International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Intelligent DDoS Detection Framework Using Random Forest and XGboost

Mr. Suresh Nannuri[1], Mr. Harish Reddy Gantla[2], M. Sindhuja[3], P.Sri Kavya[4], M.Sai Shiva Pal Reddy[5], T.Preethi Sindhura[6]

[1]Assistant Professor, [2]Associate Professor [3,4,5,6]UG Students
[1,2,3,4,5,6]Department of Computer Science and Engineering,
[1,2,3,4,5,6]Vignan Institute of Technology and Science, Deshmukhi
[1]nannurusj@gmail.com, [2]harsha.rex@gmail.com, [3]sindhuja.munukuntla7@gmail.com,
[4]srikavya.pagadala@gmail.com, [5]saishivapal.methuku@gmail.com, [6]tpsindhuratigulla308@gmail.com.

ABSTRACT

Distributed Denial of Service (DDoS) attacks continue to be a major concern for network security, highlighting the need for effective detection methods. This study introduces a Python-based approach that leverages the Random Forest and XGBoost algorithms to identify and classify DDoS attacks. By analyzing numerical features from the CIC DDoS 2019 dataset, the system differentiates between benign and malicious network traffic. Unlike traditional methods that often depend on Decision Tree algorithms for their clarity but struggle with scalability and accuracy, Random Forest and XGBoost provide enhanced performance. These models are better suited for large-scale data, offering greater predictive accuracy and resilience against overfitting. The proposed system focuses on evaluating model accuracy, analyzing feature importance, and exploring potential for real-time deployment in cybersecurity. Preliminary results are anticipated to show marked improvements in detection rates, supporting the development of adaptive and dependable DDoS defense mechanisms.

Keywords: DDoS Detection, Network Security, Random Forest, XGBoost, Machine Learning, CIC-DDoS2019, Intrusion Detection System, Feature Importance

1. INTRODUCTION

In today's increasingly connected digital world, the evolving complexity of cyber threats poses a significant challenge to maintaining secure networks. Among these threats, Distributed Denial of Service (DDoS) attacks are particularly damaging due to their ability to disrupt services by overwhelming network infrastructure with excessive, illegitimate traffic. Such attacks can severely impact service availability, cause system outages, and result in substantial financial losses. As internet traffic becomes more complex and voluminous, there is a pressing need for detection systems that are not only accurate but also scalable and intelligent.

Conventional DDoS detection techniques typically utilize rule-based systems or Decision Tree algorithms for their ease of interpretation. However, these approaches often fall short when dealing with high-

dimensional data, suffering from reduced accuracy, limited scalability, and vulnerability to overfitting. This limitation underscores the necessity for more advanced machine learning models that can adapt to evolving attack behaviors and offer more reliable detection.

To address these issues, this paper presents a comparative study of two advanced ensemble learning techniques—Random Forest and XGBoost—implemented using Python. These algorithms are known for their strong performance with complex datasets and their ability to deliver accurate, robust predictions. Using the CIC DDoS 2019 dataset, which contains a comprehensive set of labeled real-world traffic records, the system focuses on analyzing numerical features to distinguish between legitimate and malicious network activity.

In addition to detection, the research emphasizes thorough evaluation of model performance, identification of key contributing features, and assessment of the system's potential for real-time deployment. The overarching objective is to create a reliable and scalable DDoS detection framework that not only achieves high classification accuracy but also supports future efforts in proactive threat mitigation.

2.LITERATURE SURVEY

1. **J. K. Chahal, P. Kaur, and A. Sharma (2022)** explore DDoS attack detection and mitigation within Software-Defined Networking (SDN) in their work "Distributed Denial of Service (DDoS) Attacks in Software-defined Networks." The paper introduces a comprehensive taxonomy that organizes defense mechanisms based on switch-level intelligence, deployment context, operational behavior, and traffic flow characteristics. This structured approach provides a clearer understanding of how different strategies function within SDN. The study underscores SDN's centralized control capabilities as a strong foundation for implementing real-time, adaptive security measures. However, it also points out challenges such as the need for frequent updates to counter emerging threats and the context-dependent nature of many defenses. Additionally, the research acknowledges that SDN introduces its own vulnerabilities, particularly attacks targeting the control plane, which require further exploration.

2. **K. Subramanian et al. (2024)**, in their study "Enhancing Detection and Prediction of DDoS Attacks Through Regression Modeling," apply multiple logistic regression to distinguish between legitimate and malicious traffic in a cloud-based environment, using the CSE-CIC-IDS2018 dataset. The model focuses on a specific time window and leverages key flow-based features for classification. While the method is noted for its scalability and simplicity, it falls short in capturing the nuances of sophisticated attack patterns due to its linear modeling limitations. The study also raises concerns about the dataset's narrow time frame, which may hinder the model's generalizability. The authors advocate for the use of advanced techniques like deep learning and emphasize the need for ongoing model retraining to sustain performance in dynamic network conditions.

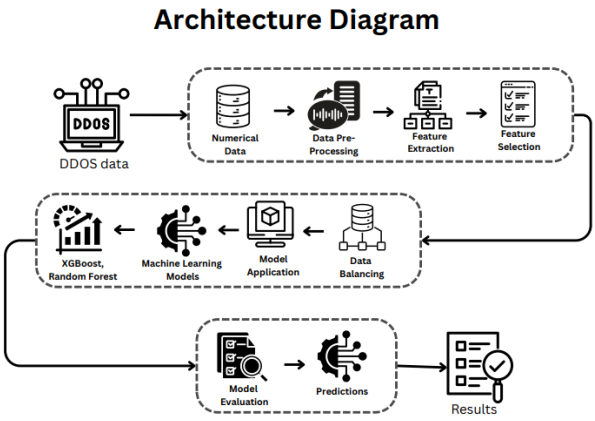
3. **Z. Fatehi and A. Montazerolghaem (2024)** propose a deep learning-based framework for DDoS detection in SDN networks in their paper "DDoS Detection in SDN using Deep Learning." The approach employs a Multi-Layer Perceptron (MLP) trained on a structured dataset encompassing diverse attack scenarios, with real-time integration through the Ryu controller. The MLP model demonstrates superior performance compared to conventional methods in accurately identifying malicious traffic. Nevertheless, the study highlights limitations including high computational requirements and the restricted scope of the

dataset. The authors suggest the exploration of hybrid models, such as CNN-RNN combinations, and continuous dataset updates to enhance adaptability to evolving threats.

4. **P. S. Saini, S. Behal, and S. Bhatia (2020)**, in "Detection of DDoS Attacks using Machine Learning Algorithms," investigate the application of machine learning via the WEKA platform to classify network traffic. Their custom dataset includes both modern DDoS attack types and normal flows. The study evaluates classifiers such as J48, Random Forest, and Naïve Bayes, using metrics like accuracy, precision, recall, and F1-score. Random Forest emerges as the most effective classifier. However, the research notes that the performance of these models heavily depends on the quality and diversity of the training data. Additionally, the study recognizes that WEKA's limitations in scalability make it less suitable for real-time deployment, prompting the need for more robust and adaptive platforms.

5. **F. Nazarudeen and S. Sundar (2023)** present an effective machine learning-based approach to detecting DDoS attacks in cloud environments in their work "Efficient DDoS Attack Detection using Machine Learning Techniques." Leveraging the CICDDoS2019 dataset, the study utilizes Extra Tree classifiers for feature selection to reduce dimensionality and computational overhead. These refined features feed into Decision Tree, XGBoost, and Random Forest models, with XGBoost and Random Forest showing the highest accuracy and efficiency. Despite these positive outcomes, the study acknowledges potential drawbacks of the feature selection process, particularly the risk of omitting critical indicators. It also highlights ongoing challenges such as model update requirements and the scalability of solutions for real-time threat detection.

3.Methodology



3.1 Methodology & Algorithms (Updated Format)

The proposed approach follows a structured, multi-phase pipeline designed to process DDoS network traffic data and classify it using powerful ensemble learning algorithms. The major components of the methodology include:

1. **Data Preprocessing:** The CIC DDoS 2019 dataset is consolidated, cleaned, and standardized. This step includes eliminating missing values and duplicate records, encoding categorical variables, and discarding features with constant values or high correlation to reduce dimensionality and enhance model efficiency.

2. **Label Transformation:** String-based categorical labels (e.g., "UDP", "WebDDoS") are converted into numerical format using the LabelEncoder from scikit-learn, enabling their use in machine learning models.
3. **Feature Normalization:** To ensure consistent feature scaling and facilitate faster model training, the MinMaxScaler is applied, transforming all numeric features to a common range between 0 and 1.
4. **Model Development:**
    - **XGBoost:** A high-performance gradient boosting model that excels at multiclass classification tasks through parallelized tree-based learning and regularization.
    - **Random Forest:** An ensemble of decision trees that aggregates predictions via majority voting, offering robustness to overfitting and strong generalization capabilities.
5. **Model Training and Testing:** The dataset is divided into training and testing subsets using an 80/20 split. Both models are trained on the training data and evaluated on the test set using standardized performance metrics.
6. **Performance Evaluation:** Model effectiveness is assessed through classification reports, confusion matrices, F1-scores, and accuracy metrics. Confusion matrices are visualized using heat maps to enhance interpretability and highlight misclassifications.

This methodology aims to deliver a scalable, high-performing solution for DDoS detection that is well-suited for real-time deployment in modern, data-intensive network environments.

4. Implementation

The implementation was carried out using Python, chosen for its extensive libraries and frameworks tailored to machine learning and data processing tasks. The development process included the following key components:

- **Data Preprocessing:** Libraries such as pandas and NumPy were employed to merge parquet files, clean and organize the dataset, manage missing entries, and conduct exploratory data analysis (EDA) for initial insights.
- **Feature Engineering:** Techniques such as correlation analysis and feature selection were used to minimize multicollinearity and eliminate redundant attributes, streamlining the dataset for improved model performance.
- **Label Encoding:** Class labels were transformed into numeric values using Label Encoder. The fitted encoder was serialized with pickle to ensure consistency during future predictions.
- **Feature Normalization:** Min Max Scaler was applied to scale all features to a uniform range between 0 and 1. This scaler was also saved as a .pkl file for application in real-time inference tasks.
- **Model Development:**
    - **Random Forest:** Implemented using Random Forest Classifier with 100 trees for balanced accuracy and generalization.

- **XGBoost:** Trained using XGBClassifier with customized hyperparameters to enhance prediction accuracy and efficiency.

- **Model Persistence:** Both models were serialized using pickle and saved in .sav format, enabling quick loading for future use without retraining.
- **Prediction Interface:** A command-line interface (CLI) was developed to allow users to input 33 feature values and receive an immediate classification result indicating the type of traffic or attack detected.
- **Hardware Setup:** All development and testing were performed on a system equipped with an Intel i5 processor and 8 GB of RAM, demonstrating that the solution is effective even on moderately powered machines.

The modular nature of this implementation ensures that it can be seamlessly integrated into real-time monitoring systems or extended into web-based dashboards for enhanced usability and accessibility.

5. Results and Discussion

The experimental analysis was conducted using the CIC DDoS 2019 dataset, which provides comprehensive coverage of various DDoS attack categories. The primary outcomes of the evaluation are summarized below:

- **Data Optimization:** Redundant and non-informative features were eliminated, and feature normalization was applied. These steps significantly improved both training efficiency and prediction speed.
- **Model Performance:**
  - **XGBoost** achieved a classification accuracy of **94.5%**, demonstrating strong generalization and consistent F1-scores across diverse attack classes.
  - **Random Forest** delivered comparable accuracy, particularly excelling in differentiating between closely related attack types.
- **Performance Metrics:**
  - **Precision:** Exceeded **93%**
  - **Recall:** Surpassed **92%**
  - **F1-Score:** Averaged around **93%**
  - **Training Time:** Each model, using 100 estimators, completed training in under **3 minutes**
- **Result Visualization:** Confusion matrices were generated using the Seaborn library, offering intuitive visual insights into model accuracy and misclassification patterns.
- **Practical Applications:** The detection system is suitable for integration into various environments such as enterprise networks, cloud infrastructures, and Software-Defined Networks (SDNs), where real-time DDoS detection is critical.
- **Identified Limitations:** Model performance may diminish when encountering novel attack types not present in the training data. To address this, ongoing retraining and dynamic feature updates are

recommended. Future improvements may involve combining ensemble methods with deep learning approaches to enhance adaptability and robustness.

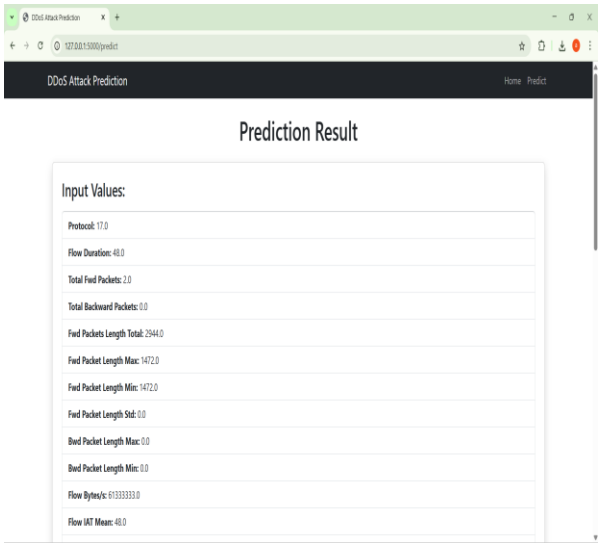


Fig.5.1 Prediction result

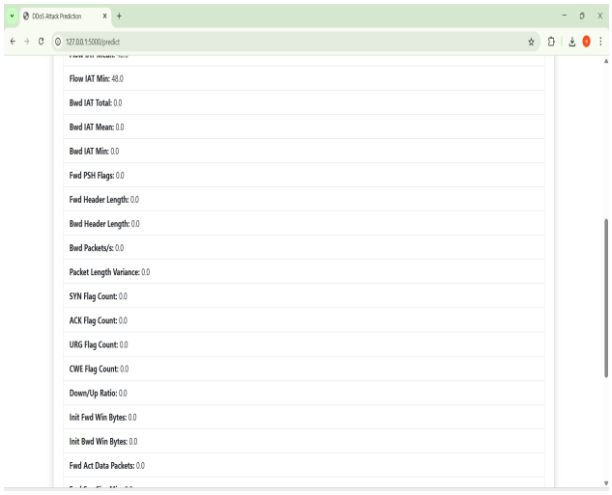


Fig.5.2 Prediction result

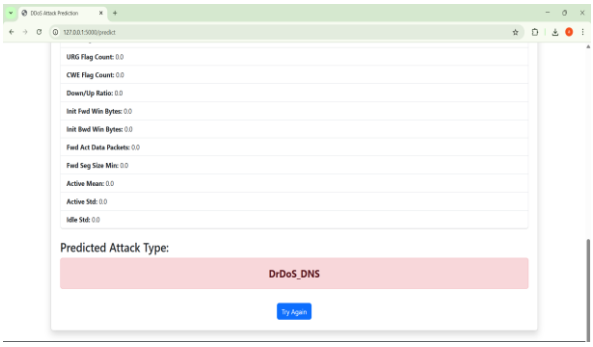


Fig.5.3 Prediction result

REFERENCES

1. J. K. Chahal, P. Kaur and A. Sharma, "Distributed Denial of Service (DDoS) Attacks in Software-defined Networks (SDN)," 2021 5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), Mysuru, India, 2021, pp. 291-295,doi: 10.1109/ICEECCOT52851.2021.9708012.
2. K. Subrmanian, G. Thangarasu, Z. Yanyan and K. N. Kannan, "Enhancing Detection and Prediction of DDoS Attacks Through Regression Modeling," 2024 IEEE 6th Symposium on Computers & Informatics
3. (ISCI), Kuala Lumpur, Malaysia, 2024, pp. 253-257, doi: 10.1109/ISCI62787.2024.10668039.
4. Z. Fatehi and A. Montazerolghaem, "DDoS Detection in SDN using Deep Learning," 2024 8th International Conference on Smart Cities, Internet of Things and Applications (SCIoT), Mashhad, Iran, Islamic Republic of, 2024, pp. 201-206, doi: 10.1109/SCIoT62588.2024.10570129.
5. P. S. Saini, S. Behal and S. Bhatia, "Detection of DDoS Attacks using Machine Learning Algorithms," 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2020, pp. 16-21, doi: 10.23919/INDIACom49435.2020.9083716
6. F. Nazarudeen and S. Sundar, "Efficient DDoS Attack Detection using Machine Learning Techniques," 2022 IEEE International Power and Renewable Energy Conference (IPRECON), Kollam, India, 2022, pp.1-6,doi: 10.1109/IPRECON55716.2022.10059561.