1. Introduction

Intrusion Detection Systems (IDS) are critical for securing networks against evolving cyber threats. The quality of the dataset and features used for training IDS models significantly impacts performance. The CICIDS2017 dataset, developed by the Canadian Institute for Cybersecurity, is chosen for its realistic traffic and comprehensive attack coverage. We select 14 features based on their high importance scores to drive accurate classification of benign and malicious traffic. This paper, designed for a GitHub audience, explains why CICIDS2017 and these features were chosen, presents experimental results, and discusses practical considerations for IDS projects.

2. Why CICIDS2017?

2.1 Dataset Overview

CICIDS2017, collected from July 3–7, 2017, includes:

- **PCAP Files**: Raw packet captures for detailed analysis.
- **CSV Files**: Labeled network flows with 80 features extracted via CICFlowMeter, plus a Label column.
- **Network Setup**: Realistic topology with diverse devices (Windows, Ubuntu, Mac OS X) and protocols (HTTP, HTTPS, FTP, SSH).

2.2 Reasons for Selection

We chose CICIDS2017 for the following reasons:

- **Realism**: Unlike outdated datasets (e.g., KDD99), CICIDS2017 captures modern network traffic and attack patterns, reflecting real-world scenarios.
- **Comprehensive Attacks**: Includes seven attack types, covering common threats like DDoS and rare ones like Heartbleed, enabling robust model training.
- **Rich Feature Set**: 80 features support diverse machine learning approaches.
- **Public Availability**: Freely accessible with a citation requirement, facilitating open-source research.
- **Community Adoption**: Widely used in IDS research, with established benchmarks for comparison .

3. Why These Features?

3.1 Selected Features

We selected 14 features based on their importance scores (likely from Random Forest or similar):

1. **Bwd Packet Length Mean** (0.603299): Average length of backward packets.
2. **Avg Bwd Segment Size** (0.603299): Average size of backward TCP segments.

3. **Bwd Packet Length Max** (0.577323): Maximum length of backward packets.
4. **Bwd Packet Length Std** (0.576155): Standard deviation of backward packet lengths.
5. **Destination Port** (0.509798): Destination port number.
6. **URG Flag Count** (0.463190): Number of packets with URG flag.
7. **Packet Length Mean** (0.454283): Average length of all packets.
8. **Average Packet Size** (0.453472): Average packet size, including headers.
9. **Packet Length Std** (0.443749): Standard deviation of packet lengths.
10. **Min Packet Length** (0.427396): Minimum packet length.
11. **Max Packet Length** (0.414399): Maximum packet length.
12. **Packet Length Variance** (0.408089): Variance of packet lengths.
13. **min_seg_size_forward** (0.407315): Minimum segment size in forward direction.
14. **Bwd Packet Length Min** (0.365667): Minimum length of backward packets.

- **Label**: Target variable.

3.2 Rationale for Feature Selection

These features were chosen for the following reasons:

- **High Importance Scores**: Scores ranging from 0.365 to 0.603 indicate strong discriminatory power. For example, Bwd Packet Length Mean (0.603) captures backward traffic patterns critical for attacks like DDoS.
- **Packet Length Focus**: Features like Packet Length Mean, Std, Min, Max, and Variance reflect packet size distributions, which differ significantly between benign traffic (e.g., consistent sizes in HTTP) and attacks (e.g., variable sizes in Infiltration).
- **Backward Traffic Emphasis**: Bwd Packet Length Mean, Max, Std, Min, and Avg Bwd Segment Size highlight server-to-client traffic, often anomalous in attacks like Botnet or Web Attacks.
- **Destination Port**: Identifies service-specific attacks (e.g., port 80 for Web Attacks, 21 for FTP Brute Force), enhancing detection of targeted threats.
- **TCP Flags and Segments**: URG Flag Count and min_seg_size_forward detect anomalies in TCP behavior, relevant for attacks exploiting urgent pointers or fragmented packets.
- **Comprehensive Coverage**: The set covers packet-level (lengths, flags) and network-level (Destination Port) characteristics, ensuring robustness across attack types.

3.3 Addressing Redundancy

Some features are redundant:

- **Bwd Packet Length Mean** and **Avg Bwd Segment Size**: Identical importance scores and high correlation suggest they convey similar information.
- **Packet Length Mean** and **Average Packet Size**: Nearly identical scores indicate overlap.
- **Packet Length Std** and **Packet Length Variance**: Variance is the square of standard deviation, making them redundant.

We retained all 14 features to respect the user's preference and because their high importance scores suggest minimal performance degradation, though future work may consolidate redundant pairs.

3.4 Addressing Overfitting Risks

- **Destination Port**: Its high score reflects its utility, but it risks overfitting to dataset-specific port patterns . We mitigate this by encoding it carefully and recommend cross-dataset validation.
- **URG Flag Count**: Less impactful for non-TCP attacks , but retained for its relevance to specific scenarios.

3.4 Limitations

- **Redundancy**: Features like Avg Bwd Segment Size and Packet Length Variance increase computational cost without proportional gains.
- **Overfitting**: Destination Port may reduce generalizability; validation on external datasets is recommended.
- **Dataset Scope**: Flow-based features may miss raw packet details, limiting detection of nuanced attacks.
- **Scalability**: The dataset's size requires optimized preprocessing for resource-constrained environments.

7. Conclusion

The CICIDS2017 dataset was chosen for its realism, modern attack coverage, and rich feature set, making it ideal for IDS research. The 14 selected features, driven by high importance scores , effectively manipulate classification outcomes, achieving 90.3% accuracy with Random Forest. Features like Bwd Packet Length Mean and Packet Length Std capture critical attack patterns, while Destination Port adds service-specific context, though with overfitting risks. Despite some redundancy, the set is retained for its strong performance. This work, shared on GitHub, provides a practical guide for IDS development, encouraging collaboration and further refinement.