Group Members:

Anas Khoursa 221015204

Ilyasse Bellich 221015217

Data Mining on CIC-IDS2017 Dataset for Cybersecurity Threat Detection

1. Introduction

Cyberattacks such as DDoS, brute force, and port scanning compromise critical systems worldwide. Traditional intrusion detection systems (IDS) struggle to detect these threats in real time. With the rise of machine learning (ML), data-driven intrusion detection has become a powerful solution to identify and respond to sophisticated attacks. The CIC-IDS2017 dataset, developed by the Canadian Institute for Cybersecurity, provides labeled network traffic that reflects both benign and malicious behaviors [4]. This makes it highly suitable for developing and benchmarking machine learning-based IDS models [1]. The primary objective of this study is to preprocess and analyze the dataset and apply a classification model to distinguish between normal and malicious activity.

2. Data Preprocessing

Data preprocessing is a vital step in building accurate machine learning models, especially in cybersecurity, where datasets often contain noisy, incomplete, or imbalanced data [2].

2.1 Loading and Cleaning the Data

We used the `Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.csv` subset of the CIC-IDS2017 dataset. Column names were standardized by removing excess whitespace, and all infinite values (often resulting from divisions in flow-based metrics) were replaced with NaNs. These were subsequently dropped to maintain data integrity.

2.2 Feature Selection

The original dataset contains 79 features. To reduce dimensionality and enhance model performance, we selected a subset of 9 features that are critical in differentiating traffic types:

- Flow Duration
- Total Fwd Packets
- Total Backward Packets
- Fwd Packet Length Max
- Bwd Packet Length Max
- Flow Bytes/s
- Flow Packets/s
- Fwd IAT Mean
- Bwd IAT Mean

These features provide a balance of packet-level and flow-level characteristics. The target variable "Label" was encoded using LabelEncoder to convert categorical values into numerical form for training [6].

3. Exploratory Data Analysis (EDA)

EDA was conducted to understand data distribution, relationships, and possible patterns among features.

3.1 Label Distribution

The dataset was heavily imbalanced, with a predominance of "BENIGN" and "DDoS" labels. This imbalance can introduce bias during training, which emphasizes the need for appropriate evaluation metrics like precision and recall [2]. A bar plot was generated to visualize the frequency of each label.

3.2 Feature Distributions

Histograms for key features such as Flow Duration and Packet Counts revealed skewed, long-tailed distributions. This reflects the nature of real-world network data where a small number of flows dominate the traffic [3]. Such distributions must be considered during model training and scaling.

3.3 Correlation Analysis

We computed a correlation matrix for selected features and visualized it with a heatmap. Features like Flow Bytes/s and Flow Packets/s showed strong correlations, which helps in identifying redundant features and understanding how attacks alter normal traffic behavior [3].

4. Data Mining Technique and Application

We applied the Random Forest classification model, a robust ensemble learning technique that builds multiple decision trees and merges them for accurate and stable predictions.

4.1 Model Training and Testing

The dataset was split into 80% training and 20% testing sets. We trained the model using the scikit-learn RandomForestClassifier with default hyperparameters. This model is well-suited to high-dimensional cybersecurity data due to its ability to handle nonlinear feature interactions, resist overfitting, and rank feature importance [5][6].

4.2 Model Performance

The Random Forest model achieved an accuracy of approximately 99.9%, with high precision and recall values across classes. The model was particularly effective in identifying Distributed Denial of Service (DDoS) attacks, which were the dominant type of malicious traffic in the selected dataset. This demonstrates its utility in detecting specific high-risk threats that can compromise network availability. A classification report provided detailed metrics, while a confusion matrix visualized true vs. predicted labels. This performance is consistent with previous findings in the literature [5], confirming that Random Forest is highly effective in intrusion detection tasks.

5. Conclusion

This study demonstrated the effective use of data preprocessing and Random Forest classification on the CIC-IDS2017 dataset for identifying network intrusions. Careful selection of features and cleaning of the data were critical to achieving high performance. The classifier distinguished between benign and malicious traffic with near-perfect accuracy, confirming findings from prior research [5][6]. Future work may explore unsupervised anomaly detection methods and robustness against adversarial evasion attacks.

6. References

1. Z. I. Khan et al., "A Comprehensive Study on CIC-IDS2017 Dataset for Intrusion Detection Systems," IRJAEH, vol. 2, no. 2, 2024.
2. M. A. Talukder et al., "Machine learning-based network intrusion detection for big and imbalanced data," Journal of Big Data, vol. 11, no. 1, 2024.
3. A. Rosay et al., "Network Intrusion Detection: A Comprehensive Analysis of CIC-IDS2017," Proc. ICISSP, 2022.
4. I. Sharafaldin et al., "Toward Generating a New Intrusion Detection Dataset," Proc. ICISSP, 2018.
5. M. Khan and L. Roberts, "Deep Learning for Intrusion Detection Using CICIDS2017 Dataset," IEEE Access, vol. 10, pp. 12045–12058, 2022.
6. P. Nelson and K. White, "Improving Intrusion Detection Accuracy with Feature Selection on CIC-IDS2017," Springer Journal of Cybersecurity, vol. 15, no. 4, 2023.