It consists of multiple layers of interconnected neurons that learn and adjust their weights based on input data [11]. NN is known for its adaptability and ability to model complex relationships between inputs and outputs, making it suitable for intrusion detection, where patterns and attack sequences can be highly variable.

3.3.4 Random Forest (RF)

RF is an ensemble learning algorithm that constructs a collection of decision trees and combines their outputs through majority voting or averaging [23]. This approach enhances the overall classification performance and reduces overfitting. In intrusion detection, RF is preferred for its robustness against noisy data and its scalability in handling large data sets, such as network traffic.

3.3.5 K-Nearest Neighbor (KNN)

KNN is a non-parametric, instance-based learning algorithm that stores the entire training data set and classifies new instances based on their proximity to existing data points. It calculates the Euclidean distance between the test instance and the k-training samples closest to it and assigns the majority class among the k-nearest neighbours [4]. In intrusion detection, KNN's versatility stems from its simplicity of implementation and ability to adapt to changes in input distributions without retraining.

4. INTRUSION DETECTION RESULT

4.1 Network Monitoring and Packet Capturing

To capture the packets, the below-mentioned web addresses have been visited and parallelly the Wireshark has been enabled to capture the network traffic packet as shown below:

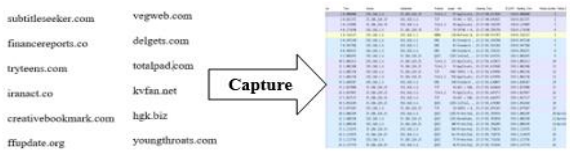

Figure 2 Network Packet Capturing

4.2 Network Packet Analysis

4.2.1 Expert Analysis

Expert analysis plays a crucial role in identifying the status of data packets and detecting any suspicious elements. The analytical process classifies all packets into four distinct categories - Error, Warning, Note, and Chat. Suspicious packets are predominantly found within the Error and Warning categories. The outcome of the analysis has been presented below:

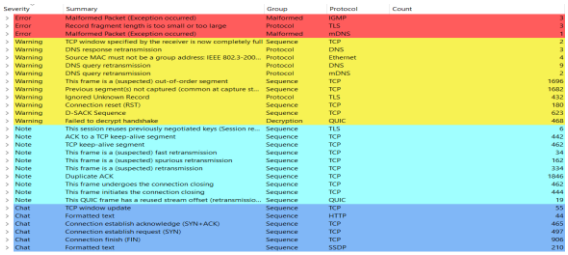

Figure 3 Expert Analysis Result

4.2.2 Extracted Strings

The identification of the network packets will be done based on their behaviours and types. To identify those, the types of packets have been determined from expert analysis. The necessary string to understand the severities of packets have been presented below:

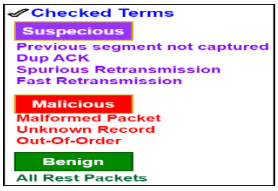

Figure 4 Network Packet Severity Strings

4.2.3 Input Output Graphs

The input and output graph generated by Wireshark offers valuable insights into the packet flow rate over time. Within this captured data, numerous protocols were involved, and their respective outcomes will be discussed in this section. Our analysis of HTTP traffic revealed that the highest rate of packet transactions reached was 1.8 packets per second. Meanwhile, UDP traffic reached 260 packets per second, QUIC traffic reached 350 packets per second, and TLS traffic peaked at 750 packets per second. This comprehensive understanding of packet flow rates provides a professional overview of the network's performance during this period.

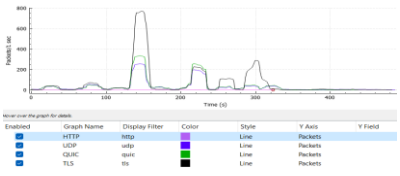

Figure 5 I/O Graph Analysis

4.3 Feature Addition and Data Transformation

4.3.1 Feature Addition

Wireshark provides the default features of packets which may not be enough to detect network intrusions. So, the necessary features have been added to the PCAP data and the final feature set has been shown below:

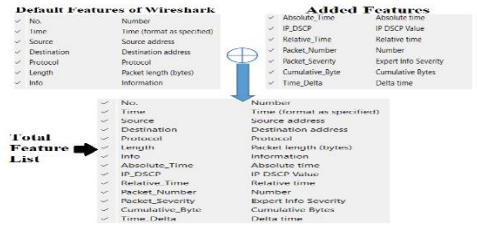

Figure 6 Feature Addition

4.3.2 Data Transformation

The PCAP data with added features have been converted to CSV data and the descriptions of the features have been presented in below table:

Table 1 Description of Data Features

Feature	Details	Feature	Details
No.	Serial of the packets captured	Absolute_Time	The absolute time is taken for the packet transaction
Time	Time of capturing	IP_DSCP	IP Priority number
Source	Source IP Address	Relative_Time	The relative time for the packet transaction
Destination	Destination IP Address	Packet_Number	Number of the packet
Protocol	Protocol of the captured packet	Packet_Severity	The severity of the packets
Length	Length of the captured packet	Cumulative_Byte	The cumulative Bytes included in the packets in a serial transaction
Info	Information of the captured packet	Time_Delta	Time difference between two transactions

4.4 Intrusion Detection

4.4.1 Feature Information

The information of the features of the CSV data has been checked wherefrom it has been observed that the data contains a mixture of feature types.

```
RangeIndex: 28889 entries, 0 to 28888
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   No.              28889 non-null  int64
 1   Time             28889 non-null  float64
 2   Source           28889 non-null  object
 3   Destination      28889 non-null  object
 4   Protocol         28889 non-null  object
 5   Length           28889 non-null  int64
 6   Info             28889 non-null  object
 7   Absolute_Time    28889 non-null  object
 8   IP_DSCP          28874 non-null  object
 9   Relative_Time    28889 non-null  float64
 10  Packet_Number    28889 non-null  int64
 11  Packet_Severity  9501 non-null   object
 12  Cumulative_Byte  28889 non-null  int64
 13  Time_Delta       28889 non-null  float64
dtypes: float64(3), int64(4), object(7)
```

Figure 7 Information of Features

4.4.2 Data Cleaning

The inspection has been done on the missing values and the identified missing values have been removed from the features.

```
No.                0          No.                0
Time               0          Time               0
Source             0          Source             0
Destination        0          Destination        0
Protocol           0          Protocol           0
Length             0          Length             0
Info               0          Info               0
Absolute_Time      0          Absolute_Time      0
IP_DSCP           15          IP_DSCP            0
Relative_Time      0          Relative_Time      0
Packet_Number      0          Packet_Number      0
Packet_Severity  19388         Packet_Severity    0
Cumulative_Byte    0          Cumulative_Byte    0
Time_Delta         0          Time_Delta         0
```

Figure 8 Cleaning Missing Values

4.4.3 Target Feature Creation

As the CSV data transformed from the PCAP data does not have any class labels, those have been created by taking the extracted strings (Figure 4) into concern as shown below:

Table 2 Target Class Creation

Target Class	Info Includes
Suspicious	Previous segment not captured Dup ACK Spurious Retransmission Fast Retransmission
Malicious	Malformed Packet Unknown Record Out-Of-Order
Benign	Otherwise

4.4.4 Outlier Treatment

The detection of outliers has been done by employing Principal Component Analysis through computing feature variances. The detected outliers (variance higher than 80%) have been removed by applying data normalisation. Using the data normalization, the variance has been reduced to 52% from 99%.

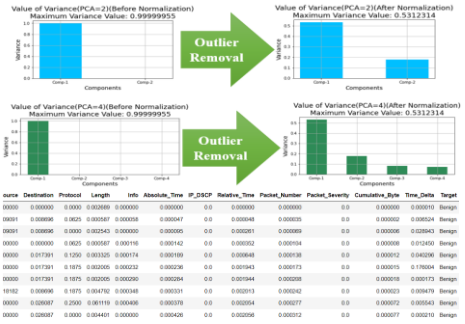

Figure 9 Outlier Removal using PCA and Normalization

4.4.5 Feature Selection

In the final stage of data preprocessing, feature selection is expertly accomplished by utilizing the SelectKBest method, which employs the Chi-Squared technique to pinpoint the desired features. The advantage of implementing SelectKBest lies in its ability to accept anticipated features as parameters, allowing for the subsequent selection of features boasting superior Chi-Squared values. This approach caters to professionals seeking a highly effective and efficient method for feature selection in data preprocessing.

```
Selected Features

Source
Destination
Protocol
Length
Absolute_Time
IP_DSCP
Packet_Severity
Cumulative_Byte
```

Figure 10 Selected Data Features

4.4.6 Data Balancing and Splitting

The data has been seen to be imbalanced which implies that the class labels are not equal. So, the data has been made balanced by features unscaling and split using a 75%-25% ratio for the extraction of train and test data.

Figure 11 Data Balancing and Splitting

4.4.7 Detection Result

The selected classifiers have been employed to detect network intrusions. The result of detection has been presented below in the table:

Table 3 Intrusion Detection Results

Classifier	Accuracy_Train	Accuracy_Test	Precision_Test	Recall_Test	F1-Score_Test	Overfit Model
Random Forest	100	99.43	99	99	99	0.57
K-Neighbors	98.45	97.25	97	97	97	1.2
MLP	92.37	92.35	92	92	92	0.02
Decision Tree	84.18	84.27	85	84	84	0.09
SVC	78.31	78.18	80	78	77	0.13

4.4.8 Comparison of Performances

The performances of the applied classifiers have been compared based on accuracy and overfitting. The comparison has been presented in graphs as follows:

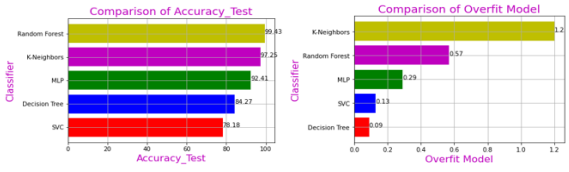

Figure 12 Performance Comparison

4.4.9 Determining Best Model

The comparison of classifiers is showing the fact that Random Forest has detected intrusive, malicious and benign packets with the highest accuracy (99.43%). So, this classifier can be said to be performing the best to detect intrusions. The confusion matrix and classification reports for random forest have been presented below:

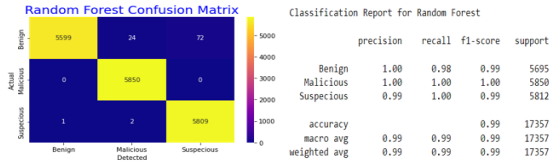

Figure 13 Best-Performing Model for Intrusion Detection

4.4.10 Model Validation

Finally, the random forest model has been validated with the test data by periodic experiments where the sample sizes are 100, 200 and so on. The complete validation result has been presented below:

Test Data	Accuracy	Precision	Recall	F1-Score
100	100.00	100.0	100.0	100.0
200	99.00	99.0	99.0	99.0
300	99.33	99.0	99.0	99.0
400	99.60	100.0	100.0	100.0
500	99.60	100.0	100.0	100.0
600	99.57	100.0	100.0	100.0
700	99.61	100.0	100.0	100.0
800	99.53	100.0	100.0	100.0

Figure 14 Model Validation

5. DISCUSSION

In reviewing previous literature, it has become apparent that the majority of researchers in the field of intrusion detection have primarily utilized collected data from sources such as NSL-KDD, KDD+, and CICIDS among others. Intrusive traffic patterns, however, are subject to change over time, rendering existing data potentially inadequate for the accurate detection of network intrusions. Consequently, the availability of updated and current data is essential for reliable detection.

Notably, data can quickly become outdated due to the dynamic nature of network intrusions. This research seeks to address this issue by utilizing real-time detection methods that capture network packets, ensuring the assessment of only the most current traffic. By implementing this approach, the detection process remains up-to-date, effectively addressing the identified gap in current research practices.

6. CONCLUSION

In the present research, the aim has been taken to identify the characteristics of network traffic and subsequently determine the nature of network packets, whether they are suspicious (intrusive), malicious, or normal (benign). To achieve this, network monitoring has been conducted using Wireshark, capturing a total of 28,889 packets. The initial analysis was performed using Wireshark to uncover network packet features, packet information, and severity levels. Subsequently, the PCAP data was converted into a CSV file, and data preprocessing was carried out to prepare it for intrusion detection. By focusing on the Info feature, the target feature labels have been created whcih are Intrusive, Malicious and Benign. Machine learning models were selected based on previous research findings. Upon applying these selected models to the data, it has been discovered that the Random Forest model delivered the highest accuracy in detecting network-based intrusions, with an impressive 99.43% accuracy rate, surpassing any previously reported approaches.

7. REFERENCES

[1] N. Z. Joodaki, M. B. Dowlatshahi and M. Joodaki, "A novel ensemble feature selection method through Type I fuzzy," In 2022 9th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS) IEEE., pp. (pp. 1-6)., 2022.

[2] S. Kejriwal, D. Patadia, S. Dagli and P. Tawde, "Machine Learning Based Intrusion Detection," Fourth International Conference on Advances in Electronics, Computers and Communications (ICAECC), pp. 1-4, 2022.

[3] X. Yao and Z.-L. Li, "Distributed stealthy traffic anomaly detection based on wavelet packet analysis," 2009 International Conference on Apperceiving Computing and Intelligence Analysis, , pp. 203-206, 2009.

[4] C. Kaushik, T. Ram and C. Ritvik, "Network Security with Network Intrusion Detection System using Machine Learning Deployed in a Cloud Infrastructure," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 701-708, 2022.

[5] G. Fu, B. Li, Y. Yang and Q. Wei, "A Multi-Distance Ensemble and Feature Clustering Based Feature Selection Approach for Network Intrusion Detection.," In 2022 International Symposium on Sensing and Instrumentation in 5G and IoT Era (ISSI) IEEE., pp. (pp. 160-164)., 2022.

[6] P.-O. Brissaud, J. Francçis, I. Chrisment, T. Cholez and O. Bettan, "Transparent and service-agnostic monitoring of encrypted web traffic," IEEE Transactions on Network and Service Management, pp. pp.842-856, 2019.

[7] G. Yedukondalu, G. H. Bindu, J. Pavan, G. Venkatesh and A. SaiTeja, "Intrusion Detection System Framework Using Machine Learning," Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1-5, 2021.

[8] B. Yang and D. Liu, "Research on network traffic identification based on machine learning and deep packet inspection," IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. pp. 1887-1891, 2019.

[9] T. Zhang and S. Bao, "A Novel Deep Neural Network Model for Computer Network Intrusion Detection Considering Connection Efficiency of Network Systems," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 962-965, 2022.

[10] P. Xiang, C. Peng and Q. Li, "Hierarchical Association Features Learning for Network Traffic Recognition," International Conference on Information Processing and Network Provisioning (ICIPN, pp. pp. 129-133, 2022.

[11] N. D. Patel and B. M. Mehtre, "Detection of Intrusions using Support Vector Machines and Deep Neural Networks," 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1-5, 2022.

[12] Z. Xi, Y. Zhou, D. Zhang, K. Gao, C. Sun, J. Cao, Y. Wang, M. Xu and J. Wu, "Newton: Intent-driven network traffic monitoring," IEEE/ACM Transactions on Networking, 30(2),, pp. pp.939-952, 2021.

[13] S. Rosidin, G. F. Shidik, A. Z. Fanani and F. Al Zami, "Improvement with Chi Square Selection Feature using Supervised Machine Learning Approach on Covid-19 Data.," In 2021 International Seminar on Application for Technology of Information and Communication (ISemantic) IEEE., pp. (pp. 32-36)., 2021.

[14] A. E. Putra and L. K. Wardhani, "Chi-Square Feature Selection Effect On Naive Bayes Classifier Algorithm Performance For Sentiment Analysis Document.," In 2019 7th International Conference on Cyber and IT Service Management (CITSM) (Vol. 7,). IEEE., pp. pp. 1-7, 2019.

[15] Y. Zhai, W. Song, X. Liu, L. Liu and X. Zhao, "A chi-square statistics based feature selection method in text classification.," In 2018 IEEE 9th International conference on software engineering and service science (ICSESS). IEEE., pp. pp. 160-163, 2018.

[16] Z. Wang and X. Shu, "Feature Selection for Bayesian Inference Network in Radar Jamming Effect Analysis.," In 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP) . IEEE., pp. (pp. 620-623), 2021.

[17] M. Lackovic, "Performance analysis of packet switched all-optical networks," Proceedings of 2003 5th International Conference on Transparent Optical Networks, , pp. 174-179, 2003.

[18] B. Mikkelsen and S. Danielsen, "Analysis of a WDM packet switch with improved performance under bursty traffic conditions due to tuneable wavelength converters," Journal of Lightwave Technology, vol. 16, no. 5, pp. 729-735, 1998.

[19] A. Singhal, A. Maan, D. Chaudhary and D. Vishwakarma, "A Hybrid Machine Learning and Data Mining Based Approach to Network Intrusion Detection," International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 1-5, 2021.

[20] S. Neupane, J. Ables, W. Anderson and S. Mittal, "Explainable intrusion detection systems (x-ids)," A survey of current methods, challenges, and opportunities, pp. pp.112392-112415, 2022.

[21] D. Hongle and Z. Lin, "Research on SDN intrusion detection based on online ensemble learning algorithm," 2020 International Conference on Networking and Network Applications (NaNA), pp. 114-118, 2020.

[22] D. Selvamani and V. Selvi, "An efficacious intellectual framework for host based intrusion detection system," Procedia Computer Science, pp. 9-17, 2019.

[23] Y. Jieun, W. Lee and H. Jeong, "Poster Abstract: A Semi-Supervised Approach for Network Intrusion Detection Using Generative Adversarial Networks," IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 1-2, 2021.

8. AUTHOR'S PROFILE

Subhadeep Chakraborty, born in 1986, is the Researcher and Developer in Artificial Intelligence. He received the B.Tech degree from Saroj Mohan Institute of Technology, WBUT, India and M.Tech degree from Kalyani Govt. Engineering College, WBUT, India in Electronics and Communication Engineering in 2008 and 2010 respectively. The author has served several engineering colleges and cionducted several training for corporate and students on Signal Processing, Machine Learbning and AI. His primary research interest includes Digital Signal Processing, Robotics, Artificial intelligence, Cybersecurity etc.