Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 1: 2025
ISSN : **1906-9685**

REAL TIME DDoS DETECTION USING MACHINE LEARNING AND DYNAMIC TRAFFIC FLOWS

Mrs. N.J.N Varsha [1], G.Adyasree[2], G.Hemanth [3], K.Dinesh [4], Ch.Prameela [5] 1:Assistant Professor, 2,3,4,5, IV-B.Tech CSE StudentsDepartment of Computer Science and Engineering, Seshadri Rao Gudlavalleru Engineering College (An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada), Seshadri Rao Knowledge Village, Gudlavalleru-521356, Andhra Pradesh, India.

ABSTRACT
DDoS attacks are one of the most dangerous threats to network security. It may easily flood systems and hinder valid services, which will result in a huge amount of damage in terms of operations and money. The proposed paper includes a machine learning-based system, using real-time dynamic flow data for detection and classification of DDoS attacks. The CatBoost model was used for binary classification on whether there was an attack and multi-class classification on the type of one of the attacks-a SYN flood, UDP flood, and HTTP flood. This was attained on several train-test splits: 80-20, 70-30, and 60-40. In all these instances, the system did well, with its highest accuracy being 92.19%. This would enable real-time detection and classification, along with high-resolution measurement of time-based metrics, packet-level statistics, and behavioral indicators to tailor mitigation measures. The paper demonstrates the use of this method of machine learning in the case of network security by building a scalable and practical means of counteracting the increasingly prevalent DDoS attack threat.

I. INTRODUCTION
DDoS attacks are still a threat to businesses since they affect services and data integrity. Detection techniques have become ineffective in producing timely and accurate answers as threats continue to become more complex. This paper uses real-time dynamic network traffic flows and machine learning algorithms to identify and classify DDoS attacks.The proposed method transforms the multi-class classification challenge problem in network security by leveraging the CatBoost algorithm, which is known to be strong when working with tabular data and unbalanced datasets. In that respect, DDoS detection research will be related to other approaches such as machine learning-based, anomaly-based, and signature-based. Although promising, SVM, Random Forest, and XGBoost techniques suffer from performance and scalability issues on unbalanced datasets. Recent developments in gradient boosting methods have reduced the accuracy gap quite significantly while also highly improving the treatment of categorical features, especially with CatBoost. This work designs a robust DDoS detection system based on early developments.

II. LITERATURE REVIEW
Time-based characteristics obtained from network traffic flow statistics were shown to be useful in identifying and categorizing DDoS assaults by Halladay et al. (2022), who achieved over 99% accuracy in binary classification and over 70% in multiclass classification across 12 attack types. They stressed that using a 25-feature time-based subset is ideal for scalable and real-time DDoS detection systems because it not only saves training time and computational overhead but also maintains accuracy on par with full-feature datasets [1]. By mixing several decision trees and minimizing overfitting by bagging, Breiman (2001) presented Random Forests, an ensemble technique that

https://doi.org/10.36893/JNAO.2025.V16I01.014

increases classification and regression accuracy. The method aggregates different tree forecasts to improve generalization and is reliable for big, high-dimensional datasets. Additionally, Random Forests provide information about the significance of features, which makes them useful for machine learning applications that can be understood [2]. XGBoost is a scalable and efficient gradient boosting technique that Chen and Guestrin (2016) introduced to optimize decision tree-based models for tasks like DDoS attack detection. Its advanced features, which include regularization techniques (L1 and L2) and sparsity-aware algorithms, can effectively handle large and unbalanced datasets, which are common in network traffic analysis. Furthermore, XGBoost's skills in missing value handling and parallelized tree construction make it an outstanding tool for real-time DDoS detection and classification systems, as they result in faster training times and superior accuracy [3]. The gradient boosting framework CatBoost was created especially to manage categorical information effectively without requiring a lot of preparation. This makes it perfect for evaluating intricate network traffic data in DDoS detection jobs. In order to minimize overfitting and improve model stability, it integrates sophisticated approaches like ordered boosting and minimal variance sampling. This ensures dependable performance even with highly imbalanced datasets, which are frequently encountered in network security. For real-time DDoS detection and classification systems, CatBoost is a strong and effective option because it also offers quick training times, high accuracy, and reliable handling of real-world categorical data [4]. In order to create reliable detection systems, a thorough taxonomy of DDoS attacks and mitigation strategies provide a fundamental understanding of attack kinds, tactics, and their effects on network security. It draws attention to how DDoS attacks are changing, moving from volumetric floods to application-layer attacks, and stresses the necessity of flexible, real-time detection techniques. Furthermore, the classification of defense mechanisms into proactive and reactive tactics emphasizes the significance of utilizing machine learning approaches, like those employed in this project, to improve detection accuracy and successfully neutralize a variety of attack types [5]. A overview of DDoS attack methods in cloud computing investigates the weaknesses of cloud settings, where resources are dispersed and vulnerable to many kinds of attacks. It highlights how DDoS tactics take advantage of cloud-specific characteristics like auto-scaling and multi-tenancy by classifying them into volumetric, protocol, and application-layer attacks. The survey highlights how important it is to have real-time detection tools, including machine learning-based methods, in order to detect and stop DDoS attacks. This is in line with the project's goals of improving detection and classification in dynamic network environments [6]. This research examines the performance of machine learning and deep learning techniques for intrusion detection systems in detecting sophisticated threats such as DDoS attacks. sIt demonstrates how machine learning methods can handle massive network traffic and adjust to changing attack patterns, which makes them ideal for real-time detection systems. It also highlights the importance of deep learning in managing high-dimensional data and revealing hidden patterns, both of which are essential for boosting intrusion detection systems' scalability and accuracy and are in line with the project's objective of enhancing DDoS detection and categorization [7]. An investigation of machine learning for network intrusion detection looks at the difficulties and possibilities of using these methods in practical settings. It draws attention to the shortcomings of conventional strategies, including rule-based systems, in defending against complex and dynamic attacks like DDoS. According to the study, improving the efficacy and dependability of machine learning-based intrusion detection systems requires careful consideration of feature selection and model interpretability. These observations support the project's goal of using machine learning to detect and classify DDoS attacks in real time in a way that is precise and scalable [8].

III. METHODOLOGY

A. Dataset: CICDDoS2019

The CICDDoS2019 dataset is large-scale network traffic data, specifically labeled for use in researching and finding Distributed Denial of Service attacks. More than 80 features are included; among them, some important ones that are the characteristics in the network traffic like bytes and packets per flow, protocol type, source and destination IPs and ports, flow duration, packet size, and TCP flags. The dataset is diverse in terms of traffic patterns because of the mix of benign traffic and

https://doi.org/10.36893/JNAO.2025.V16I01.014

16 types of DDoS attacks, some of which are known attacks, such as the SYN Flood and UDP Flood, while others are less known: DNS Amplification and TFTP attacks. In addition, the dataset starts from more than one million rows per category of attack, hence ample volume for training or testing purposes. It is offered in two formats: CSV files, which give flow-based data appropriate for machine learning applications, and PCAP files, which contain raw packet-level data. This dataset provides a strong basis for identifying and describing DDoS assaults because to its extensive feature set and diversity, which make it perfect for binary and multi-class classification tasks.

B. Model Description

CatBoost model is a very strong machine learning approach specifically designed to deal effectively with tabular datasets with categorical features. This technique is applied when the DDoS attacks are found to be greatly underrepresented for certain types of attacks; CatBoost falls under the gradient boosting family and is excellent with imbalanced datasets. This model utilizes a dataset of 60 features. Some of the features include packet lengths, flow durations, source ports, and destination ports of the network characteristics. The network traffic flow recording depends on such properties in distinguishing between malicious and benign activity.

CatBoost natively handles categorical data, so one-hot encoding is not as intensive, thus simplifying preprocessing without losing the natural structure of the data. Three different data splits—80-20, 70-30, and 60-40—are used for training and testing to determine the consistency of the model's performance. During training, the model learns to identify network traffic patterns and characteristics that may indicate DDoS attacks. Its performance level has been optimized to achieve high accuracy, precision, recall, and F1-score in the recognition of even minority attack types. The predictability power of the trained model is well confirmed by very comprehensive testing. The best model is selected based on accuracy and other evaluation metrics for real-time implementation. After this process, this ideal model is used for predicting whether an assault has occurred and what kind of attack it is. Through CatBoost's features, the team has developed scalable and efficient detection for real-time DDoS across changing network conditionSs.

C. Implementation

Fig: Flowchart of Real-Time DDoS Detection and Classification System"

1. *Data Collection*:

The data source for this project is the CICDDoS2019 dataset. It is a big dataset specifically designed to detect DDoS-attacks and contains traffic records as well as benign traffic from 14 types of DDoS attacks that include SYN Flood, UDP Flood, and DNS Amplification among others. There were initially greater than 1 million rows for each type of attack, and more than 80 features capturing key characteristics of the network traffic: such as TCP flags, packet size, protocol type, and flow time.

2. *Data Preprocessing:*

We used the correlation-based feature selection method, maximizing the training dataset. All features with a correlation value above 0.8 were fetched using the formula (mentioned below) for the

correlation coefficient. It was left with only 60 highly linked features in the finalized dataset. Besides that, it trimmed the data to 430,000 rows so that there was efficient processing without jeopardizing the model performance.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

3. *Data Splitting*:

The whole stage was split into three configurations: 80-20, 70-30, and 60-40 for training and testing. It aided in checking the generalization and consistency of the model across different proportions of data.

4. *Training*:

This work uses the CatBoost algorithm that was designed with the purpose of addressing tabular data specifically with categorical features in gradient-boosted machines. Thus, the architecture employed in this work is that for

- Binary classification: whether there was a DDoS attack or not.
- Multi-Class Classification: Classification of attack types.

5. *Loss Function:*

The log-loss function is used as the target in training by CatBoost. This measures the goodness of fit of a classification model whose output is a probability between 0 and 1. The more that the decision trees are optimized so that their log-loss decreases, the more capable the model is of being correct in categorizing data. In this approach, the real labels are best approximated by the projected probabilities, thus maximizing classification performance.

6. *Hyperparameter Tuning*:

Some of the model's hyperparameters that are changed during training to increase accuracy and decrease overfitting include learning rate, depth, and iterations.

7. Model Evaluation:

It uses the precision, accuracy, recall, and F1-score measures as evaluation metrics for measuring the performance of the learned model on the test set. Finally, it selects the best model according to the configuration that produced maximum accuracy.

8. Attack Prediction:

It categorizes the risk into whether or not an attack will occur, and if an attack does occur, what type of attack it is. This model acts as the basis of real-time risk identification in networks.

9. Visualization:

With visualizations created through Matplotlib and Seaborn on confusion matrices and accuracy graphs, the model is analyzed for insights into performance and error.

IV. RESULTS

Over 500 epochs, the model was validated on 3,44,400 samples after being trained on 4,30,000 training samples. As the epochs went by, the training procedure steadily improved accuracy and loss measures. A brief synopsis of the findings is provided below:

Epochs	Learning Rate	Learning Time
0	0.6384	20.7s
100	0.7167	33m 22s
200	0.7264	1h 5m 50s
300	0.7304	1h 38m 51s
400	0.7328	2h 11m 42s
499	0.7343	2h 44m 25s

https://doi.org/10.36893/JNAO.2025.V16I01.014

With an accuracy of 92.19% and a learning rate of 0.7343 in the first epoch, these data demonstrate the model's learning curve and show a notable improvement in performance over epochs.

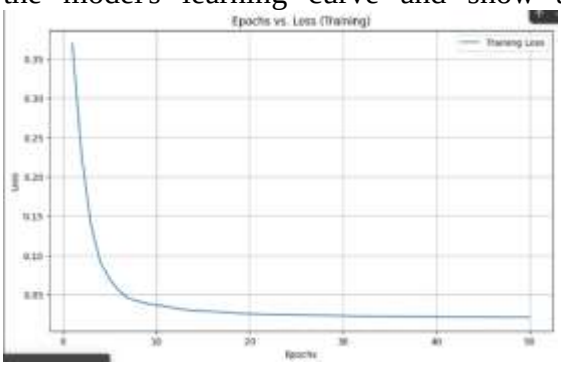


Fig 2. The graph shows the training loss over 500 epochs, illustrating a steady decrease in loss as the model improves its performance.

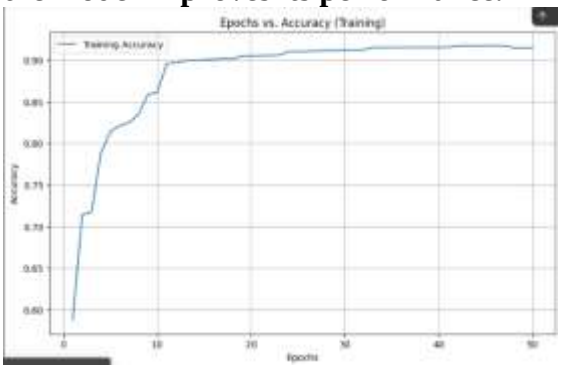


Fig 3. The graph shows the training accuracy over 500 epochs, demonstrating a steady increase in accuracy as the model learns and improves.

The graphs provided below illustrate the change in model performance over epochs. The loss of training remains steadily low on the Epochs vs. Loss graph, suggesting that the model is indeed decreasing errors as it learns. The loss quickly drops off during the early phases of training, which suggests much learning occurred over those epochs. The loss level stabilizes after a while with epochs, which suggests that the model is approaching convergence with little further error reduction.

The model improves the probability of correct predictions with training, and this is reflected in the Epochs vs. Accuracy graph, which also describes a general increase in training accuracy. The model attains the high train accuracy and stabilizes to converge on the training data when the accuracy curve dramatically shoots up through the first few epochs before gradually leveling off. When graphed together.

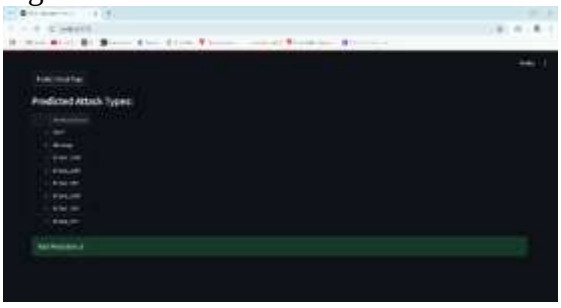
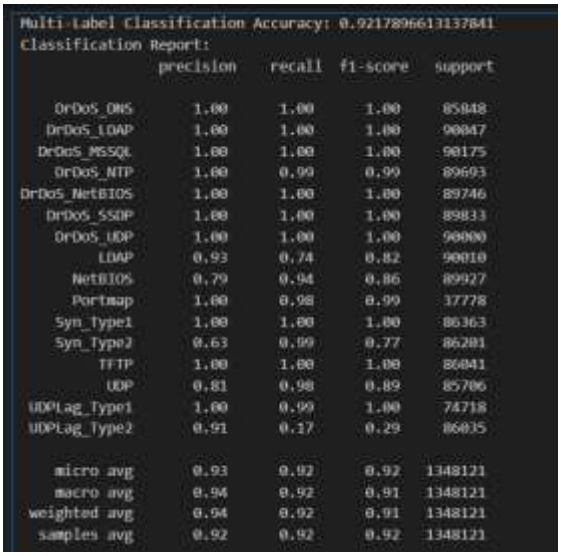


Fig 4. DDoS Attack Classification

The above image showcases the real-time DDoS Attack Detection System, implemented using CatBoost and deployed with Streamlit. The interface provides an intuitive way to predict and display various Distributed Denial-of-Service (DDoS) attack types based on network traffic data. By analyzing multiple test files, the system classifies incoming network traffic into specific attack categories such as DrDoS_LDAP, DrDoS_NTP, and Portmap. The prediction results are dynamically displayed in a structured format, ensuring clarity and accessibility. This system enhances network security by identifying malicious traffic patterns, enabling proactive defense mechanisms against cyber threats.

https://doi.org/10.36893/JNAO.2025.V16I01.014

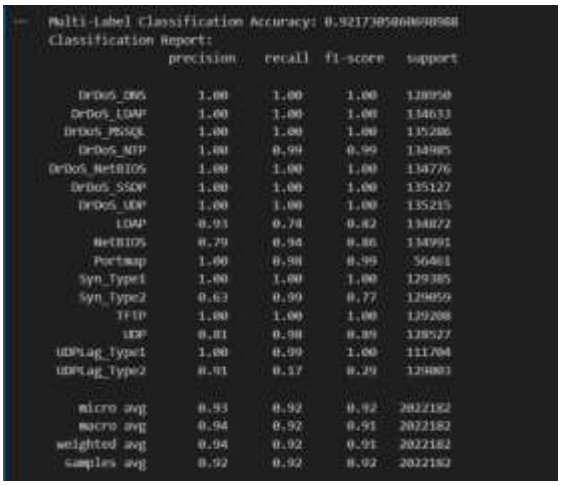
V. CONCLUSION



Multi-Label Classification Accuracy: 0.9217896613137841
Classification Report:

	precision	recall	f1-score	support
DrDoS_DNS	1.00	1.00	1.00	85848
DrDoS_LDAP	1.00	1.00	1.00	90047
DrDoS_MSSQL	1.00	1.00	1.00	90175
DrDoS_NTP	1.00	0.99	0.99	89693
DrDoS_NetBIOS	1.00	1.00	1.00	89746
DrDoS_SSDP	1.00	1.00	1.00	89833
DrDoS_UDP	1.00	1.00	1.00	90000
LDAP	0.93	0.74	0.82	90010
NetBIOS	0.79	0.94	0.86	89927
Portmap	1.00	0.98	0.99	37778
Syn_Type1	1.00	1.00	1.00	86363
Syn_Type2	0.63	0.99	0.77	86201
TFTP	1.00	1.00	1.00	86041
UDP	0.81	0.98	0.89	85706
UDPLag_Type1	1.00	0.99	1.00	74718
UDPLag_Type2	0.91	0.17	0.29	86035
micro avg	0.93	0.92	0.92	1348121
macro avg	0.94	0.92	0.91	1348121
weighted avg	0.94	0.92	0.91	1348121
samples avg	0.92	0.92	0.92	1348121

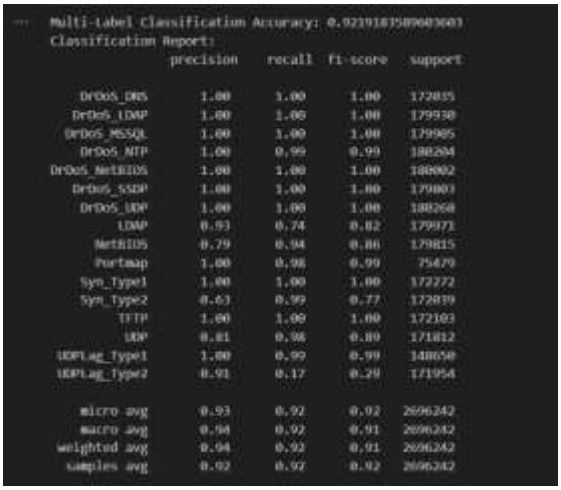
Fig 1. Accuracy and Classification Report of CatBoost model with 80% train and 20% test datasets



Multi-Label Classification Accuracy: 0.9217305860640968
Classification Report:

	precision	recall	f1-score	support
DrDoS_DNS	1.00	1.00	1.00	128950
DrDoS_LDAP	1.00	1.00	1.00	134633
DrDoS_MSSQL	1.00	1.00	1.00	135206
DrDoS_NTP	1.00	0.99	0.99	134985
DrDoS_NetBIOS	1.00	1.00	1.00	134776
DrDoS_SSDP	1.00	1.00	1.00	135127
DrDoS_UDP	1.00	1.00	1.00	135215
LDAP	0.93	0.74	0.82	134872
NetBIOS	0.79	0.94	0.86	134991
Portmap	1.00	0.98	0.99	56681
Syn_Type1	1.00	1.00	1.00	129385
Syn_Type2	0.63	0.99	0.77	129059
TFTP	1.00	1.00	1.00	129208
UDP	0.81	0.98	0.89	128527
UDPLag_Type1	1.00	0.99	1.00	111704
UDPLag_Type2	0.91	0.17	0.29	129003
micro avg	0.93	0.92	0.92	2022182
macro avg	0.94	0.92	0.91	2022182
weighted avg	0.94	0.92	0.91	2022182
samples avg	0.92	0.92	0.92	2022182

Fig 2. Accuracy and Classification Report of CatBoost model with 70% train and 30% test datasets



Multi-Label Classification Accuracy: 0.9219183589603603
Classification Report:

	precision	recall	f1-score	support
DrDoS_DNS	1.00	1.00	1.00	172035
DrDoS_LDAP	1.00	1.00	1.00	179930
DrDoS_MSSQL	1.00	1.00	1.00	179985
DrDoS_NTP	1.00	0.99	0.99	180204
DrDoS_NetBIOS	1.00	1.00	1.00	180002
DrDoS_SSDP	1.00	1.00	1.00	179803
DrDoS_UDP	1.00	1.00	1.00	180268
LDAP	0.93	0.74	0.82	179971
NetBIOS	0.79	0.94	0.86	179815
Portmap	1.00	0.98	0.99	75479
Syn_Type1	1.00	1.00	1.00	172272
Syn_Type2	0.63	0.99	0.77	172039
TFTP	1.00	1.00	1.00	172103
UDP	0.81	0.98	0.89	171012
UDPLag_Type1	1.00	0.99	0.99	148850
UDPLag_Type2	0.91	0.17	0.29	171954
micro avg	0.93	0.92	0.92	2696242
macro avg	0.94	0.92	0.91	2696242
weighted avg	0.94	0.92	0.91	2696242
samples avg	0.92	0.92	0.92	2696242

Fig 3. Accuracy and Classification Report of CatBoost model with 60% train and 40% test datasets

There were three different training and testing splits applied to train the dataset using the CatBoost model, which are 80%-20%, 70%-30%, and 60%-40%. The accuracy for the 60%-40% split was a little more at 92.19%, but the accuracy of 80%-20% and 70%-30% divisions was the same at 92.17%.

https://doi.org/10.36893/JNAO.2025.V16I01.014

The strength of the CatBoost model to handle massive datasets in multi-class classification tasks is seen by the stability of performance in the different splits. The 60%-40% split was the one that performed best among the three splits, hence showing its potential to generalize to new data properly. The outcomes of the project show how effective the CatBoost algorithm is in real-time DDoS detection and classification. The stability in gathering data from the model is shown by low variation in accuracy between splits.

VI. REFERENCES

[1]. J. Halladay, D. Cullen, N. Briner, J. Warren, K. Fye, R. Basnet, J. Bergen, & T. Doleck, (2022). Detection and Characterization of DDoS Attacks Using Time-Based Features. IEEE Access, 10, 49794–49807. https://doi.org/10.1109/ACCESS.2022.3173319.
[2]. L. Breiman (2001). Random Forests. Machine Learning. DOI:10.1145/2939672.2939785.
[3]. T. Chen, & C. Guestrin, (2016). XGBoost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). DOI:10.1145/2939672.2939785.
[4]. A.V. Dorogush, V. Ershov & A. Gulin, (2018). CatBoost: Gradient boosting with categorical features support. In Advances in Neural Information Processing Systems (NeurIPS).
[5]. J. Mirkovic & P. Reiher, (2004). A taxonomy of DDoS attack and DDoS defense mechanisms. ACM SIGCOMM Computer Communication Review. DOI: 10.1145/997150.997156.
[6]. K. Rohit & V. Tripathi, (2016). A survey on DDoS attack techniques in cloud computing. IEEE International Conference on Computing, Communication, and Automation (ICCCA).
DOI: 10.1109/CCAA.2016.7813728.
[7]. H. Liu & B. Lang, (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. Applied Sciences.
DOI: 10.3390/app9114396.
[8]. R. Sommer & V. Paxson, (2010). Outside the closed world: On using machine learning for network intrusion detection. In Proceedings of the IEEE Symposium on Security and Privacy.
DOI: 10.1109/SP.2010.25.