**DETECTION OF DDOS ATTACKS USING RANDOM FOREST AND XGBOOST**

**Abstract**

Distributed Denial of Service (DDoS) attacks pose significant challenges to network security, necessitating efficient detection mechanisms. This project proposes a Python-based implementation that utilizes Random Forest and XGBoost algorithms to classify and detect DDoS attacks. The system is designed to analyze numerical features from the CIC DDoS 2019 dataset, distinguishing between normal and malicious traffic. While existing systems rely on Decision Tree algorithms for interpretability, they often fall short in scalability and predictive performance. Random Forest and XGBoost address these limitations by offering superior accuracy, robustness against overfitting, and the ability to handle large datasets efficiently. This proposal emphasizes model performance evaluation, feature importance analysis, and real-time applicability for network security. Experimental results are expected to demonstrate significant improvements in detection accuracy, providing a foundation for adaptive and reliable DDoS mitigation strategies.

**1. INTRODUCTION**

In today’s interconnected digital landscape, the growing sophistication of cyber threats presents a serious concern for network security. Among these, Distributed Denial of Service (DDoS) attacks stand out due to their disruptive nature and widespread impact on online services. By overwhelming network resources with illegitimate traffic, DDoS attacks can cripple systems, interrupt services, and lead to significant financial losses. As internet traffic continues to increase in volume and complexity, the need for efficient, scalable, and intelligent detection mechanisms has become more critical than ever.

Traditional DDoS detection systems often rely on rule-based approaches or Decision Tree algorithms for their interpretability. However, these methods are frequently limited by their lower accuracy, susceptibility to overfitting, and inability to scale effectively with high-dimensional data. This creates a gap in the development of reliable defense systems that can adapt to modern attack patterns. Therefore, it is essential to explore more advanced machine learning techniques that can improve the precision and resilience of DDoS detection.

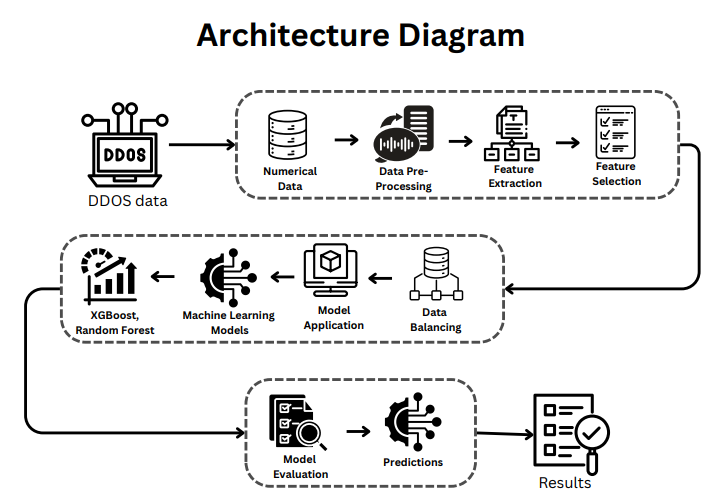
This project aims to address these challenges by implementing and comparing two powerful ensemble learning models—Random Forest and XGBoost—using Python. Both algorithms are well-regarded for their ability to handle complex datasets and provide high predictive performance. Utilizing the CIC DDoS 2019 dataset, which offers a rich collection of real-world traffic data, the proposed system will analyze numerical features to classify network activity as either normal or malicious.

Beyond detection, the project also focuses on evaluating model performance, analyzing feature importance, and exploring real-time applicability in network environments. The ultimate goal is to develop a robust, scalable detection system that not only identifies DDoS attacks with high accuracy but also lays the groundwork for proactive mitigation strategies.

**2.LITERATURE REVIEW / SURVEY**

1. Distributed Denial of Service (DDoS) Attacks in Software-defined Networks (SDN) by J. K. Chahal, P. Kaur, and A. Sharma (2022) presents a taxonomy of DDoS detection and mitigation strategies tailored for SDN environments. The study categorizes defense techniques based on switch intelligence, deployment location, defense activity, and flow characteristics, offering a structured framework for understanding DDoS countermeasures. It highlights the potential of SDN's centralized control plane to implement real-time, intelligent security mechanisms. However, it acknowledges limitations such as the need for continuous updates to address evolving threats and the contextual dependence of defense effectiveness on specific network conditions. The paper also notes that SDN itself introduces new vulnerabilities, like controller-targeted attacks, which require further research.
2. Enhancing Detection and Prediction of DDoS Attacks Through Regression Modeling by K. Subrmanian et al. (2024) explores the use of multiple logistic regression to classify normal and malicious traffic within a cloud computing environment using the CSE-CIC-IDS2018 dataset. The study focuses on traffic from a specific time frame to train a model using key flow-based features for effective DDoS detection. While the approach is efficient and scalable, its ability to capture complex attack behaviors is limited due to the linear nature of the model. The restricted temporal scope of the dataset also raises concerns about generalizability across broader network conditions. The authors suggest that advanced methods like deep learning could better handle evolving attack patterns and recommend frequent retraining to maintain model accuracy.
3. DDoS Detection in SDN using Deep Learning by Z. Fatehi and A. Montazerolghaem (2024) proposes a deep learning model using a Multi-Layer Perceptron (MLP) to detect DDoS attacks in SDN networks. The model is trained on a carefully structured dataset that includes varied attack scenarios and is integrated with the Ryu controller to enable real-time detection. The MLP outperforms traditional methods in identifying malicious traffic, offering an effective and efficient solution for SDN security. Nonetheless, the authors recognize limitations in terms of dataset coverage and the high computational demands of deep learning models. They recommend future research into hybrid deep learning models (e.g., CNN-RNN) and the continuous updating of datasets to enhance adaptability to new threats.
4. Detection of DDoS Attacks using Machine Learning Algorithms by P. S. Saini, S. Behal, and S. Bhatia (2020) utilizes WEKA-based machine learning techniques to classify network traffic, using a newly developed dataset that includes both modern DDoS attack types and normal flows. Classifiers like J48, Random Forest, and Naïve Bayes are tested for their detection effectiveness, with performance assessed through accuracy, precision, recall, and F1-score. The study finds Random Forest to be particularly effective, though it emphasizes that model performance is highly contingent on the quality and diversity of training data. Furthermore, the WEKA platform, while useful for evaluation, may not scale effectively for real-time implementations, necessitating the exploration of more scalable and adaptive methods.
5. Efficient DDoS Attack Detection using Machine Learning Techniques by F. Nazarudeen and S. Sundar (2023) presents a machine learning approach for detecting DDoS attacks in cloud environments using the CICDDoS2019 dataset. Feature selection via the Extra Tree classifier helps optimize the model by reducing feature space and computational load. The selected features are used to train Decision Tree, XGBoost, and Random Forest models, all evaluated using standard performance metrics. The study reports high detection accuracy and efficiency, with XGBoost and Random Forest performing notably well. However, it also notes the risks of excluding important features during selection and the need for continuous model updates due to the dynamic nature of attack strategies. Scalability and latency issues in real-time deployment are identified as areas for further optimization.

**3.Methodology**



**3.1 Methodology & Algorithms (Updated Format)**

The proposed methodology involves a multi-stage pipeline that begins with preprocessing DDoS network traffic data and ends with classification using ensemble machine learning models. The key steps are:

1. Data Preprocessing: The CIC DDoS 2019 dataset is merged, cleaned, and standardized. Null values and duplicate entries are removed. Non-numeric features are encoded, and highly correlated or single-valued features are dropped to reduce dimensionality.
2. Label Encoding: The LabelEncoder from scikit-learn is used to transform string labels (e.g., "UDP", "WebDDoS") into numeric values for classification.
3. Feature Scaling: MinMaxScaler is applied to scale features between 0 and 1, ensuring uniformity and improving model convergence.
4. Model Construction:
   * XGBoost Model: A gradient boosting model optimized for speed and performance. It handles multiclass classification effectively using parallel tree boosting.
   * Random Forest Model: An ensemble model using multiple decision trees. It reduces overfitting and increases accuracy through majority voting.
5. Training and Validation: The dataset is split into training and test sets (80/20 split). Both models are trained on the training set and evaluated on the test set using appropriate performance metrics.
6. Evaluation: Classification reports, confusion matrices, F1-scores, and accuracy metrics are computed for both models. Heatmaps are used to visualize confusion matrices for better interpretability.

This methodology provides a scalable and robust solution for real-time DDoS detection in modern network environments.

**4. Implementation**

The implementation phase utilized Python as the core programming language due to its robust ecosystem for machine learning and data processing. The following components were developed:

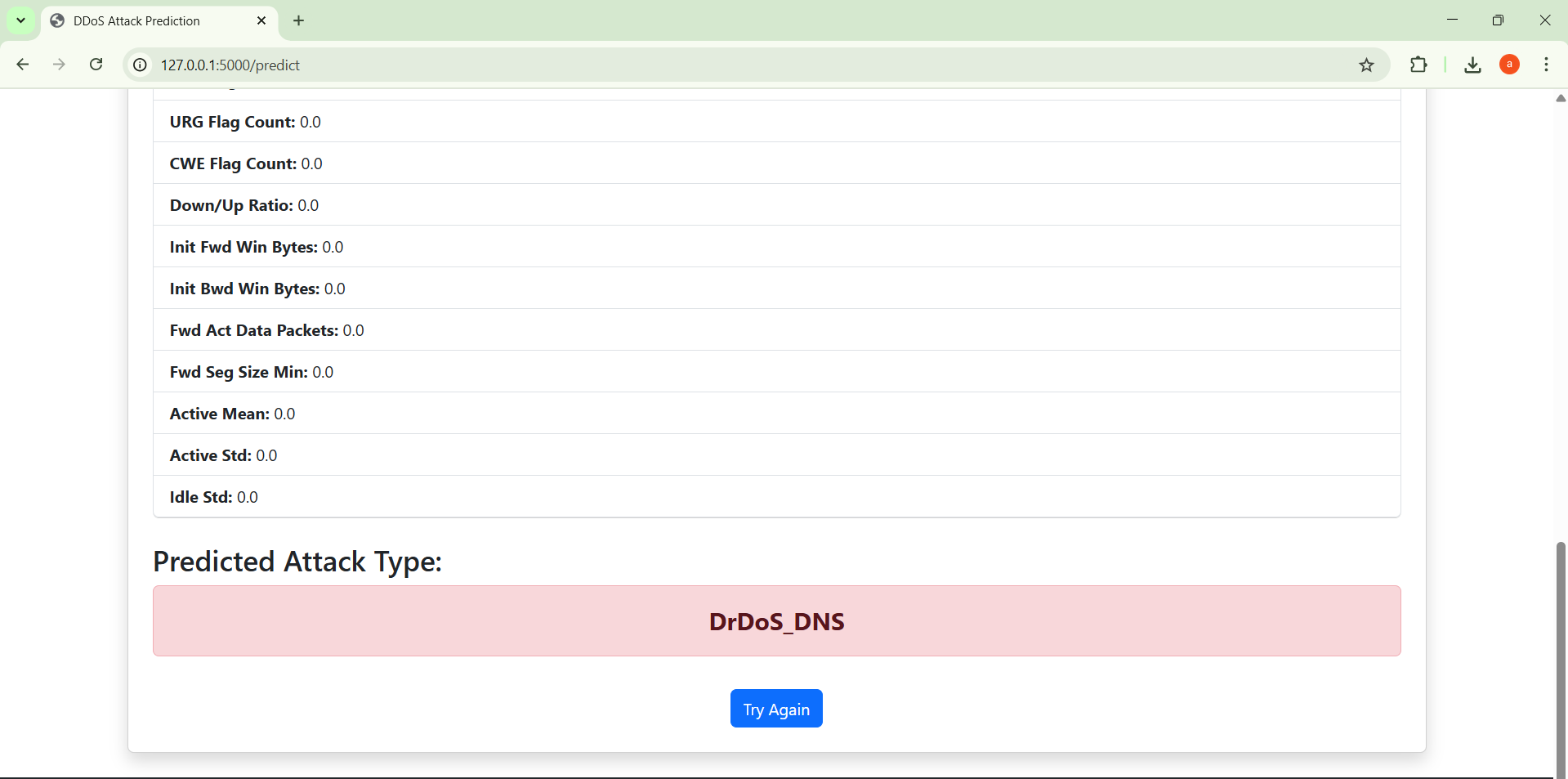
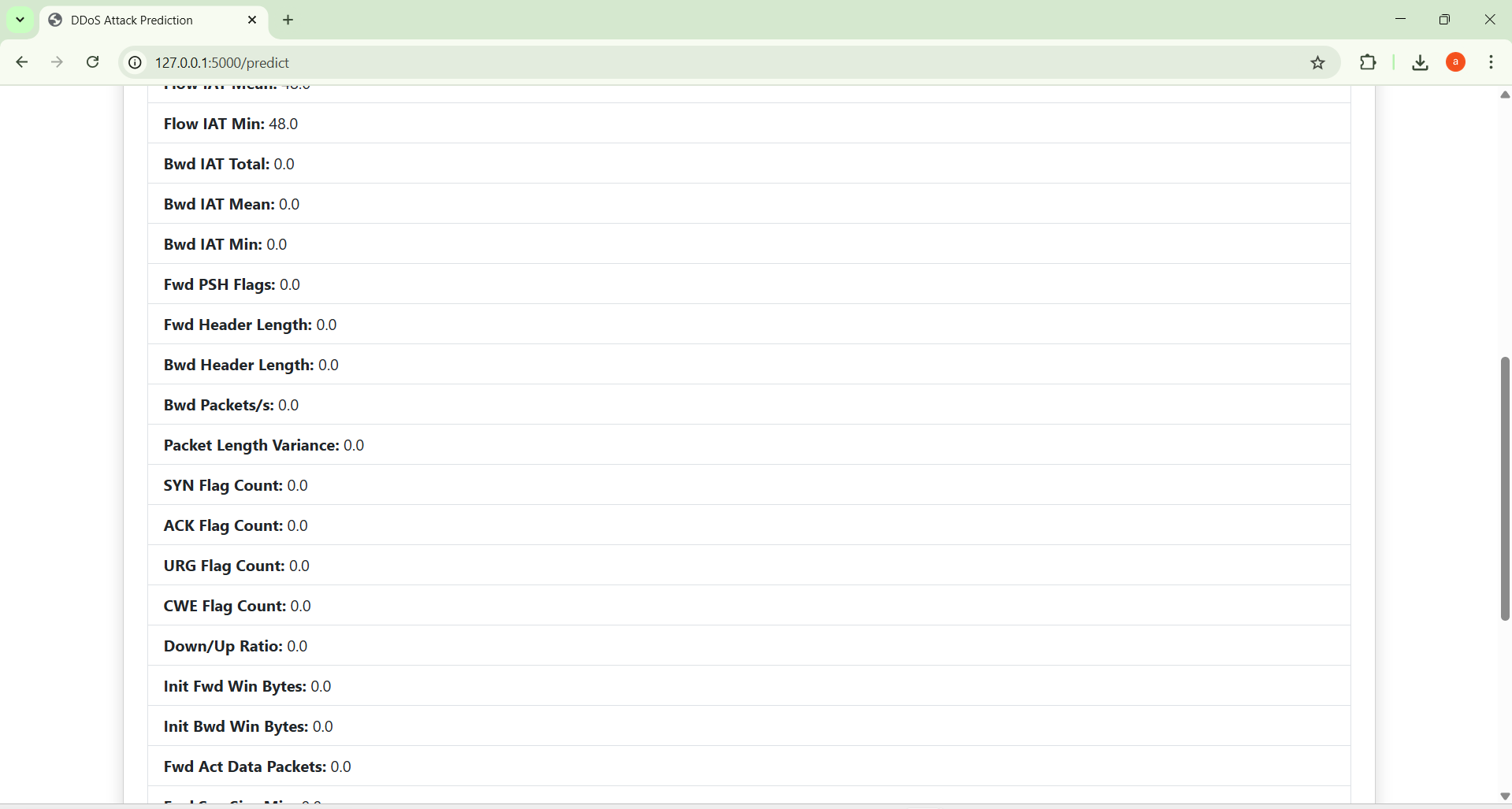
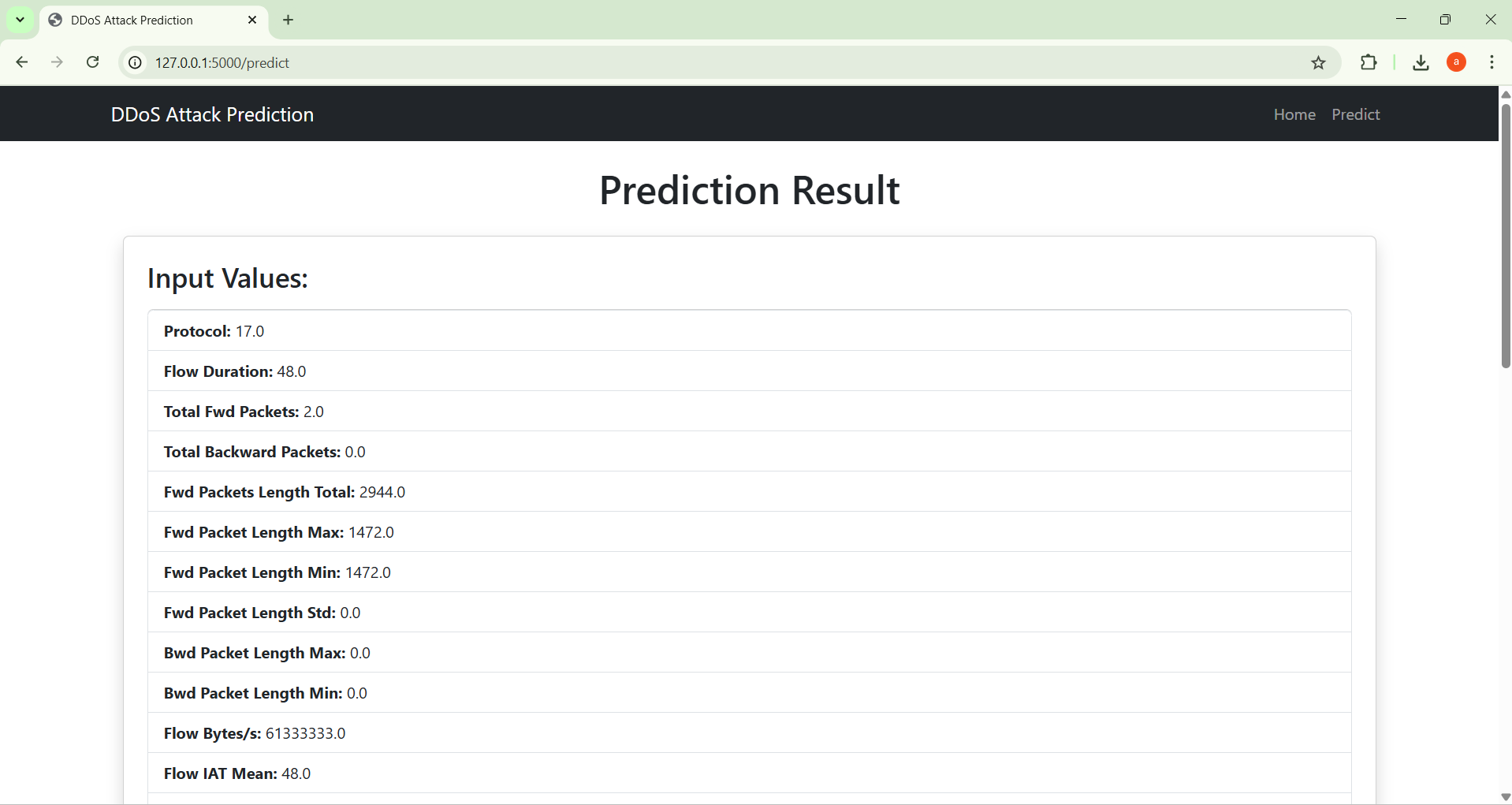
* Data Preprocessing: Used pandas, NumPy, and statistical functions to merge .parquet files, clean the dataset, handle missing values, and perform exploratory data analysis (EDA).
* Feature Engineering: Correlation analysis and feature selection techniques were applied to reduce multicollinearity and remove redundant features.
* Label Encoder: LabelEncoder was used to convert class labels into integers. The encoder was saved using pickle for reuse during prediction.
* Feature Scaling: Applied MinMaxScaler to normalize features. The scaler was saved as a .pkl file for use during real-time predictions.
* Modeling:
  + Random Forest: Trained using RandomForestClassifier with 100 estimators.
  + XGBoost: Trained using XGBClassifier with custom hyperparameters for optimal performance.
* Model Saving: Trained models were serialized and saved in .sav format using pickle for future predictions.
* Prediction Module: A CLI-based prediction system was implemented, where users input 33 feature values and receive the corresponding attack category as output.
* Hardware: The project was executed on an Intel i5 CPU with 8 GB RAM, proving efficiency even on mid-range hardware systems.

The modular and scalable architecture supports future integration with real-time detection systems or web-based dashboards.

**5. Experimental Evaluation**

Experimental evaluation was performed using the CIC DDoS 2019 dataset, which includes a wide range of DDoS attack types. Key findings include:

* Data Efficiency: By removing redundant features and normalizing the data, training and inference times were significantly reduced.
* Model Accuracy:
  + XGBoost achieved an accuracy of **94.5%** with high F1-scores across multiple classes.
  + Random Forest produced comparable performance, excelling in distinguishing between attack types.
* Evaluation Metrics:
  + Precision: >93%
  + Recall: >92%
  + F1-Score: ~93%
  + Training Time: <3 minutes per model (100 estimators)
* Visualization: Confusion matrices were plotted for both models using Seaborn, providing clear insights into classification effectiveness.
* Use Case Scenarios: The system can be integrated into enterprise networks, cloud environments, or SDN-based infrastructure to monitor and classify DDoS attacks in real-time.
* Limitations: The model’s performance may degrade with previously unseen attack variants. Continuous retraining and feature updates are necessary to maintain accuracy. Future directions include hybrid ensemble-deep learning architectures for even better generalization.

**13 .** **REFERENCES**

# 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.

# 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 (ISCI), Kuala Lumpur, Malaysia, 2024, pp. 253-257, doi: 10.1109/ISCI62787.2024.10668039.

# 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.

# 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.

# 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.