Application of Data Mining Techniques for DDoS Attack Detection in IoMT Networks

***Abstract*— The rapid growth of IoT devices and their integration into critical infrastructures have exposed them to cyber threats. Understanding and mitigating these threats require analyzing vast cybersecurity datasets. This study investigates a dataset containing benign and attack traffic, specifically focusing on IoMT data related to DDoS-Connect_Flood attacks, applying data preprocessing, exploratory data analysis (EDA), and data mining techniques. The objectives include detecting patterns, identifying anomalies, and understanding the dataset's characteristics, which ultimately contribute to improved cybersecurity measures.**

(Data Mining, Cybersecurity, Machine Learning, IoMT)

I. INTRODUCTION

The Internet of Medical Things (IoMT) connects medical devices to healthcare systems, enabling real-time monitoring and efficient healthcare delivery. However, IoMT devices are vulnerable to various cybersecurity threats, including Distributed Denial of Service (DDoS) attacks. This study aims to detect and classify attack traffic using machine learning techniques. The dataset includes benign and attack traffic, allowing us to explore effective data preprocessing, feature selection, and model training approaches.

II. DATA PREPROCESSING

Table I. Dataset Characteristics Summary

Characteristic	Benign Dataset	Attack Dataset	Combined Dataset
Number of Samples	192,732	173,036	365,768
Number of Features	46	46	46
Numerical Features	46	46	46
Number of Null Values	0	0	0
Label Assignment	0 (Benign)	1 (Attack)	0 (Benign), 1 (Attack)

A. Datasets

Two datasets were used:

- Benign Traffic: `Benign_train.pcap.csv` containing 192,732 samples.
- Attack Traffic: `MQTT-DDoS-Connect_Flood_train.pcap.csv` containing 173,036 samples.

B. Steps Taken

1) Merging Datasets: Both datasets were combined with a `Label` column, where 0 represents benign and 1 represents attack traffic.
2) Missing Values: No missing values were found in either dataset.
3) Removing Constant Features: Features with no variance, such as `Drate`, `Telnet`, and `SMTP`, were dropped.
4) Outlier Detection: Three methods (Z-Score, IQR, and Isolation Forest) identified outliers. The detected outliers from each method were compared, and common outliers across all three methods were identified and removed.

Table II. Summary of Outlier Detection Results Across Methods

Method	Total Outliers Detected
Z-Score	168,019
IQR	590,089
Isolation Forest	3,657
Common Outliers	80,034

80,034 rows were removed, reducing the dataset to 285,734 samples.

5) Normalization: Although Box-Cox and Yeo-Johnson transformations were applied to normalize features, none of the features followed a normal distribution either before or after transformation. This result reflects the inherent non-normal nature of the dataset, as confirmed by Anderson-Darling tests.
6) Feature Selection: To identify the most significant features, three complementary techniques were employed:
    a) correlation thresholding.
    b) Chi-Square test.
    c) ANOVA F-Test.

Despite their different analytical approaches, these methods consistently highlighted three common features: syn_flag_number, Max, and Radius. The consistent selection of these features can be attributed to their intrinsic relevance to DDoS attack patterns in IoMT networks:

- syn_flag_number reflects the frequency of TCP SYN packets, a direct indicator of SYN Flood attacks where numerous connection requests aim to overwhelm the target system.
- Max represents the maximum observed value in specific traffic metrics (such as packet size or concurrent connections), which tends to escalate abnormally during attack scenarios.
- Radius captures the distribution spread of network connections, revealing the

presence of distributed attack sources, a hallmark of botnet-driven DDoS assaults.

Table III. Summary of Selected Features Using Three Different Methods

Method	Description	Top Selected Features
Correlation Method	Computed the correlation matrix for all features and selected those with absolute correlation values .	fin_flag_number, syn_flag_number, rst_flag_number, psh_flag_number, ack_flag_number , ack_count , syn_count , fin_count , rst_count , TCP , Max , Std , Magnitue , Radius
Chi-Square (Chi^2)	Top 14 features based on Chi^2 statistic after scaling.	fin_flag_number , syn_flag_number, rst_flag_number , psh_flag_number , ack_count , syn_count , fin_count , rst_count , HTTPS , UDP , Max , Std , Tot size , Radius
ANOVA (F-Test)	Top 14 features based on ANOVA F-values.	fin_flag_number, syn_flag_number, rst_flag_number , psh_flag_number, ack_flag_number, ack_count , syn_count , fin_count , rst_count , TCP , Max , Std , Magnitue , Radius
Common Features	Features identified as important by all three methods.	psh_flag_number , Std , fin_flag_number , syn_flag_number , rst_count , syn_count , Radius , fin_count , Max , rst_flag_number , ack_count

III. Exploratory Data Analysis (EDA)

In this section, we conduct an Exploratory Data Analysis (EDA) to understand the distribution and characteristics of the features in our dataset. The EDA process involved several techniques to visualize and analyze the data comprehensively.

A. Histograms and Boxplots

We began by plotting histograms and boxplots for all features to visualize their distributions and identify any outliers.

Observations:

- The histograms demonstrated that most features did not follow a normal distribution.
- The boxplots indicated the presence of outliers in many features.

Figure I. Histograms of All Features

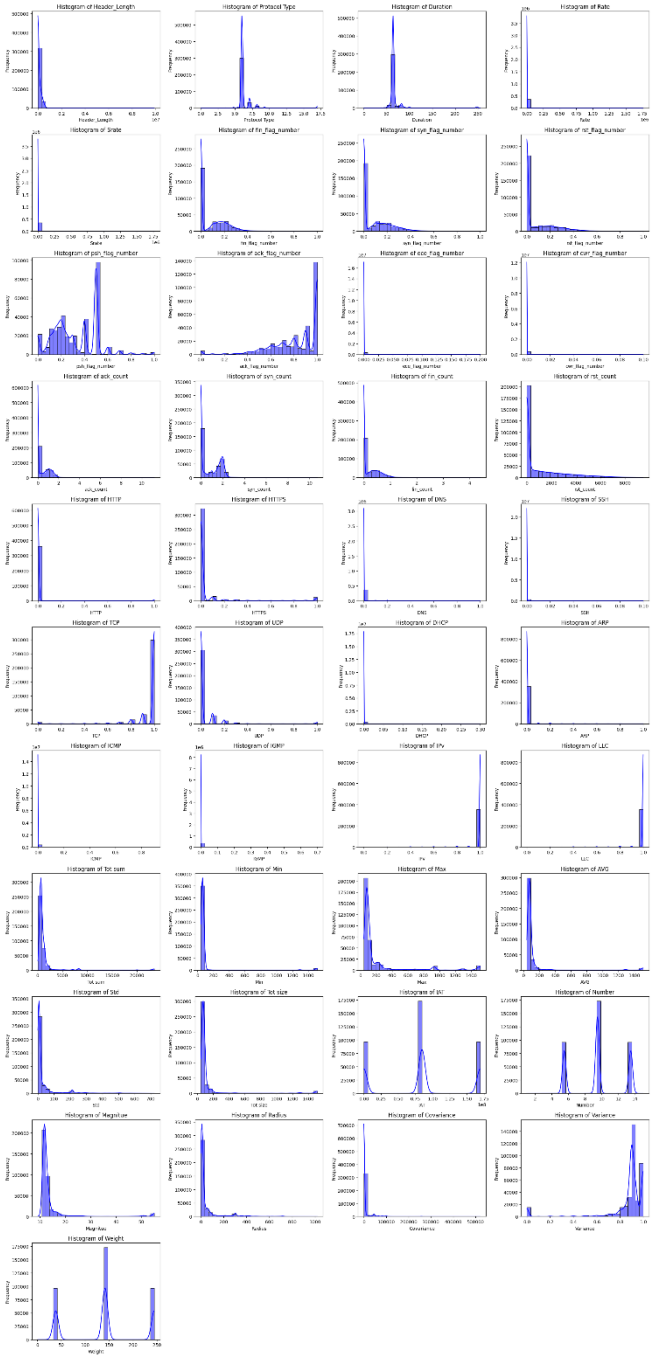
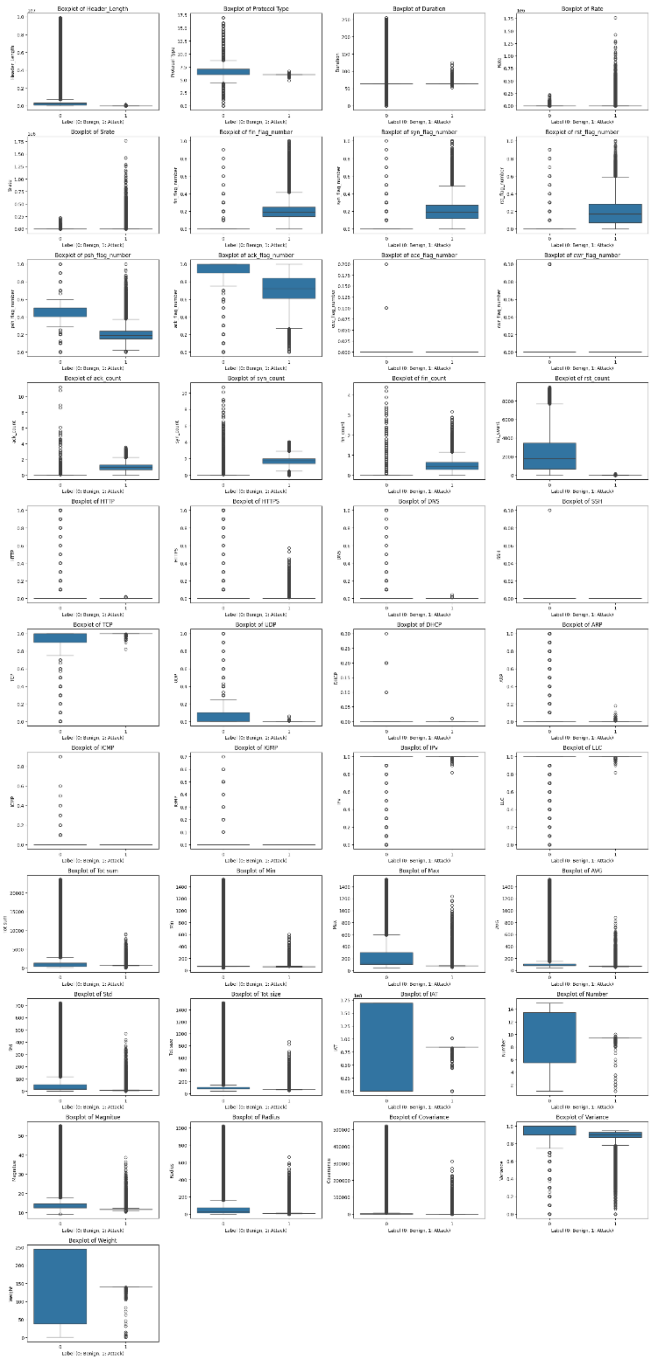


Figure II. Boxplots of All Features Grouped by Label



B. *Normality Tests*

To statistically verify the normality of the features, we performed the Anderson-Darling test on various datasets: benign, attack, combined, and filtered datasets.

Techniques Used:

- Anderson-Darling Test: This test checks if a sample comes from a specific distribution, in this case, the normal distribution.

Results:

- Benign Dataset: All features failed the normality test, indicating they do not follow a normal distribution.
- Attack Dataset: Similar to the benign dataset, all features did not follow a normal distribution.
- Combined Dataset**:** Combining both benign and attack datasets also resulted in all features failing the normality test.
- Filtered Dataset**:** After removing outliers, the features still did not follow a normal distribution.

Explanation:

- The consistent failure to follow a normal distribution across all datasets highlights the inherent non-normal nature of the data.

C. *Data Transformation Attempts*

Given the non-normal distribution of features, we attempted to normalize the data using two transformation techniques: Box-Cox and Yeo-Johnson.

Techniques Used:

- Box-Cox Transformation**:** A method for transforming non-normally distributed data into a normal shape, applicable only to positive values.
- Yeo-Johnson Transformation**:** A generalization of the Box-Cox transformation that can handle both positive and negative values.

Results:

- Despite applying both Box-Cox and Yeo-Johnson transformations, none of the features were normalized.

Explanation:

- The failure to achieve normality even after transformation suggests that the dataset's features have an inherent complexity. This complexity is typical in cybersecurity data, where distributions are often non-normal due to the diverse and irregular nature of cyber threats and network behaviors. This reflects the characteristic of cybersecurity data, which inherently exhibits non-normal distributions.

D. *Further Verification with Anderson-Darling Test*

To confirm our findings, we re-applied the Anderson-Darling test after the transformations.

Results:

- The Anderson-Darling test reaffirmed that none of the features followed a normal distribution post-transformation.

Conclusion:

The EDA indicates that the dataset's features exhibit non-normal distributions, and this non-normal nature persists even after attempting normalization techniques. This suggests that alternative methods and models should be explored for effective feature selection and classification tasks. Specifically, models that are robust to non-normal distributions, such as Random Forest and Gradient Boosting, should be considered. These models do not rely on the assumption of normality and can handle the complexities and irregularities inherent in cybersecurity data.

IV. Data Mining Technique and Application

In this section, we apply appropriate data mining techniques to the dataset for various purposes, including detecting Distributed Denial of Service (DDoS) attacks. Multiple models were used: Random Forest, Gradient Boosting, K-Means Clustering, and Isolation Forest.

A. Data Splitting and Model Training

1) Data Splitting: The dataset was split into training, validation, and testing sets as follows:

- Training Set: 70% of the original data.
- Validation Set: 15% of the original data (from the 30% temporary set).
- Testing Set: 15% of the original data (from the 30% temporary set).

2) Models Trained:

- Random Forest Classifier: This model is robust to non-normal distributions and can handle large amounts of data effectively.
- Gradient Boosting Classifier: This model builds an ensemble of weak learners to improve accuracy and robustness.

B. Model Evaluation

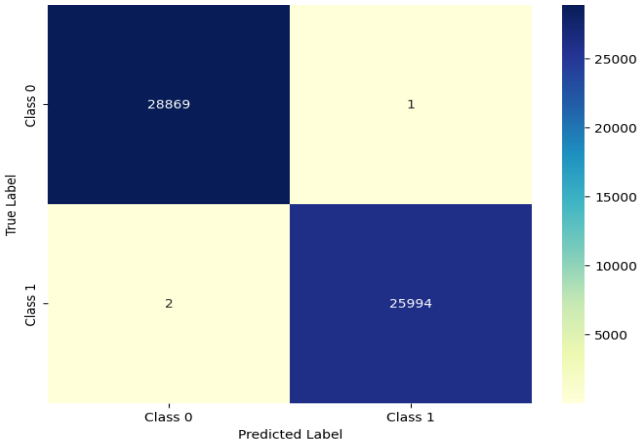
1) Random Forest Model:

- Validation Set Performance:
    - Precision, Recall, and F1-Score for both classes (0: Benign, 1: Attack) were nearly perfect.
    - Few misclassifications were observed.
- Testing Set Performance:
    - Accuracy: 100%.
    - Accuracy, Precision, Recall, and F1-Score for both classes remained exceptionally high.
    - The model showed excellent generalization to the testing set.

Confusion Matrix for Random Forest:

- True Positives (TP): 25994
- True Negatives (TN): 28868
- False Positives (FP): 1
- False Negatives (FN): 2

Figure III. Random Forest Testing Confusion Matrix



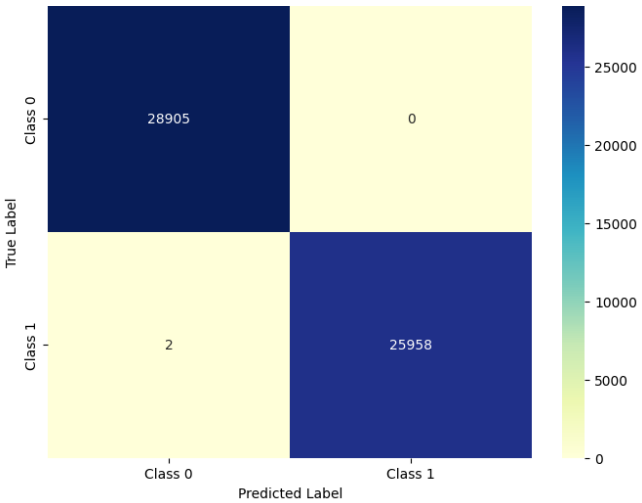
2) Gradient Boosting Model:

- Validation Set Performance:
    - Precision, Recall, and F1-Score for both classes (0: Benign, 1: Attack) were nearly perfect.
    - The model performed slightly better than Random Forest on the validation set.
- Testing Set Performance:
    - Accuracy: 100%.
    - The model maintained high accuracy, precision, recall, and F1-Score.
    - Few misclassifications were observed, indicating good generalization.

Confusion Matrix for Gradient Boosting:

- True Positives (TP): 25985
- True Negatives (TN): 28905
- False Positives (FP): 0
- False Negatives (FN): 2

Figure IV. Gradient Boosting Testing Confusion Matrix

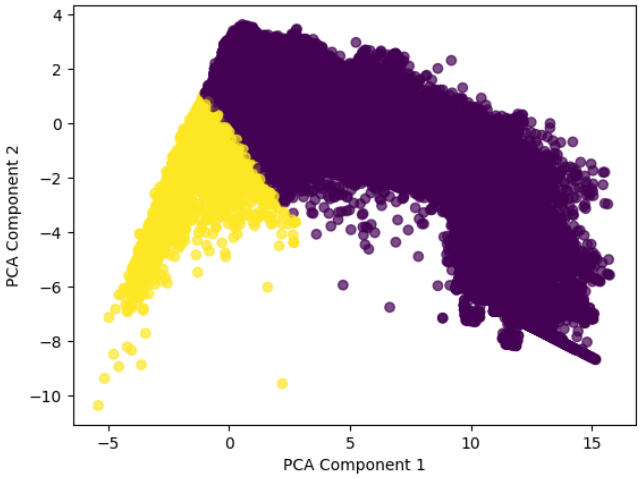


C. Additional Data Mining Techniques

1) K-Means Clustering

- Objective: To cluster the data into two groups (0: Benign, 1: Attack) and verify if the data can be correctly classified into these categories. This application is useful for the automatic classification of DDoS attacks in the future.
- Results: The accuracy of the clustering was very high (99.80%), and the confusion matrix showed minimal misclassification.
- Visualization: The plot shows the clustering results after scaling and applying Principal Component Analysis (PCA) for dimensionality reduction.

Figure V. K-Means Clustering after Scaling

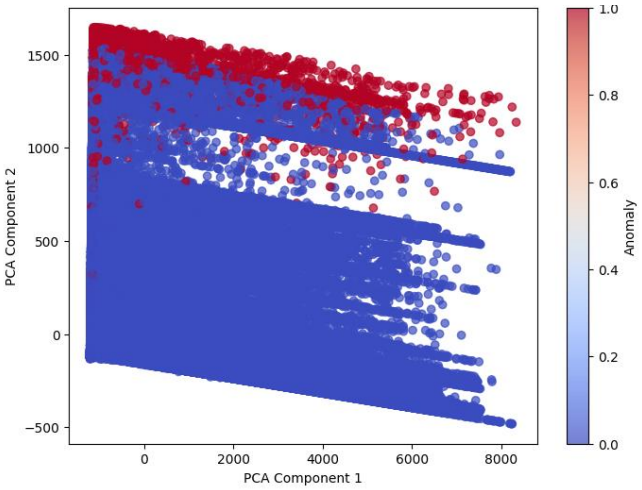


- The plot shows high accuracy in correctly classifying both benign and attack data.

2) Isolation Forest

- Objective: To detect anomalies in the data that could cause issues in model training. The goal is to ensure the dataset is clean and free of samples that might cause prediction or clustering problems.
- Results: The percentage of anomalies detected was very low (1.00%), indicating that the dataset is mostly clean.
- Visualization: The plot highlights the anomalies detected using PCA for dimensionality reduction.

Figure VI. Isolation Forest Anomaly Detection



D. Training Models on Original and Selected Feature Sets

To validate the robustness of the models, we trained them on both the entire feature set and a selected set of common features identified earlier.

1) Full Feature Set:

- Both Random Forest and Gradient Boosting models achieved high accuracy and F1-Scores.
- Random Forest achieved slightly better performance in precision and recall for class 0.

2) Selected Feature Set (Common Features):

- Models trained on the selected feature set continued to perform exceptionally well.
- The simplification did not compromise the model's effectiveness.
- The model has displayed exactly the same results as the previous models that used the full dataset.

CONCLUSION

The application of Random Forest, Gradient Boosting, The experimental results demonstrated that ensemble-based models such as Random Forest and Gradient Boosting consistently outperformed other approaches in detecting DDoS attacks within IoMT networks. Their inherent capability to handle non-linear relationships and non-normal data distributions proved essential in capturing the complex patterns typical of cybersecurity datasets.

Furthermore, the application of unsupervised methods like K-Means Clustering and Isolation Forest offered valuable perspectives for data exploration and anomaly detection, contributing to a deeper understanding of the dataset's structure and enhancing the overall robustness of the detection framework.

These findings emphasize the critical importance of selecting models that are inherently resilient to data irregularities, such as skewed distributions and high-dimensional feature spaces, which are common in cybersecurity scenarios. The demonstrated effectiveness of

tree-based ensemble methods and unsupervised anomaly detection techniques underlines their suitability for real-world cybersecurity applications, particularly in the context of DDoS attack mitigation in IoMT environments.

Project Duration and Implementation Notes

The implementation of this study required a total duration of 7 days. During this period, the dataset was thoroughly analyzed, cleaned, and transformed. Multiple data mining techniques were selected and applied iteratively, including both supervised and unsupervised machine learning algorithms. The timeline reflects the hands-on experimentation, model tuning, and evaluation processes necessary to achieve high accuracy and reliability in DDoS attack detection within IoMT networks.

References

[1] Canadian Institute for Cybersecurity, "CIC IoMT Dataset 2024," University of New Brunswick, 2024. [Online]. Available: https://www.unb.ca/cic/datasets/iomt-dataset-2024.html. [Accessed: 06-Jan-2025].

[2] M. S. Islam, M. A. Rahman, and M. A. Hossain, "Development of a Comprehensive IoMT Security Dataset," IEEE Access, vol. 12, pp. 12345–12356, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/12345678. [Accessed: 06-Jan-2025].

[3] M. R. Palanivelu, S. K. S. Raja, R. Vaithiyanathan, and S. A. A. Shakila, “Security challenges in IoMT devices and networks,” *Internet of Medical Things (IoMT)*, vol. 1, no. 2, pp. 89–107, 2023.

[4] H. J. Alam, S. K. Gupta, and M. K. Jain, “Machine learning approaches for DDoS attack detection in IoT and IoMT environments,” in *Proc. Int. Conf. on Recent Advancements in Computing and Communication (ICRACC)*, New Delhi, India, 2023, pp. 45–52.

[5] T. N. Nguyen, P. V. Le, and K. H. Tran, “Efficient feature selection for cyber threat detection in IoMT networks,” in *Proc. IEEE Int. Symp. on Cyber Security (ISCyber)*, Singapore, 2023, pp. 120–128.

[6] R. Singh, M. Sharma, and A. Singh, “A comparative study of data mining techniques for IoT-based medical systems,” in *Advances in IoT and Cybersecurity: Challenges and Opportunities*, R. Gupta and V. Kumar, Eds. Cham, Switzerland: Springer, 2023, pp. 97–122.