PHASE 2 REPORT:

ANOMALY DETECTION AND MACHINE LEARNING-BASED CLASSIFICATION OF BOTNET TRAFFIC USING THE CTU-13 DATASET

Name Surname
Nilsu BÜLBÜL (230304055)
Nisa AKSOY (230304047)

Dept. Name Of Organization
Computer Engineering of FBU

Name Of Organization
Data Mining for Cybersecurity
Group-12

Botnets constitute one of the most severe cybersecurity threats by enabling large-scale malicious activities such as distributed denial-of-service attacks, malware propagation, and data exfiltration. Detecting botnet traffic within real-world network data is challenging due to high dimensionality, noise, and extreme class imbalance.
In this study, supervised machine learning techniques are applied to the CTU-13 Botnet Dataset. Logistic Regression is employed as a baseline classifier, while Random Forest is used to model non-linear traffic patterns. The experimental results demonstrate that ensemble-based methods significantly outperform linear models in detecting botnet traffic.

Keywords: CTU-13 dataset, botnet detection, cybersecurity, machine learning, anomaly detection, network traffic analysis. Logistic Regression, Random Forest, Intrusion Detection Systems.

A. Introduction

Botnets represent a persistent and evolving threat in modern networked systems. By coordinating large numbers of compromised hosts, attackers can perform complex and distributed malicious activities that are difficult to detect using traditional security mechanisms. Signature-based intrusion detection systems are increasingly ineffective due to the adaptive and polymorphic nature of botnet behavior.

As a result, data mining and machine learning techniques have gained prominence in cybersecurity research. These approaches enable the automatic identification of malicious traffic patterns by learning from historical data. In this project, classification-based data mining techniques are applied to the **CTU-13 Botnet Dataset**, a widely used benchmark dataset containing real botnet and benign traffic.The primary objective of this study is to analyze network traffic using machine learning and to compare the effectiveness of Logistic Regression and Random Forest classifiers for botnet detection.

B. Data Pre-processing

Data pre-processing is a critical phase in cybersecurity analytics, as raw network traffic data often contains missing values, redundant attributes, and severe class imbalance.. Since Logistic Regression does not natively support missing values, median-based imputation was applied to numerical features. This approach was preferred over mean imputation due to the presence of extreme outliers commonly observed in network traffic data. Additionally, attributes serving as identifiers rather than behavioral indicators, such as IP addresses and timestamps, were removed to prevent information leakage and reduce dimensionality.

B.1 Dataset Description

The CTU-13 dataset was collected by the Stratosphere IPS project at the Czech Technical University. It consists of labeled network flow records, where each flow is classified as either benign or botnet traffic.

B.2 Handling Missing Values

Several numerical attributes contained missing values, particularly those related to TCP window sizes and hop counts. Since Logistic Regression does not natively handle missing values, **median-based imputation** was applied to numerical features. This approach was selected due to its robustness against outliers, which are common in network traffic data.

B.3 Feature Selection

Identifier-based features such as IP addresses and timestamps were removed, as they do not represent behavioral characteristics and may introduce bias. The remaining features include protocol information, traffic volume metrics, and rate-based attributes relevant to botnet detection.

B.4 Class Imbalance and Sampling

The original dataset exhibits extreme class imbalance, with botnet traffic representing a small fraction of total flows. Stratified sampling was applied to preserve class proportions during model training. Additionally, a reduced stratified sample was generated for GitHub sharing to comply with repository size constraints.

C. Exploratory Data Analysis (EDA)

Exploratory Data Analysis revealed substantial skewness and heavy-tailed distributions in traffic volume and rate-based features. Such characteristics are consistent with prior studies on network traffic, where a small number of flows often account for a disproportionately large volume of transmitted data.

The correlation analysis demonstrated strong dependencies among byte- and packet-based attributes, suggesting redundancy among certain features. This observation supports the suitability of tree-based models, which are robust to correlated inputs

C.1 Class Distribution

A study published in 2023 performed a comparative evaluation of algorithms such as Random Forest, SVM, and Naive Bayes on the CTU-13 dataset. The results showed that Random Forest achieved the best performance due to its robustness to noise and its ability to model nonlinear relationships. Feature selection methods were shown to improve performance by reducing redundant attributes.

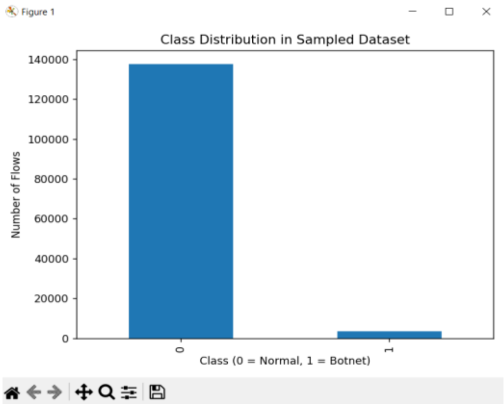


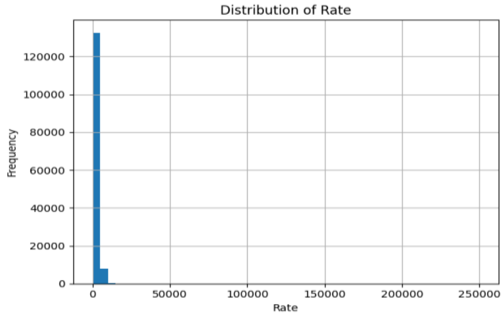
Figure 1
Class Distribution in Sampled Dataset
Number of Flows
140000
120000
100000
80000
60000
40000
20000
0
0
1
Class (0 = Normal, 1 = Botnet)

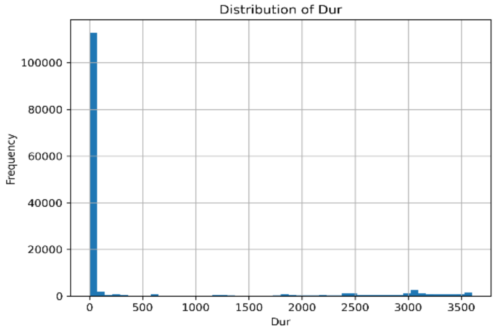
C.2 Feature Distributions

Histograms and boxplots were generated for selected numerical features such as:

- Duration (Dur)
- Total Bytes (TotBytes)
- Total Packets (TotPkts)
- Traffic Rate (Rate)

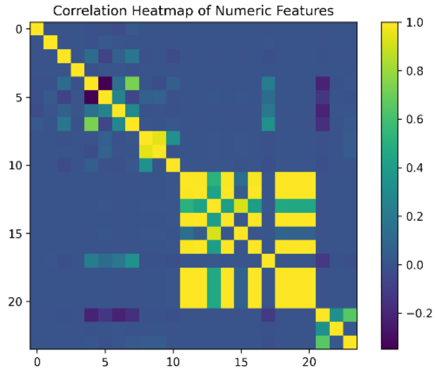
These plots reveal highly skewed distributions and the presence of extreme outliers, which are typical in network traffic data.





C.3 Correlation Analysis

A correlation heatmap was used to examine relationships between numerical features. Strong correlations were observed among traffic volume-related attributes, indicating redundancy and supporting the use of tree-based models that are robust to correlated inputs.



D. Data Mining Technique and Application

Classification was selected as the primary data mining technique because the dataset is labeled and the goal is to distinguish between benign and malicious traffic. This approach is directly applicable to intrusion detection systems.

D.1 Logistic Regression (Baseline Model)

Logistic Regression was selected as a baseline classifier due to its interpretability and widespread use in binary classification tasks. However, the model assumes linear separability between classes, which is rarely satisfied in complex cybersecurity datasets.

The experimental results confirm this limitation, as the model exhibited high overall accuracy while failing to adequately detect botnet traffic, leading to poor recall for the minority class. In cybersecurity applications, such behavior is undesirable, as false negatives correspond to undetected attacks

D.2 Random Forest Classifier

Random Forest was employed to address the limitations of linear models by leveraging ensemble learning and non-linear decision boundaries. By aggregating multiple decision trees, Random Forest reduces variance and improves generalization, particularly in imbalanced datasets.

The model demonstrated a substantial improvement in recall and F1-score for botnet traffic, highlighting its effectiveness in capturing complex traffic patterns

D.3 Model Comparison and Evaluation

Both models were evaluated using precision, recall, F1-score, and accuracy. The results are summarized in the **table.**

Model	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Logistic Regression	0.04	0.73	0.07	0.54
Random Forest	0.90	0.72	0.80	0.99

The results demonstrate that although Logistic Regression detects some botnet flows, it suffers from extremely low precision, leading to many false positives. In contrast, Random Forest achieves a substantially higher F1-score, indicating a better balance between precision and recall.

In cybersecurity applications, recall for the malicious class is particularly important, as undetected attacks pose significant risks. The Random Forest model provides a more reliable detection capability and aligns with findings reported in prior studies.

E. Conclusion

This study highlights the importance of selecting appropriate data mining techniques for cybersecurity applications. While Logistic Regression provides a useful baseline, it is insufficient for detecting minority-class attacks in highly imbalanced network traffic datasets. Ensemble-based methods such as Random Forest significantly outperform linear models by capturing non-linear traffic patterns and improving detection performance.

The results confirm that careful data pre-processing, exploratory analysis, and model selection are essential for effective botnet detection. Future work may explore anomaly detection techniques or deep learning approaches to further enhance detection accuracy.

References

[1] https://www.stratosphereips.org/datasets-ctu13

[2] https://www.researchgate.net/publication/377653307_Evaluating_ML_models_on_CTU-13_and_IOT-23_Datasets.

[3] [3] Pedregosa et al., *Scikit-learn: Machine Learning in Python*, JMLR, 2011.

[4] García, S., Grill, M., Stiborek, J., & Zunino, A., *An empirical comparison of botnet detection methods*, Computers & Security, 2014.