IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Novel Machine Learning Approaches For Benign And Malicious Network Traffic

[1]Mr. Rakesh V.S., [2]Amrutha Varshini Challa, [3]Mamata G, [4]Precilla Mary B, [5]S D Shruthi

[1]Assisstant Professor,

[1]Cambridge Institute of Technology,

[1]Bengaluru, India

***Abstract:*** Distributed denial of service (DDoS) threats represents a significant cybersecurity challenge, constituting a variant of denial of service (DoS) in which IP addresses are exploited to launch attacks to a specific host or victim. DDoS attacks, characterized by meticulous coordination, exploit compromised secondary victims to target one or more victim systems, ranging from large-scale enterprise servers to less. These threats incur significant bandwidth and power costs, leading to the compromise of confidential data. Therefore, developing advanced algorithms to accurately detect various DDoS cyber threats, while considering computational load, has become urgent. The majority of the research currently in publication approaches DDoS threat detection as a binary classification problem, that is ascertaining whether or not an attack has started. However, to effectively protect the network and minimize significant damage, it's critical to distinguish the specific type of DDoS attack targeting the network or system. This study presents a comprehensive classifier that combines the strengths of the four best-performing algorithms. A comparative analysis is performed, comparing the Classifier with different artificial intelligence and machine learning (AI and ML) algorithms. Its goal is to improve the identification of various kinds of DDoS threats by transforming the problem into a multi-label classification scenario. Through this approach, the research aims to contribute to the refinement of cybersecurity strategies, ensuring a deeper understanding and proactive defence against various DDoS cyber threats.

***Index Terms*** - DDoS, cybersecurity, classification, multi-label, detection, attacks, algorithms, proactive defence, threat identification, network security.

## I. INTRODUCTION

Any malicious activity intended to interfere with the regular operation, availability, integrity, or functioning of computer networks or the data transferred across them is known as a network attack. These assaults can appear in a variety of ways, from straightforward ones like denial-of-service (DoS) assaults to more complex ones like data interception, virus propagation, and network vulnerability exploitation. Network attacks pose serious hazards to people, businesses, and entire systems to the internet. They can be executed by hackers, cybercriminals, or even hostile insiders. Network assaults can be carried out for a multitude of causes, such as espionage, money, political activism, or just general mayhem. Strong cybersecurity defences, such as intrusion detection systems, firewalls, encryption protocols, and frequent security updates to reduce vulnerabilities, are necessary to fend against network attacks.

Safeguarding sensitive data, avoiding unwanted access, preserving operational continuity, and adhering to legal obligations all depend on network protection. Organizations may reduce the risk of malware infections, data breaches, and other cyber threats by putting security measures like firewalls, encryption, access controls, antivirus software, and frequent security updates into place. In the eyes of clients, partners, and stakeholders, an organization's reputation and dependability are upheld by effective network security, which also protects the availability and integrity of data.

A crucial component of the architecture of cybersecurity is the intrusion detection system (IDS). It keeps an eye on network traffic and system events in order to spot any odd activity or potential dangers and alert administrators. By placing IDS at the network perimeter, on certain hosts, or inside designated segments, an organisation can strengthen its security posture and fend against cyberattacks, identity theft, and data breaches.

The below figure is an example of a network attack. The attacker overwhelms the victim with network traffic. Which leads to the server crashing down. It has more traffic than a typical Dos attack.

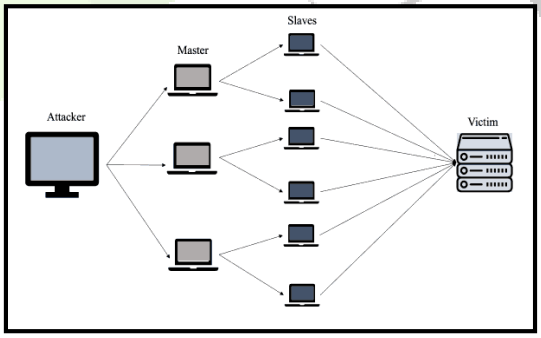


Fig1: DDos Attack

## II. RELATED WORK

The author of [1] proposed an unsupervised model that identifies network traffic patterns, enhancing security by detecting anomalies in real-time without labelled data.

The primary goal of [2] is detecting anomalies in software-defined networks to enhance security and dependability by tracking traffic patterns and evaluating forwarding choices for early threat identification.

The author of [3] proposed a system that detects intrusions by analyzing flow-based network data using the Inverse Potts Model, enhancing security through simulated flow relationships.

The output of [4] addresses the issue of preserving IoT user privacy by analyzing encrypted communication using symmetric cryptography and bolstering security without compromising sensitive data.

The author of [5] enhances cybersecurity situational awareness for intelligent power substations through multi-scale network data analysis, fortifying defences against infrastructure intrusions.

The author of [6] uses network traffic entropy and chaos analysis to promptly mitigate DDoS attacks by identifying chaotic patterns and improving network availability and dependability.

DAE-GAN with feature extraction enhances anomalous traffic detection by generating artificial samples and extracting discriminative features from network data as shown in [7].

The technique in [8] improves flow-based network attack detection accuracy through natural nearest neighbour classification and fuzzy entropy weighting, addressing traffic significance and uncertainty.

The author of [9] utilizes convolutional and recurrent neural networks to classify Internet of Things traffic, streamlining network administration and security through automated, precise classification.

The approach in [10] utilizes a GAN-based one-class classifier for proactive intrusion detection by analyzing traffic patterns and enhancing network security.

The research [11] enhances network traffic anomaly detection by combining feature extraction with DAE-GAN, improving accuracy through effective data augmentation for diverse patterns.

The research [12] enhances anomaly detection for proactive network security measures by analyzing temporally correlated traffic patterns and temporal relationships in network data.

The research [13] enhances volatility forecast precision by combining ARIMA and XGBoost methods, aiding informed financial decision-making.

The research [14] integrates CNNs and PCA to enhance anomaly detection in WSNs, boosting security resilience against cyberattacks.

In [15], the author examines and presents a cutting-edge method for flow-based intrusion detection using the Inverse Potts Model to provide timely and precise responses to security breaches.

The work of [16] supports urban planning and transportation decision-making by providing an accurate estimation of urban traffic density utilizing ultrahigh-resolution UAV video data and DNNs.

## III. DATASET AND METHODOLOGY

To distinguish between malicious and benign network traffic, this section describes the dataset and methods that were employed. ML methods are employed in the classification process, and the dataset utilized is CICDDOS2019

### *A. Description of dataset*

The CICDDoS2019 dataset is an extensive collection of network traffic information created to mimic both benign background traffic and actual DDoS attacks. Focusing on modern DDoS attacks against PortMap, NetBIOS, LDAP, MSSQL, UDP, and more, it includes PCAPs recording both benign and malicious activity. 25 users' behaviours were modelled using a B-Profile system to achieve realism over an array of protocols, including FTP, SSH, HTTP, HTTPS, and email. The dataset includes 12 DDoS assaults, including NTP, DNS, SYN, and WebDDoS, that were completed throughout the training phase and 7 during testing. Every day's data consists of event logs for each machine and raw network traffic. CICFlowMeter-V3 is used to extract over 80 traffic features, which are then recorded in CSV format. To create and assess detection and mitigation strategies, this dataset gives researchers access to important insights into DDoS assault patterns and benign network behaviour.

| Total number of rows | Benign | Malicious |
|---|---|---|
| 1953286 | 4097 | 1949189 |

Table 1:Total dataset

The above table shows the number of rows in the dataset considered. The dataset is a combination of SYN and UDP-Lag datasets with different attacks as shown below.

| Type of attack | No of row |
|---|---|
| SYN | 1582289 |
| UDP-lag | 366461 |
| WebDDoS | 439 |

Table 2: The Types of attack and number and rows

| Type of dataset | Benign | Malicious |
|---|---|---|
| Train set | 2868 | 1197328 |
| Test set | 1197 | 513173 |

Table 3:The data being used

After preprocessing the dataset, the dataset is split into a training set and a testing set. 70% of the data are a part of the training set. And 30% of the dataset is located in the training set post-cleaning and feature selection.

### *B. ML Approaches*

Random Forest is a machine learning approach that creates several decision trees during training and provides the average prediction for regression or the classes' mode for classification. Through the aggregate of individual tree projections, it decreases overfitting and increases accuracy.

$$h(x)=\sum i=1Nwi\cdot I(x\in Ri) \quad (1)$$

A machine learning approach called a decision tree divides the dataset recursively into subsets based on input features. At each node, it selects the feature and threshold for the best split to maximize purity. The resulting tree structure represents a series of if-else conditions for making predictions. Decision trees are interpretable but can overfit, so techniques like pruning are used. They are commonly employed in ensemble methods like Gradient Boosting Machines and Random Forests for enhanced performance.

Naive Bayes is a widely applied machine learning algorithm for classification tasks, such as text classification and spam filtering. It predicts the probability of a given data instance belonging to a particular class using probability theory. The algorithm makes an assumption of independence between features, which is why it is called "naive". Despite this simple assumption, Naive Bayes often performs surprisingly well in practice, especially with large datasets. It is computationally efficient, requires minimal training data, and is suitable for real-time applications and scenarios where computational resources are limited.

$$P(A|B) = P(B|A) * P(A) / P(B) \quad (2)$$

XGBoost is a popular and powerful ensemble learning algorithm in machine learning. It sequentially creates weak decision trees to correct the errors of its predecessors, optimizing an objective function that combines a loss function and a regularization term. XGBoost is widely used for its accuracy, efficiency, and versatility in both machine-learning competitions and real-world applications.

$$F(x) = \sum_{i=1}^{N} f(x) \quad (3)$$

LightGBM is an advanced gradient-boosting system that employs tree-based learning techniques. Its algorithm is founded on the Gradient Boosting Decision Tree (GBDT), much like CatBoost and XGBoost. Nonetheless, what sets LightGBM apart is its utilization of a leaf-wise approach, enabling it to manage sizable datasets and attain swifter training times.

$$Y=Base\_tree(X)-lr*Tree1(X)-lr*Tree2(X)-lr*Tree3(X) \quad (4)$$

### *C. Methodology*

Preprocessing the data is the initial step in the procedure. This entails employing label encoding to transform categorical data into numerical representation and standardising numerical properties through normal scaling. The most crucial attributes are then determined via feature selection using the ExtraTreeClassifier approach. This enhances the model's functionality and helps to decrease the number of dimensions.

The dataset is split into training and testing sets to train the machine learning models. Thirty percent of the dataset is set aside for testing, leaving the remaining seventy percent for training. Next, the training dataset is used to train five models: XGBoost, Random Forest, Decision Tree, Naïve Bayes, and LightGBM.

Important measures like Accuracy, Precision, Recall, and F1-Score are utilised to assess the trained models. This helps choose which model is best for the job at hand and offers insightful details on several facets of the models performance. Preparing the data is the first step in this process. This includes employing label encoding to translate categorical data into numerical format and standardising numerical properties with normal scaling. The ExtraTreeClassifier approach is used in the next stage to pick features, and lowering dimensions and improve the performance of the model by highlighting the most important qualities.

In order to instruct the machine learning models, the dataset is then divided into training and testing sets. 30% of the data is set aside for testing, and the remaining 70% is dedicated to training. On the training dataset, five models—Random Forest, Decision Tree, Naïve Bayes, XGBoost, and LightGBM—are trained.

Important measures like Accuracy, Recall, Precision, and F1-Score are utilised to assess the trained models. These metrics help identify the best model for the job at hand by offering insightful information about different facets of model performance.

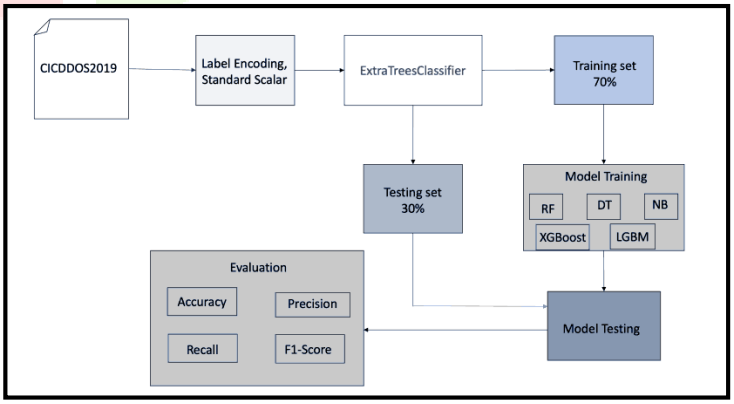


Fig 2: Methodology

*D. Evaluation Parameters*

Accuracy: Measures overall correctness by considering true positives and true negatives.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision: Quantifies the accuracy of positive predictions, minimizing false positives.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall (Sensitivity): Measures the ability to capture actual positive instances, minimizing false negatives.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1 Score: Harmonic mean of precision and recall, providing a balanced metric.

$$\text{F1 Score} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

Confusion Matrix: A summary table detailing true positives, true negatives, false positives, and false negatives to provide a detailed breakdown of model performance.

## IV. RESULTS AND DISCUSSIONS

LightGBM and XGBoost are two of the leading candidates in the assessment of algorithms for network traffic classification. XGBoost boasts an impressive accuracy rate of 99%, which indicates its effectiveness in classifying various network events. On the other hand, LightGBM is a strong contender, securing the second position out of the five algorithms tested and achieving a commendable accuracy rate of 98%.

It's interesting to note that LightGBM beats XGBoost in recall yet XGBoost is superior in total accuracy. This distinction draws attention to the distinct advantages of each algorithm. Even though XGBoost is quite accurate at classifying network traffic, LightGBM is better at catching more relevant instances and reducing false negatives. Finding all pertinent patterns is vital in network security and anomaly detection, where this characterstic is especially important.

| Algorithm | Accuracy Score | Precision | Recall | F1-score |
|---|---|---|---|---|
| Random Forest | 82.23% | 93% | 92% | 92% |
| Decision Tree | 82.02% | 78% | 54% | 64% |
| Naïve Bayes | 83.42% | 83% | 87% | 76% |
| XGBoost | 99.93% | 99% | 93% | 97% |
| LightGBM | 98.96% | 96% | 99% | 97% |

Table 4: The results obtained

The observed difference in performance between LightGBM and XGBoost highlights the significance of considering metrics other than accuracy. In real-world situations, striking a balance between precision and recall is vital, depending on the specific objectives and constraints of the application. LightGBM's superior recall can prove beneficial in scenarios where reducing missed detections is critical, while XGBoost might be more suitable in situations where precision is the primary concern.

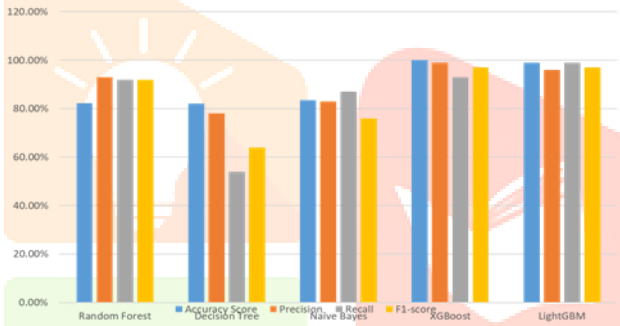


Fig 3: Results

Overall, the evaluation results highlight the necessity for a sophisticated approach to algorithm selection that considers various performance measures and matches them to the application's particular needs. Both LightGBM and XGBoost show remarkable potential in classifying network traffic, offering useful instruments to improve network administration and security.

## V. CONCLUSION

The top algorithms are XGBoost and LightGBM, which outperform Random Forest, Decision Tree, and Naïve Bayes in prediction, according to a comparative study of accuracy ratings. This implies that boosting technique-based ensemble approaches, such as XGBoost and LightGBM, are very good at managing large, complicated datasets and finding patterns that are important for precise prediction. They are suitable choices for various types of machine learning applications due to their resilience and capacity to manage non-linear interactions. XGBoost and LightGBM use ensemble learning techniques like boosting, which have proven to be quite effective in various applications. These algorithms capture complicated correlations in data by iteratively combining weak learners to generate strong learners. Particularly XGBoost has been quite well-liked because of its scalability, adaptability, and outstanding performance in machine learning contests.

Similar to this, LightGBM has proven remarkably accurate and effective, particularly when handling huge datasets, thanks to its unique gradient-based approach to tree splitting and leaf-wise growth strategy. On the other hand, although Random Forest and Decision Tree algorithms show reasonable accuracy, they could have trouble identifying complex relationships in the data. While Random Forest is an ensemble method, it is not as good at capturing subtle patterns as boosting techniques because it creates several decision trees with bagging and feature randomness. On the other hand, decision trees' overfitting limits their ability to make predictions, particularly when working with noisy or high-dimensional data.

Although Naïve Bayes is known for its simplicity and efficiency, it may not perform as well in some situations. This could be because of its oversimplified assumption of feature independence. While this assumption holds true for certain types of data, such as text categorization, Naïve Bayes may struggle with more complex datasets.

The findings indicate the significance of selecting the right algorithms based on the unique features and specifications of the dataset in order to accomplish the greatest possible prediction performance. While ensemble techniques like XGBoost and LightGBM are great at handling diverse datasets and identifying complex relationships, other algorithms such as Random Forest, Decision Tree, and Naïve Bayes can still be useful in certain situations. Therefore, it is essential to choose an algorithm with care and conduct thorough testing and evaluation to ensure accurate and reliable predictions in machine learning tasks.

## REFERENCES

[1] Ren-hung Hwang, "An Unsupervised Deep Learning Model for Early Network Traffic Anomaly Detection",2020,

[2] Camila F. T. Pontes "A New Method for Flow-Based Network Intrusion Detection Using the Inverse Potts Model",2022

[3] Peng Zhang, "Network-Wide Forwarding Anomaly Detection and Localization in Software Defined Networks,"2021,

[4] Dajiang Chen, "Privacy-Preserving Encrypted Traffic Inspection with Symmetric Cryptographic Techniques IoT",2022,

[5] Weijie Hao, "Multi-Scale Traffic Aware Cybersecurity Situational Awareness Online Model for Intelligent Power Substation Communication Network",2023,

[6] Xinlei Ma, "DDoS Detection Method Based on Chaos Analysis of Network Traffic Entropy ",2014

[7] Traffic Feature Extraction and DAE-GAN With Efficient Data Augmentation",2023

[8] Liangchen Chen introduces, "FEW- NNN: A Fuzzy Entropy Weighted Natural Nearest Neighbor Method For FlowBased Network Traffic Attack Detection",2020

[9] Manuel Lopez-Martin1 Belen Carro1, "Network Traffic Classifier With Convolutional and Recurrent Neural Networks for Internet of Things",2017

[10] Taehoon Kim, "Early Detection of Network Intrusions Using a GAN-Based One-Class Classifier",2022

[11] Bin Xiao, “Abnormal Traffic Detection: Traffic Feature Extraction and DAE-GAN With Efficient Data Augmentation”,2023

[12] Ido Nevat, “Anomaly Detection and Attribution in Networks With Temporally Correlated Traffic”,2022

[13] Yan Wang, “Forecasting Method of Stock Market Volatility in Time Series Data Based on Mixed Model of ARIMA and XGBoost”,2021

[14] Chengpeng yao, “This Traffic Anomaly Detection in Wireless Sensor Networks Based on Principal Component Analysis and Deep Convolution Neural Network”, 2022

[15] Camila F. T. Pontes, “ A New Method for Flow-Based Network Intrusion Detection Using the Inverse Potts Model”,2012

[16] Jiasong Zhu, “Urban Traffic Density Estimation Based on Ultrahigh-Resolution UAV Video and Deep Neural Network”,2022

[17] KHAN ZEB, “Anomaly Detection Based on LRD Behavior Analysis of Decomposed Control and Data Planes Network Traffic Using SOSS and FARIMA Models”,2017

[18] Peng Zhang, “Network-Wide Forwarding Anomaly Detection and Localization in Software Defined Networks”, 2021

[19] Yazhou Zhang, “A Multimodal Coupled Graph Attention Network for Joint Traffic Event Detection and Sentiment Classification”,2023

[20] ROTEM BAR, “SimCSE for Encrypted Traffic Detection and Zero-`qDay Attack Detection”,2022