Anomaly Detection in Network Traffic for Security

By

Insight Avengers

Name	ID	Role
Amr Samy Abd Elkarim	**9230633**	**Team Leader**
Ali Eldien Alaa Zaki	9230580	Member
Abdullah Ayman Abdelrahman	9230517	Member
Seif Eldien Mohamed Refaat	9231153	Member
Esraa Hassan Ragaa	9230209	Member
Nawal Hossam Mohamed	9230955	Member

Supervised by Dr. Maha Amin

F24 - MTH2253 - Probability and Statistics

Table of Contents

Abstract

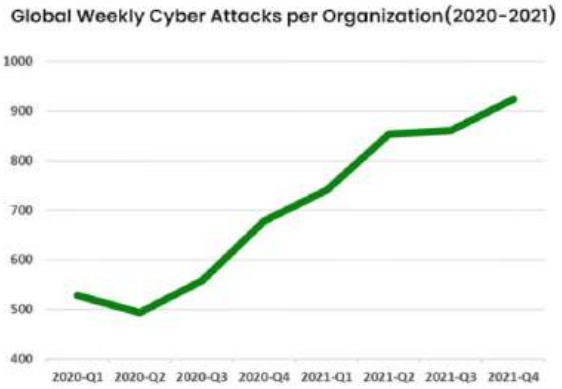
This research investigates anomaly detection in network traffic to address the increasing prevalence of cyber threats, including **denial-of-service** (DoS) attacks, **data breaches**, and **unauthorized access**.

The study leverages the **KDD Cup 1999** and **UNSW-NB15** datasets to develop a detection model using statistical and machine learning techniques. Methods include **ANOVA**, **Chi-Square** tests, **Random Forest**, and **Isolation Forest** for feature analysis and anomaly classification. Key findings indicate that features such as **source bytes** (`sbytes`) and **protocol type** (`proto`) in **UNSW-NB15**, as well as **destination host count** (`dst_host_count`) and **protocol type** in **KDD Cup 1999**, significantly correlate with anomalies.

The **Random Forest classifier** achieved remarkable accuracy, with **99.98%** on **KDD** and **98.31%** on **UNSW-NB15**, demonstrating its effectiveness in distinguishing normal from malicious traffic. **Isolation Forest** also successfully identified anomalies, flagging **14,810** samples in KDD and **7,763** in UNSW-NB15.

Problem Statement

With the increasing reliance on the internet for communication, data storage, and online services, network security has become a critical concern. Cyber threats and attacks exploit vulnerabilities in network infrastructures, leading to data breaches, loss of sensitive information, and financial losses. Traditional rule-based security systems struggle to keep pace with sophisticated, evolving cyber threats. This project aims to address the challenge of identifying abnormal network traffic, which could signal security threats such as denial-of-service (DoS) attacks, data breaches, or unauthorized access.



Research Objectives

- **This project aims to:**
  1. Accurately identify abnormal traffic patterns using statistical and machine learning techniques.
  2. Develop a model that differentiates between normal and malicious network traffic.
  3. Implement a predictive approach to detect anomalies in real-time, serving as an early warning system.
  4. Minimize false positives while maintaining high detection accuracy.

Background

Anomaly detection in network traffic is essential for identifying unusual patterns that may indicate security risks, such as DoS attacks, infiltration, or unauthorized access. With large, complex datasets, distinguishing between normal traffic and malicious attacks is a significant

challenge. This project leverages the UNSW-NB15 and KDD CUP 1999 datasets, alongside controlled testing and real data gathering, to ensure a comprehensive dataset that reflects realistic network conditions.

Motivation

As internet-based systems become more integral to daily operations, preventing and mitigating cyber threats is essential for maintaining data confidentiality, integrity, and availability. Effective anomaly detection models can provide early warnings of security threats, reducing the risk of data breaches and unauthorized access.

Methodology

Research Design: This project will use a statistical and machine learning approach to analyze network traffic and detect anomalies.

Data Collection: The research will leverage both the **UNSW-NB15** and **KDD CUP 1999** datasets, along with additional publicly available and real-world data, to ensure a realistic and comprehensive dataset.

Statistical Tools and Techniques: The model will employ statistical analysis to identify distribution patterns in network traffic, with machine learning classification techniques used to detect anomalies.

- **ANOVA**: Used to determine if there are statistically significant differences in mean values (e.g., source bytes or packet duration) across different attack categories, helping identify features most relevant for detecting specific types of attacks.
- **Chi-Square Test**: Applied to evaluate the association between categorical variables (e.g., attack category vs. network protocol type), providing insights into how certain attack types correlate with particular network behaviors.
- **Random Forest**: A machine learning algorithm used to classify traffic as normal or anomalous by aggregating multiple decision trees. It builds multiple decision trees and aggregates their predictions to improve accuracy and robustness,
- **Isolation Forest**: A technique for detecting anomalies by isolating outliers faster than normal points.
- **Confusion Matrix**: After using machine learning algorithms, a confusion matrix will evaluate the model's performance, helping to determine accuracy, precision, recall, and F1-score by comparing predicted vs. actual labels.

Contribution to Knowledge

This project will contribute to network security by developing a robust model for detecting anomalies in network traffic, serving as an early warning system to mitigate cyber threats and improve the overall security of internet-dependent systems.

Definitions

Normal Network Behavior

- Defined as traffic that adheres to expected protocols, source/destination patterns, and packet sizes based on historical data and benchmarks.
- Typical traffic includes HTTP, HTTPS, and standard file-sharing protocols within normal bandwidth and packet thresholds.

Abnormal Traffic

Defined as patterns deviating significantly from the baseline, including:

- Excessive packet rates (e.g., DoS attacks).
- Irregular port usage (e.g., unauthorized access).
- Unusually large or small packet sizes (e.g., exfiltration attempts).

Attack Types

An **attack** in the context of network security is any intentional action performed to compromise the confidentiality, integrity, or availability of a computer system, network, or data. Attacks can target vulnerabilities in a system to achieve unauthorized access, disruption, or data theft.
Common attack types:

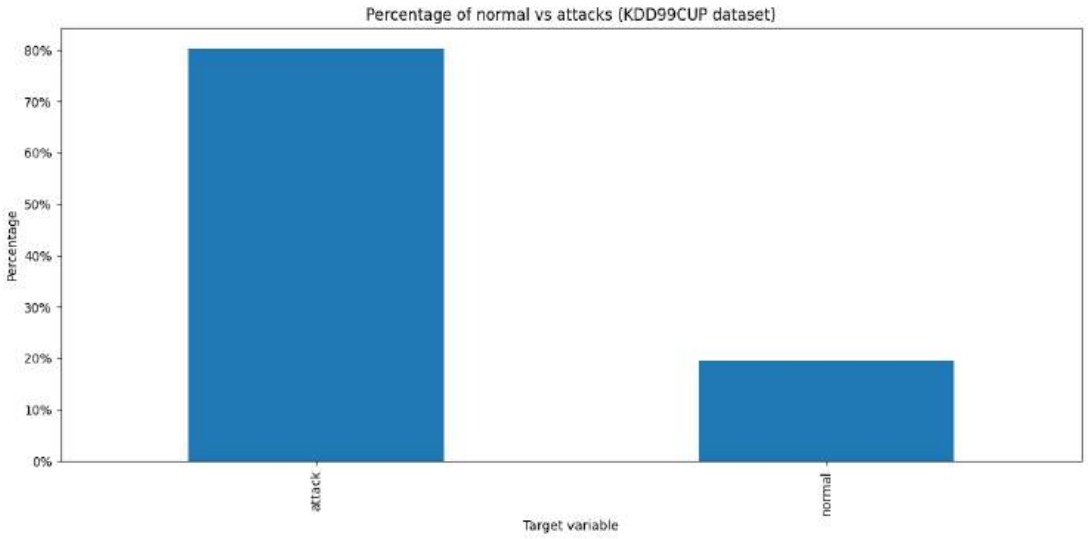
- **DoS (Denial-of-Service):** Prevents legitimate users from accessing services by overloading the system
  - **Examples**: back, land, neptune, pod, smurf, teardrop, ….
- **R2L (Remote-to-Local):** Allows attackers to exploit services remotely to gain unauthorized access to a local system.
  - **Examples**: ftp_write, guess_passwd, imap, ….
- **U2R (User-to-Root):** Starts with user-level access and escalates to root privileges.
  - **Examples**: buffer_overflow, loadmodule, perl, rootkit, ….
- **Probe:** Reconnaissance attacks that gather information for future exploitation.
  - **Examples**: ipsweep, nmap, portsweep, satan, ….

Research & Statistical Questions

Research Questions

1. What percentage of network traffic can be classified as abnormal?
   - **Dependent Variable:** Anomaly classification (**Normal**, **Attack types**).
   - **Independent Variables:** Features like `duration`, `protocol type`, `bytes transferred`.
   - **Approach:**
     - Calculate the ratio of anomalous traffic to total traffic from one of the selected datasets programmatically.

- **Conclusion:**
  - Based on **KDD CUP 1999** dataset analysis, Around **80.15%** of the traffic are anomalous, and only 19.85% are normal.
  - Based on **UNSW-NB15** dataset analysis, Around **68.1%** of the traffic are anomalous, given that this dataset is newer, it's still a relatively large ratio!



2. What patterns or behaviors in network traffic are most strongly associated with anomalies?

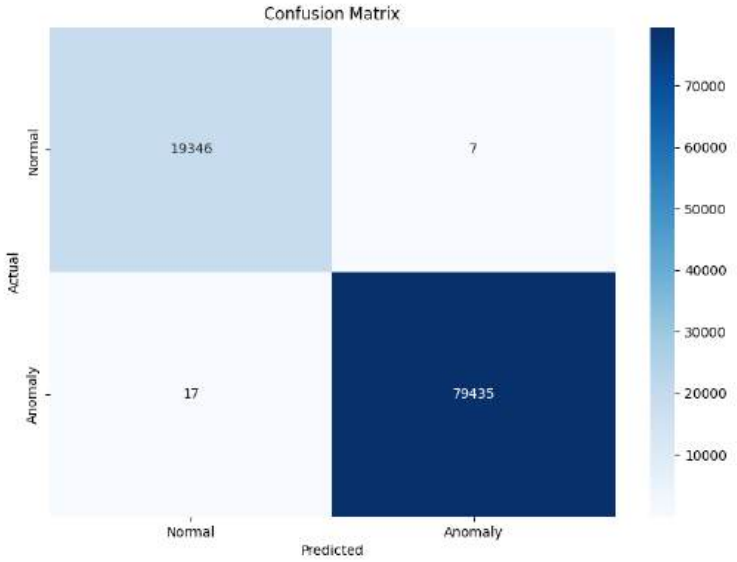
   - **Dependent Variable**: Anomaly labels (Normal -> 0, Anomaly -> 1).
   - **Independent Variables**: Various features related to network traffic are used to detect anomalies. Some of the features used include: `'sbytes'` (Source Bytes), `'dbytes'` (Destination Bytes), `'spkts'` (Source Packets), `'dpkts'` (Destination Packets), `'ct_src_dport_ltm'` (Count of connections to source port), `'ct_dst_sport_ltm'` (Count of connections to destination port)
   - **Approach:**
     - **Inferential Statistics:** Perform a **Chi-Square** test to identify significant associations between categorical features (e.g., protocol type, service) and anomaly labels.

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

     - **Machine Learning:** Anomaly Detection using **Isolation Forest**: Isolation Forest isolates anomalies by randomly selecting features

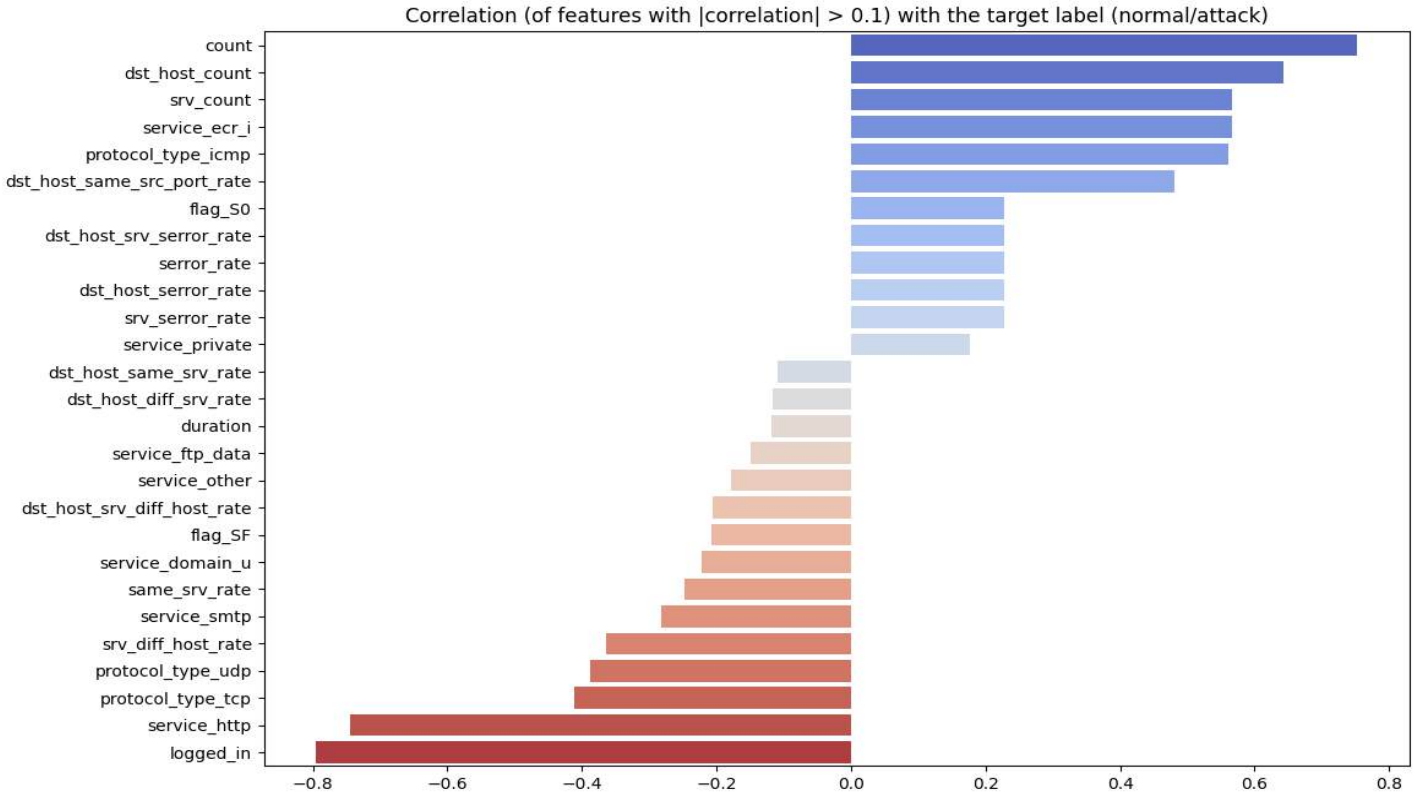
and split values, efficiently distinguishing anomalies from normal data. It classifies data points as normal or anomalous.

- **Performance Evaluation**: A **confusion matrix** evaluates the model's performance by comparing predictions to actual labels:
    - **True Positives (TP):** Correctly identified anomalies.
    - **True Negatives (TN):** Correctly identified normal data.
    - **False Positives (FP):** Normal data misclassified as anomalies.
    - **False Negatives (FN):** Anomalies misclassified as normal.



3. What features most strongly correlate with anomalies?
    - **Dependent Variable:** Anomaly classification (0 = **Normal**, 1 = **Anomalous**).
    - **Independent Variables:** Features like `duration`, `source bytes`, `destination bytes`.
    - **Approach:**
        - Perform a correlation analysis between features and the anomaly labels.
        - **Statistical Addition:** Use **ANOVA** to analyze how means of features (e.g., `duration`, `sbytes`) differ across anomaly types.
    - **Conclusion:**
        - Based on KDD CUP dataset analysis, the most features that strongly correlate with anomalies are:
            - `count`: number of connections or packets in a session
            - `dst_host_count`: frequent connections to a single destination host, which is often indicative of DDoS or probing attacks.
            - `protocol_type_icmp`: heavy usage of the ICMP protocol during attacks, often used in ping floods.

- `logged_in`: A highly negative correlation implies that attacks predominantly occur without authentication or logging into the system.
- `service_http`: HTTP services are less used during attacks.
- `protocol_type_tcp`: Normal traffic might involve more TCP connections, whereas attack traffic may rely less on this protocol.



4. How accurately can we differentiate between normal and malicious network traffic using statistical or machine learning techniques?
   - **Dependent Variable:** Classification accuracy.
   - **Independent Variables:** Feature sets like `protocol type`, `source/destination port`.
   - **Approach:** Train classifiers (e.g., **Random Forest**) and compute metrics like **accuracy**, **precision**, **recall**.

5. What thresholds or detection criteria minimize false positives while maximizing true positives in network anomaly detection?
   - **Dependent Variable**: Anomaly detection performance (e.g., accuracy, precision).

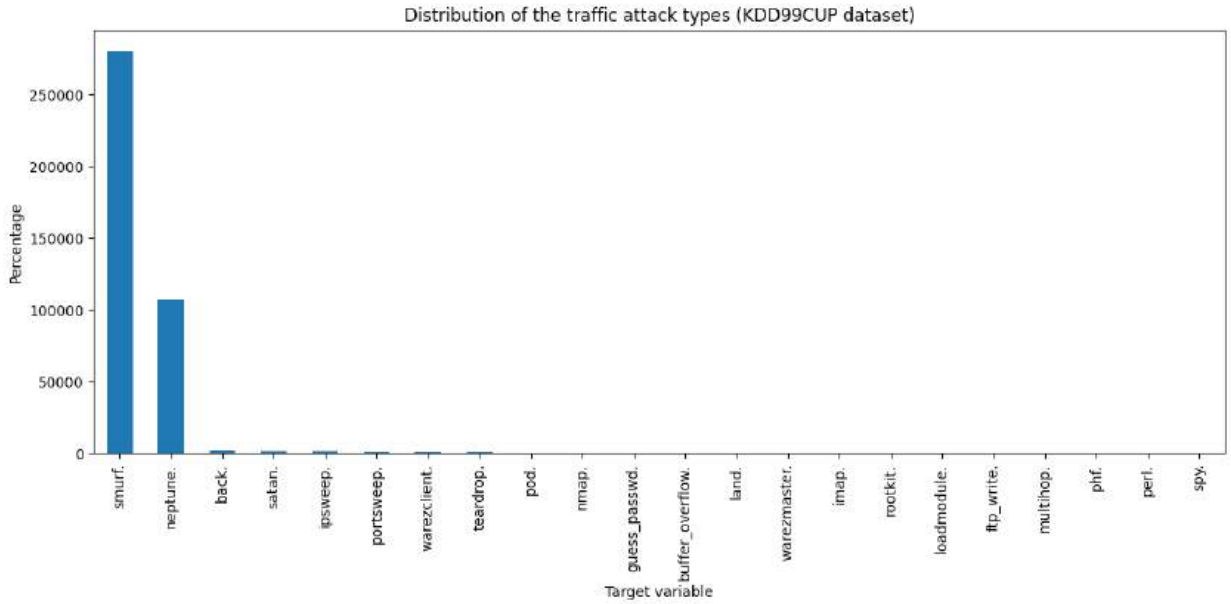
- **Independent Variables**: Threshold values for key features (`bytes transferred, packet duration`).
- **Approach:**
  - Use ROC-AUC to identify optimal thresholds for separating normal and anomalous traffic, and evaluate thresholds using confusion matrices.

6. Does the type of attack depend on the amount of incoming traffic?
   - **Dependent Variable**: Type of Attack (`DoS, Normal, Exploits, etc.`).
   - **Independent Variables**: Amount of Incoming Traffic (`sbytes`): The numeric variable representing the amount of source bytes transferred, which varies for different attacks.
   - **Approach:**
     - **ANOVA** (`Analysis of Variance`): Perform **ANOVA** to determine if mean `sbytes` values differ significantly across attack categories.

$$F = \frac{Between - Group\ Variance}{Within - Group\ Variance}$$

Statistical Questions

7. What types of anomalies are most common (e.g., denial-of-service attacks, data exfiltration attempts, unauthorized access) in the network traffic?
   - **Dependent Variable**: Frequency of anomaly type (e.g., DoS, probe, R2L, U2R).
   - **Independent Variables**: Traffic characteristics like `source IP, destination port`.
   - **Approach**:
     - Analyze anomaly type distribution in the dataset.
     - Based on KDD 1999 dataset, it shows that smurf (**DoS**) is the most common attack type in network traffic.

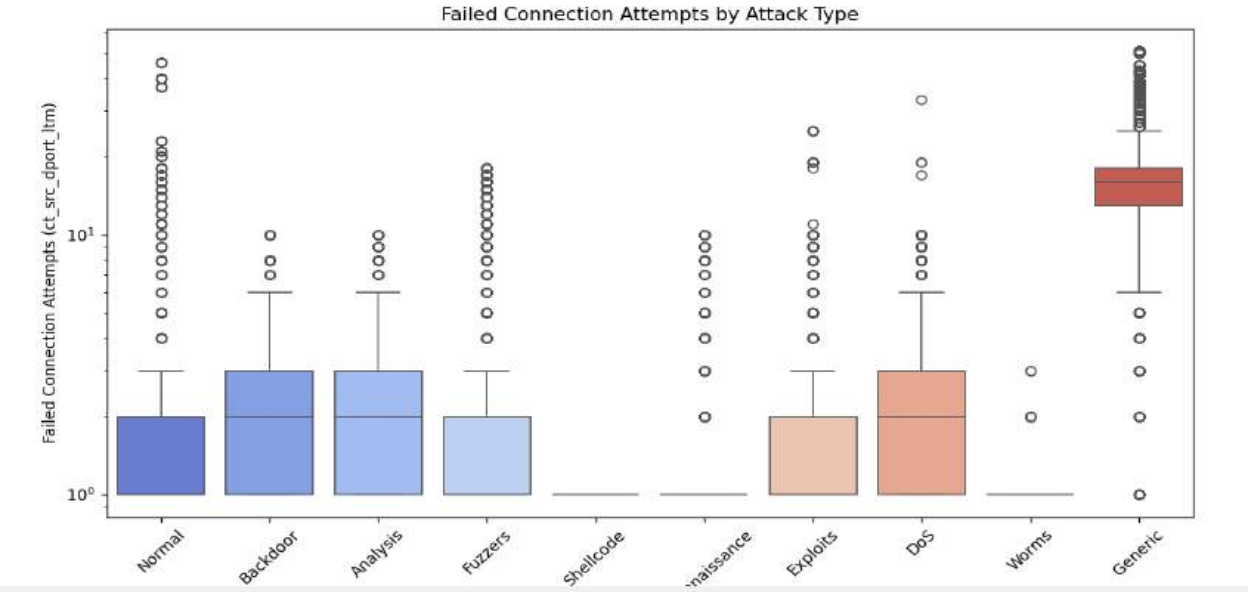


8. Are there correlations between the number of failed connection attempts and the type of attack?

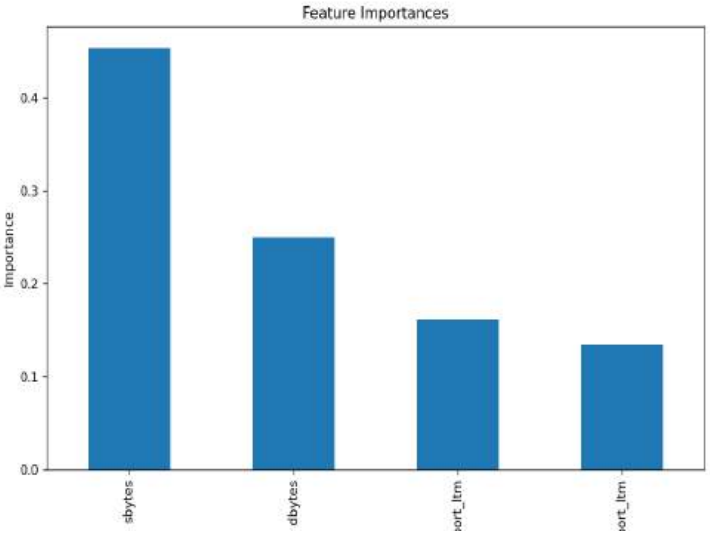
   - **Dependent Variable**: Number of Failed Connection Attempts (`ct_src_dport_ltm`): A numerical variable representing the number of failed connection attempts from the source.
   - **Independent Variables**: Attack Category: A categorical variable representing the type of attack.

   - **Approach:**

     - **ANOVA** (`Analysis of Variance`): ANOVA was used to test whether the mean number of failed connection attempts differs significantly across different attack categories.
       This method is suitable because it compares means across multiple groups (attack categories) while assessing whether the differences are statistically significant.



9. Can we predict the likelihood of an anomaly occurring based on past network activity?
    - **Dependent Variable**: Attack Category
    - **Independent Variables**:
        - Source Bytes (`sbytes`): Amount of bytes sent by the source.
        - Destination Bytes (`dbytes`): Amount of bytes received by the destination.
        - Packet Size (`pkt_size`): (if available) Represents the size of packets transmitted.
        - Flow Time (`dur`): Duration of the connection or network flow.
    - **Approach:**
        - **Random Forest Classifier**: Random Forest is a supervised machine learning algorithm used for classification. It builds multiple decision trees during training and aggregates their predictions to make a robust and accurate classification. This model is effective for handling datasets with mixed data types and can capture nonlinear relationships between the features and the target variable.



Datasets

KDD Cup 1999 Dataset

The KDD Cup 1999 data set is one of the most commonly used benchmarks for network intrusion detection research. It was derived from the 1998 DARPA Intrusion Detection Evaluation Program data and contains simulated network traffic data.

Timeframe: The data was collected over a period of **7 weeks in 1999,** The original data was generated by simulating a network environment with normal and attack traffic over those 7 weeks.

The data consists of 41 features (plus the label), which describe various attributes of a network connection, **including**:

- **Categorical features (symbolic):** e.g., `protocol_type`, `service`, `flag`.
- **Continuous features:** e.g., `src_bytes`, `dst_bytes`, `duration, etc`.

The **label** indicates whether the connection is normal or an attack, and if an attack, its type (e.g., `neptune`, `satan`).

1. Variables

- Basic Features:
  - Duration of the connection.
  - Protocol type (e.g., TCP, UDP, ICMP).

- Service type (e.g., HTTP, SMTP, FTP).
- Flag (e.g., SF, REJ).
- Number of bytes sent/received.

- Content Features:
  - Number of failed logins.
  - Root access attempts.
  - Number of file creation operations.
- Traffic Features:
  - Count of connections to the same host as the current connection within a specified time window.
  - Count of connections with the same service as the current connection within a specified time window.
- Target Variable:
  - Class label identifying normal or one of the specific attack types (e.g., DOS, R2L, U2R, probe).

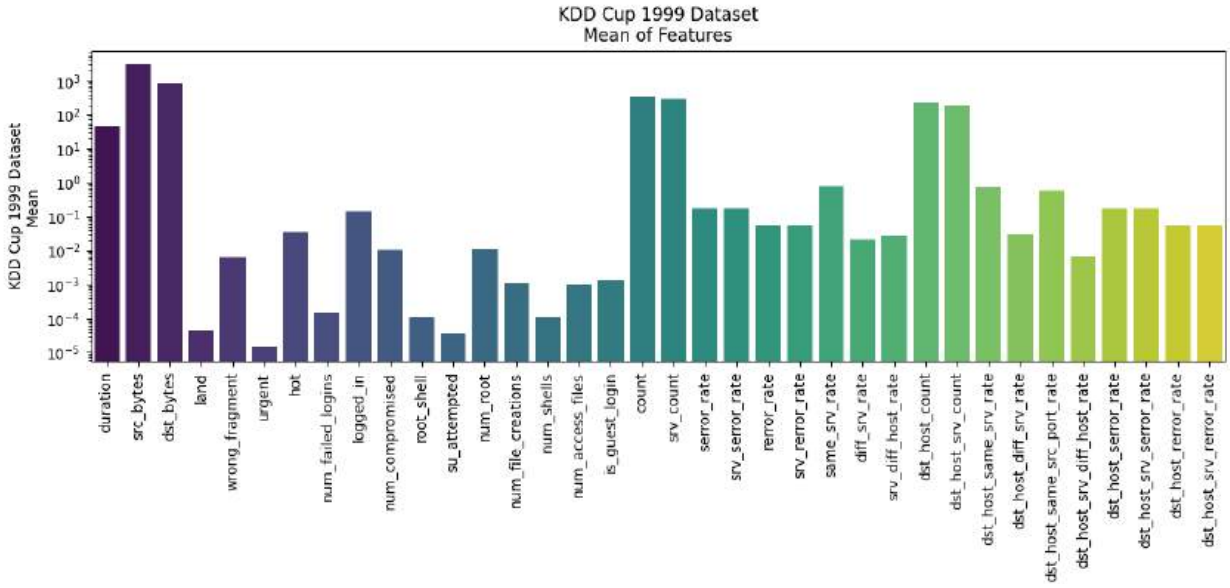
2. Sample Size

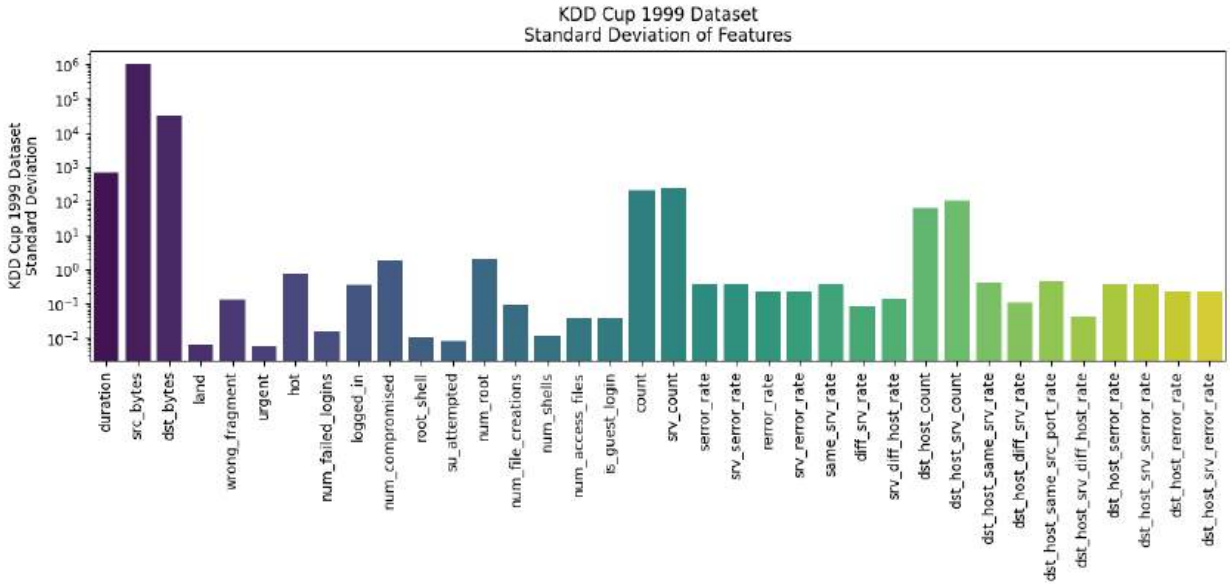
- The data set contains approximately 4.9 million connection records in the full training set.
- The 10% subset, which is often used for research, contains around 494,021 records.
- The records are labeled as either "normal" or one of 22 different types of attacks (e.g., Neptune, Smurf, Satan).

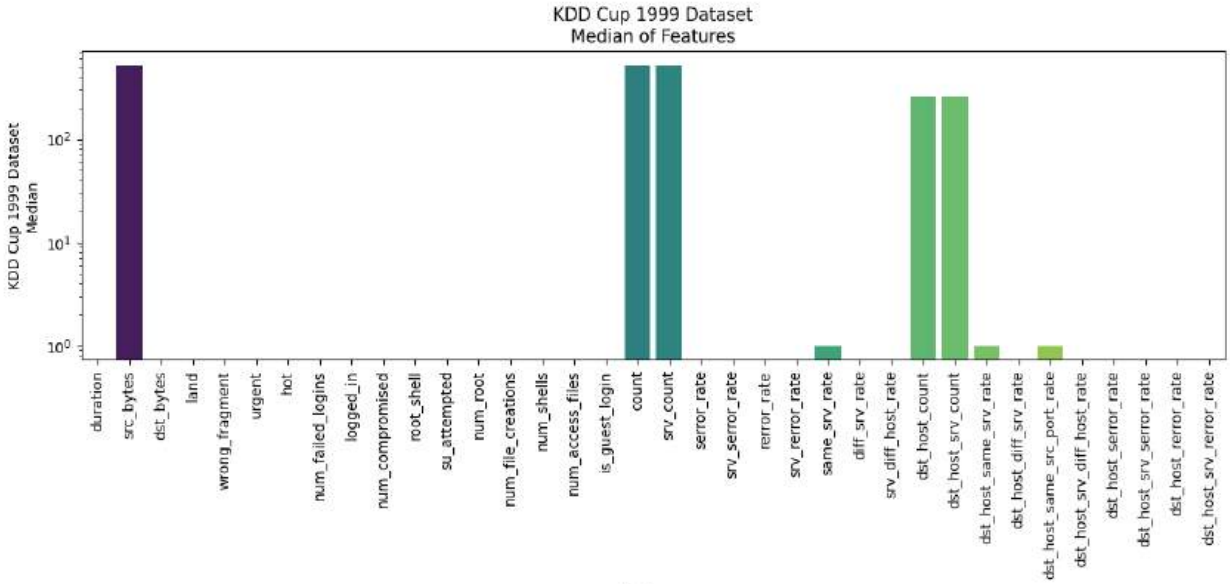
3. Categorization of Attacks

- **DOS** (Denial-of-Service): E.g., smurf, teardrop, back.
- **Probe**: E.g., portsweep, nmap.
- **R2L** (Remote-to-Local): E.g., guess_password, ftp_write.
- **U2R** (User-to-Root): E.g., buffer_overflow, rootkit.

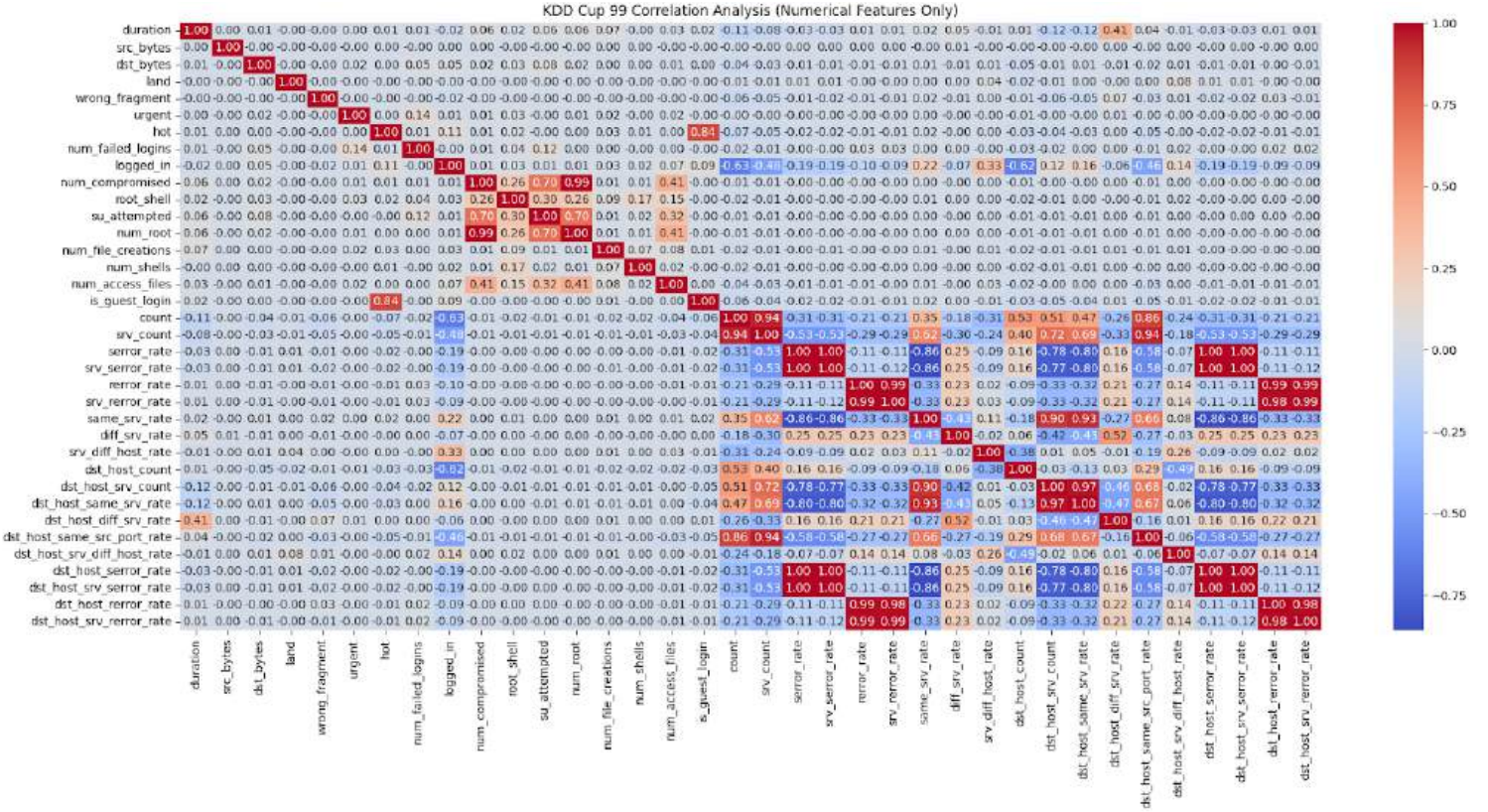
Central Tendency of Features (a 10% sample)







Correlation Heatmap of Features



UNSW-NB15 Dataset

The **UNSW-NB15 dataset** is a widely used benchmark for research in network intrusion detection. It was created by generating synthetic network traffic and incorporating modern attack scenarios, ensuring it reflects contemporary network behaviors. The dataset contains features that describe network activities and classifies them as either normal or specific attack types.

Timeframe: Created in **2015** using the **IXIA PerfectStorm tool** to generate a modern hybrid of real-world normal and attack traffic in a period of **(31 hours)**.

This dataset is relatively newer than the **KDD Cup 1999** one.
The **UNSW-NB15** was released in 2015 while the **KDD Cup 1999** was released in 1999

The data consists of (49 total) features :

- Includes attributes such as flow characteristics, basic features, content features, and time-based statistical features, (e.g., `srcip, sport, dstip, dsport, proto, state, dur, sbytes, dbytes, etc.`), **attack_cat** and **one label.**
    - A binary label indicating whether the record is **normal (0)** or **malicious (1)**.

1. Variables

- **Basic Features:**
  - Source and Destination IP Address: Identifiers for the originating and receiving endpoints of the network connection.
  - Source and Destination Port: Port numbers associated with the connection.
  - Protocol: Network protocol used (e.g., `TCP, UDP, ICMP`).
  - Packet and Byte Counts: Total number of packets and bytes transmitted in the connection.
- **Statistical Features**
  - Flow Rates: Metrics such as packets per second or bytes per second.
  - Percentage-Based Metrics: Proportions of packets exchanged in different directions or flags triggered.
- **Label Features :** The dataset provides a target variable that identifies whether the connection is:
  - Normal Traffic
  - Attack Traffic **(classified into multiple categories)**.
- **Time-Based Features**
  - Average Packet Size: Mean size of transmitted packets.
  - Interarrival Times: Time gap between consecutive packets in the connection.

2. Data Size and Format

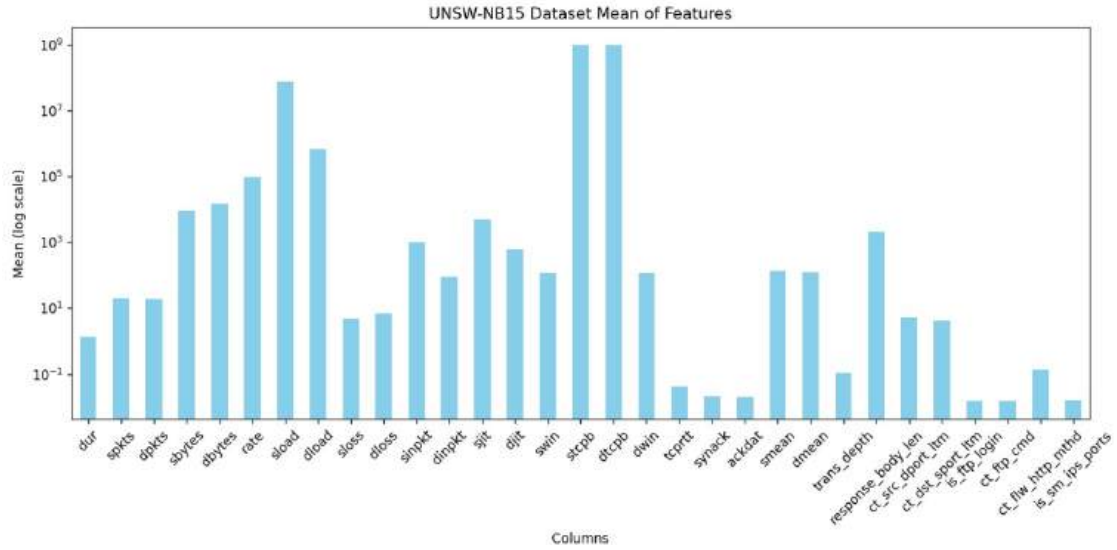
- The dataset contains 2,540,044 records, split across training and testing sets.
- The training set includes 175,341 records, while the testing set comprises 82,332 records.
- Each record has 49 features (plus the label).

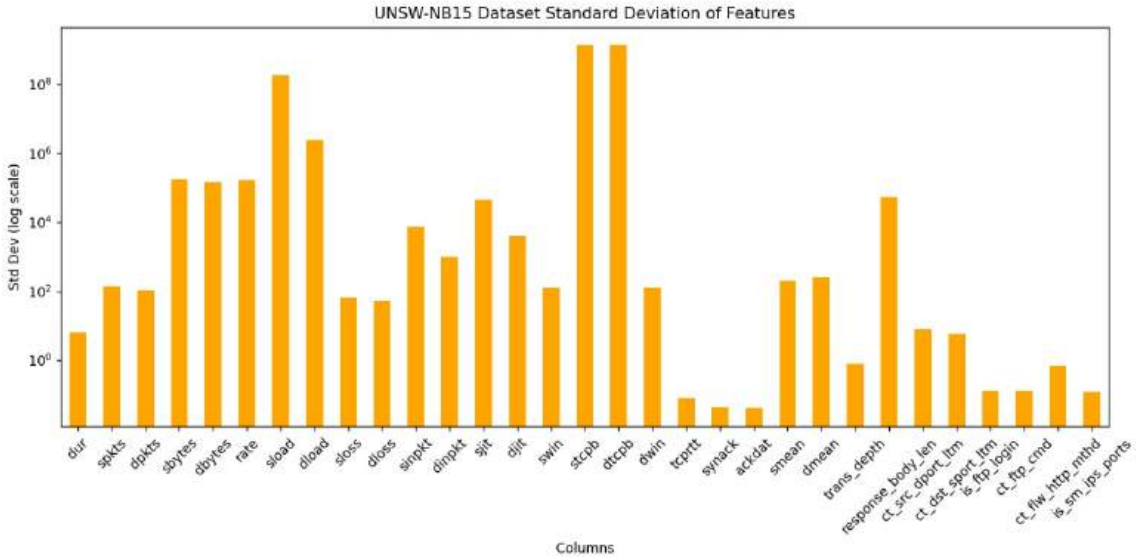
3. Categorization of Attacks

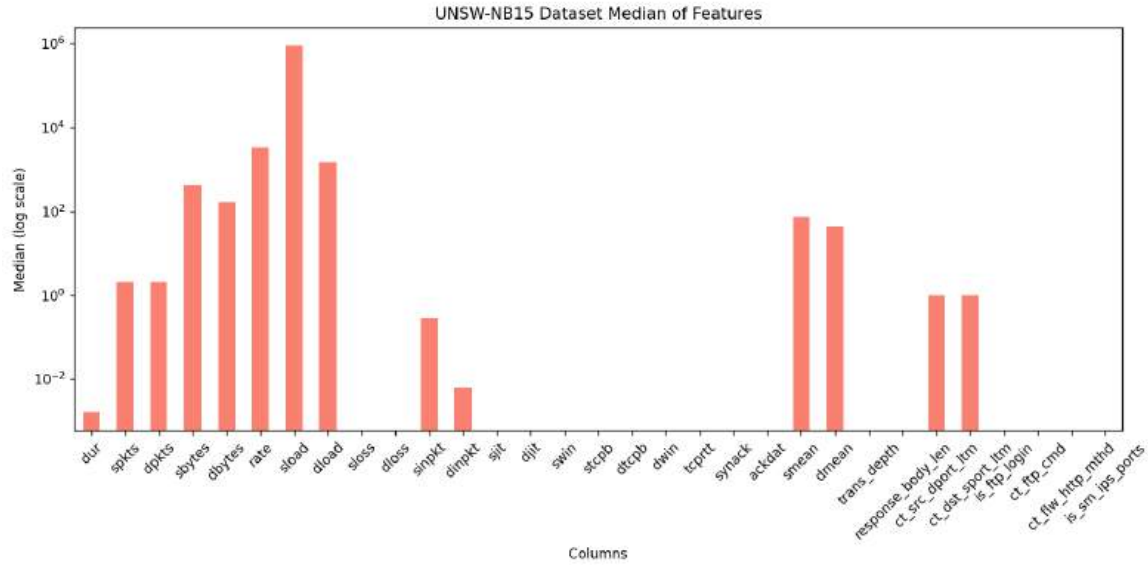
UNSW-NB15 categorizes attack traffic into nine distinct types:

- **Fuzzers**: Exploits software vulnerabilities by sending unexpected input.
- **Analysis**: E.g., probing and network reconnaissance activities.
- **DoS (Denial-of-Service)**: Disrupts services by overwhelming resources.
- **Backdoors**: Remote control mechanisms for unauthorized access.
- **Exploits**: Exploitation of software vulnerabilities.
- **Generic**: Attacks not specific to protocols or services.
- **Reconnaissance**: Scans or probes for network vulnerabilities.
- **Shellcode**: Malicious code executed as a command shell.
- **Worms**: Self-propagating malware over the network.

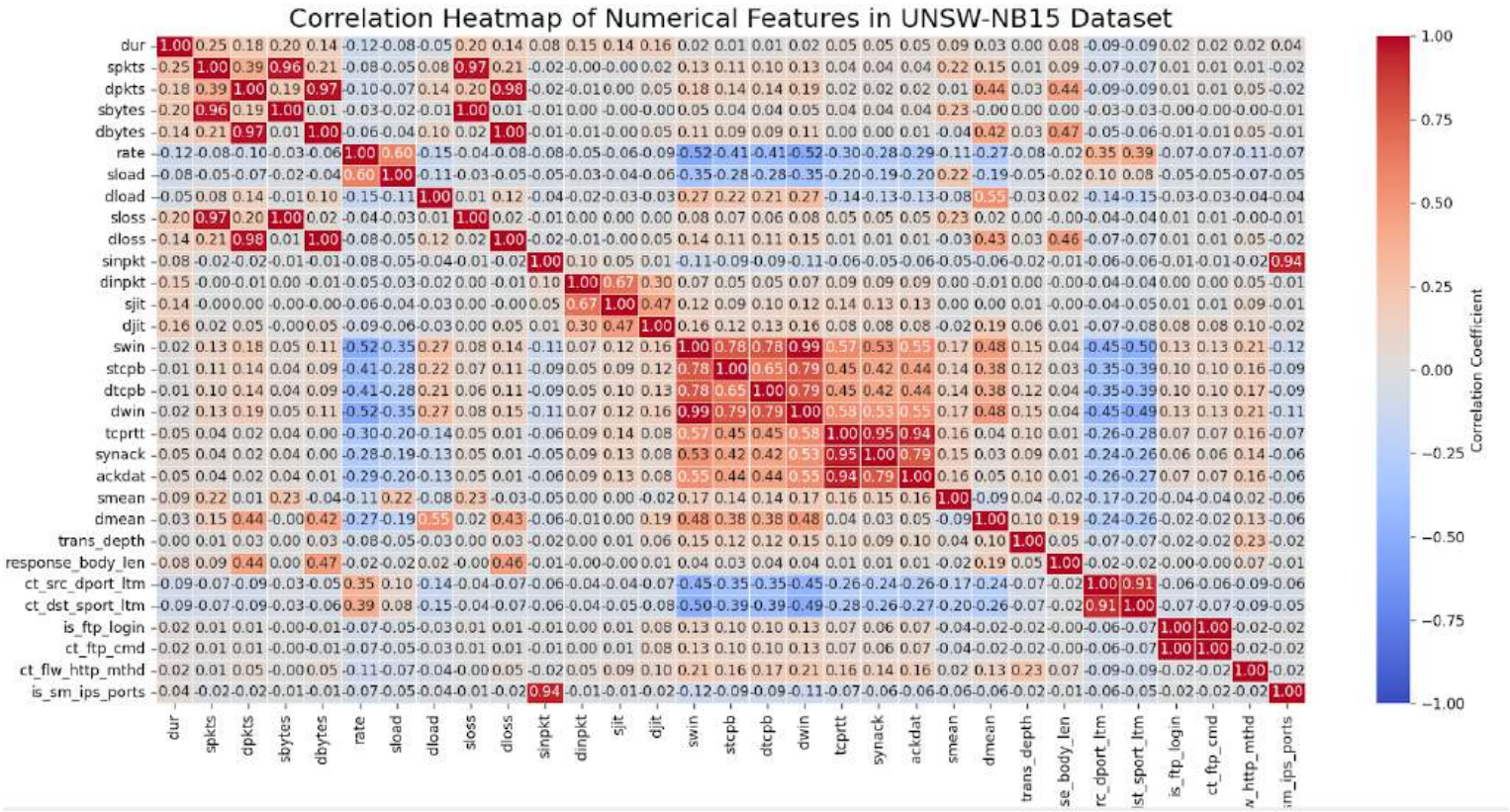
Central Tendency of Features and some statistical graphs (a 10% sample)





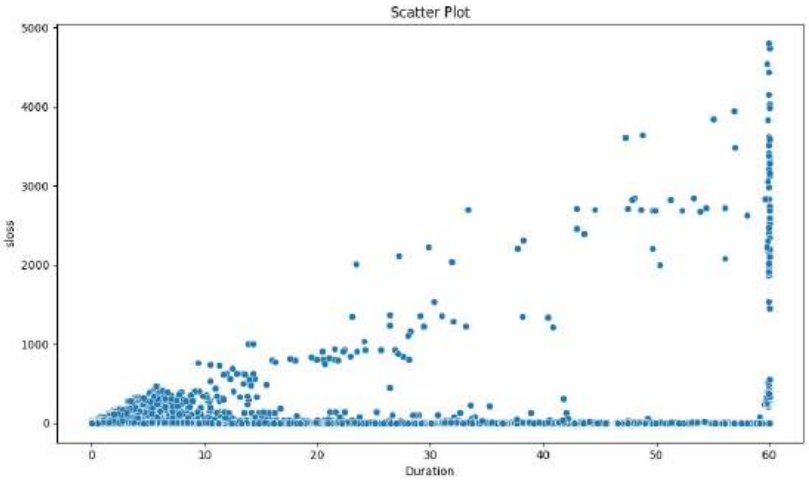


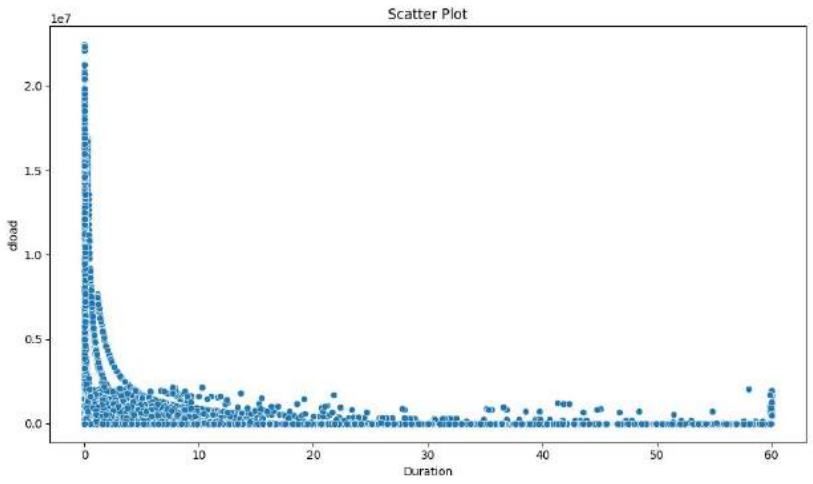
Correlation Heatmap of Features



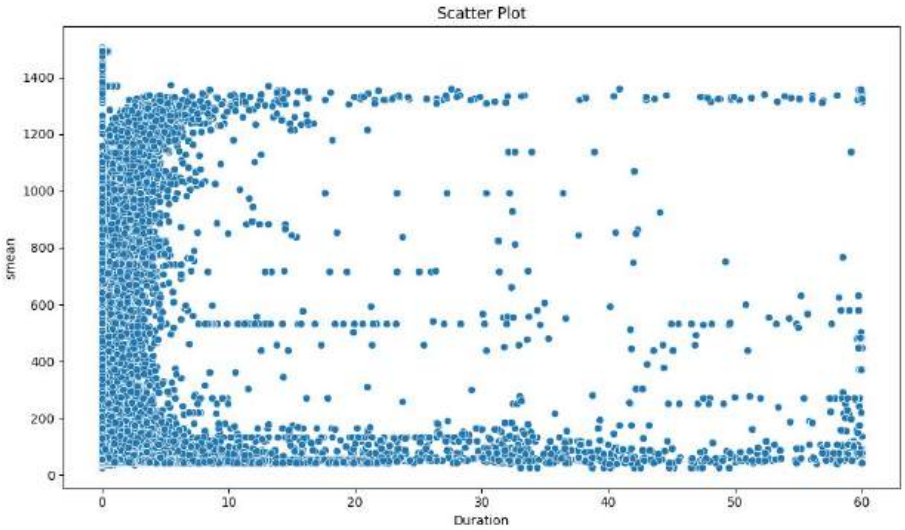
Investigating the relation between Duration & Features

Features: **sloss**, **dloss** (amount of **data lost or dropped** at the source/destination)

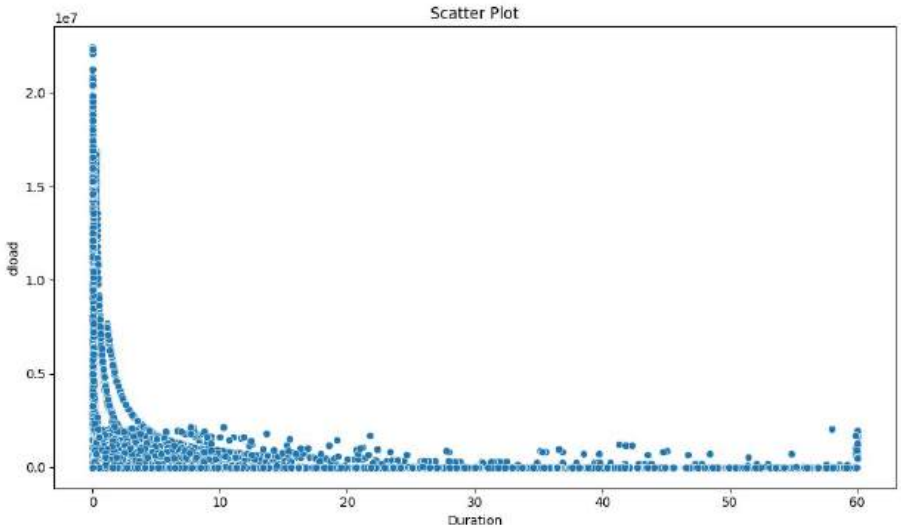




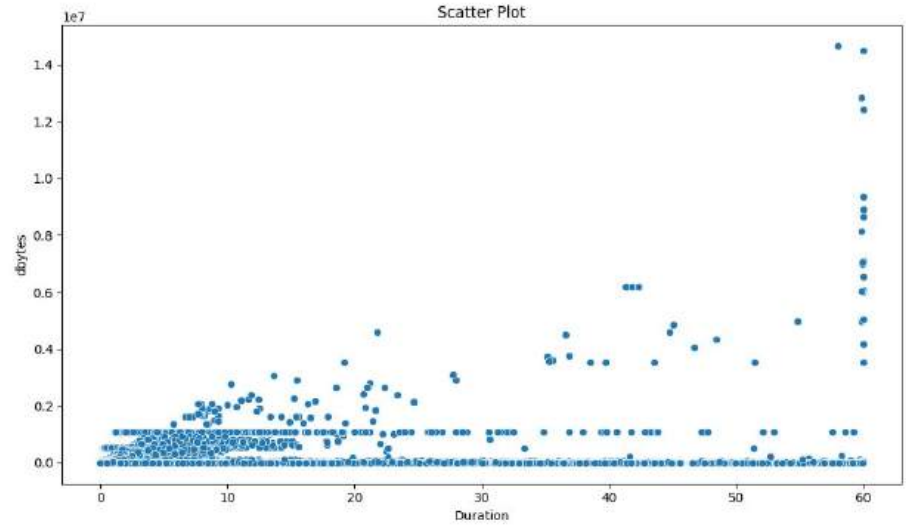
Feature: **smean** (the **mean size of packets** sent by the source)



Feature: **dload** (the **rate of data transfer** to the destination)



Feature: **dload** (the **rate of data transfer** to the destination)

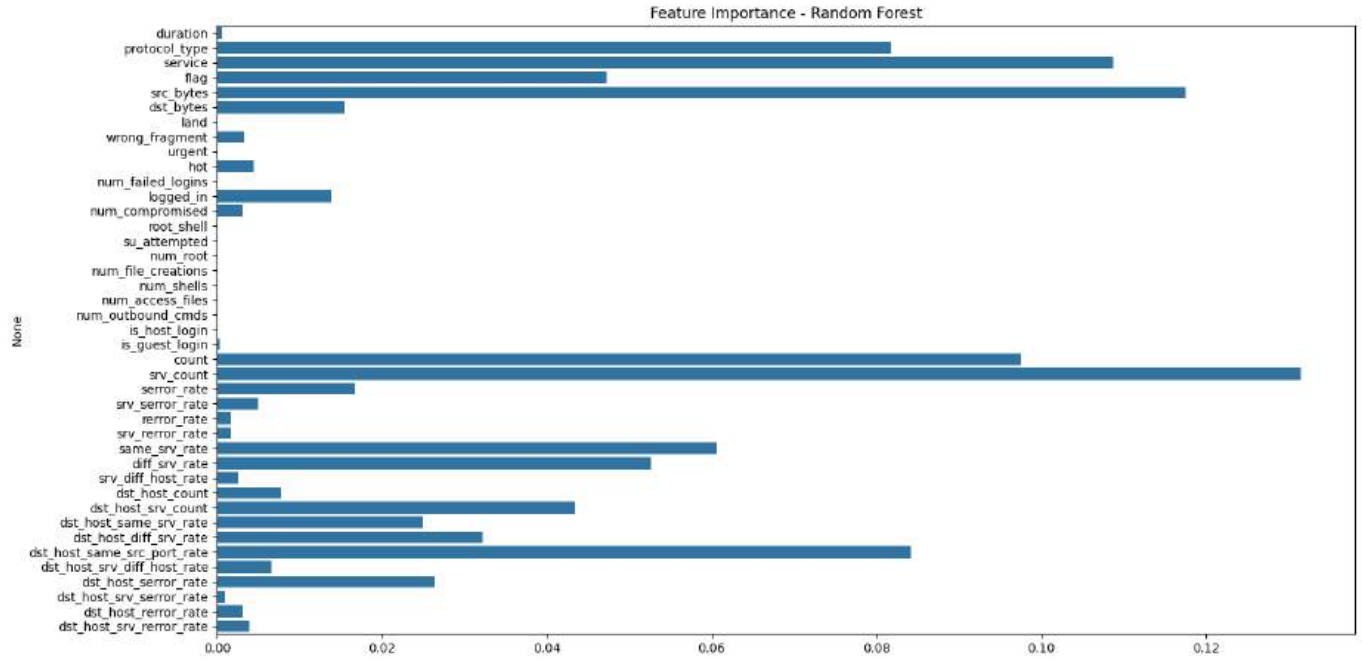


Results and Discussion

1. **KDD Cup 1999 Dataset**
    - **ANOVA Result:**
        - **F-statistic:** `25.93`, **P-value:** `1.5487e-106`
        - Significant differences in `src_bytes` across attack categories, indicating its relevance for anomaly detection.
    - **Chi-Square Test:**
        - **Chi-Square Statistic:** `564811.83`, **P-value:** `0.0`
        - Strong correlation between `protocol_type` and anomaly labels, highlighting `protocol_type` as a key feature.
    - **Random Forest Performance:**
        - **Accuracy:** `99.98%`, **Precision:** `99.98%`, **Recall:** `99.98%`, **F1-Score:** `99.97%`
        - The model demonstrated near-perfect performance, effectively classifying anomalies with minimal errors.

```
ANOVA Result: statistic= 25.932257171380268 P-value= 1.5487281408042723e-106
Chi-Square Test Result: 564811.8303096127 P-value: 0.0
Random Forest - Accuracy: 0.9997773369499211
Precision: 0.9997761402987123 Recall: 0.9997773369499211 F1 Score: 0.9997478166543043
```

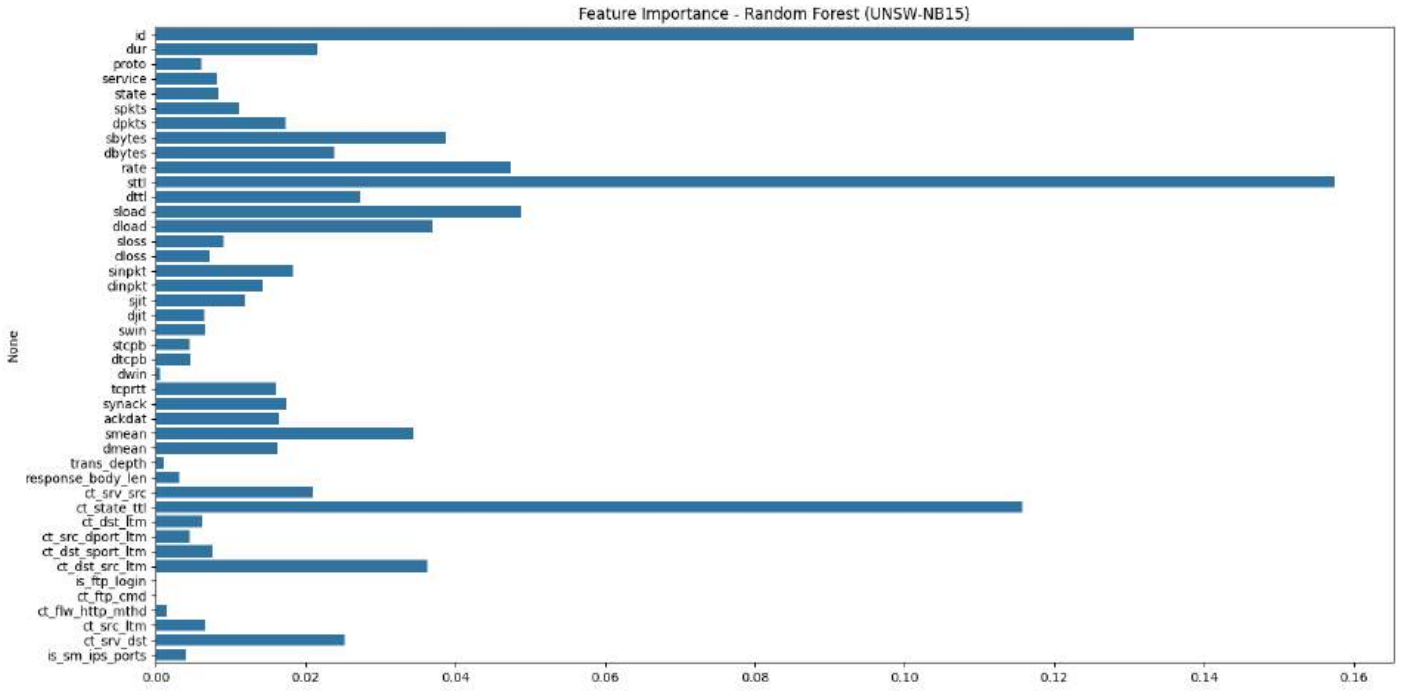
Explanation:
The KDD Cup 1999 dataset yielded exceptional results, with Random Forest achieving near-perfect classification accuracy. Features like `src_bytes` and `protocol_type` were highly predictive of anomalies.



2. **UNSW-NB15 Dataset**
    - **ANOVA Result:**
        - **F-statistic:** `96.78`, **P-value:** `7.83e-23`
        - Significant differences in `sbytes` across attack categories, underscoring its importance.
    - **Chi-Square Test:**
        - **Chi-Square Statistic:** `52144.47`, **P-value:** `0.0`
        - Strong correlation between `proto` (protocol type) and anomaly labels, confirming its predictive value.
    - **Random Forest Performance:**
        - **Accuracy:** `98.31%`, **Precision:** `98.31%`, **Recall:** `98.31%`, **F1-Score:** `98.31%`
        - The model achieved excellent classification performance, effectively identifying anomalies.

```
ANOVA Result: statistic= 96.77791650270673 P-value= 7.82787548141833e-23
Chi-Square Test Result: 52144.47214314193 P-value: 0.0
Random Forest - Accuracy: 0.9830922873923055
Precision: 0.983092684411472 Recall: 0.9830922873923055 F1 Score: 0.9830924843027471
```

Explanation:
The UNSW-NB15 dataset results also demonstrated high performance, with Random Forest achieving over 98% accuracy. Features like `sbytes` and `proto` were identified as critical for anomaly detection.



Challenges

1. **Dataset Limitations**
    a. **Redundancy**
    - Approximately 78% of records in the training set are duplicates, leading to bias and overfitting during model training.
        - **Impact**: Models may focus on memorizing duplicate patterns rather than learning generalizable features.
        - **Solution**: Remove duplicate entries to ensure a more diverse dataset.

    b. **Feature Relevance**
    - Some features are irrelevant, redundant, or require domain-specific expertise to interpret effectively.
        - **Impact**: Irrelevant features can dilute the model's performance.
        - **Solution**: Use feature selection techniques to identify and retain only important features.

2. **Challenges in Random Forest Training**
    a. **High Dimensionality**
    - The dataset includes 41 features, many of which are categorical or redundant.
        - **Impact**: Increases computational complexity and can lead to model inefficiency.
        - **Solution**: Preprocess the data by encoding categorical variables, normalizing numerical features, and using feature selection methods.

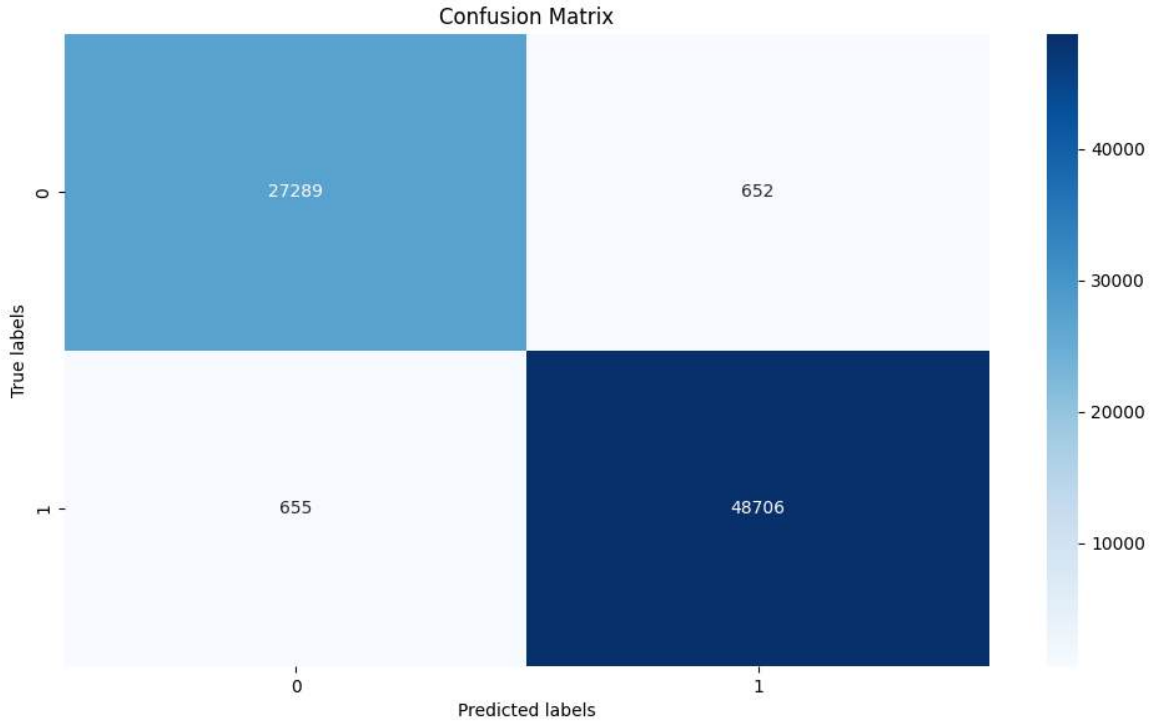
    b. **Categorical Variables**
    - Categorical features (e.g., "`protocol_type`", "`service`", "`flag`") require preprocessing to make them usable by Random Forests.
        - **Impact**: Encoding can introduce additional challenges, such as high cardinality in "service".
        - **Solution**: Use one-hot encoding or target encoding, but be mindful of data dimensionality.

Conclusion

This research successfully developed anomaly detection models using statistical and machine learning techniques, achieving high accuracy and reliability across two prominent datasets. The models effectively identified significant features such as `sbytes` and `proto` for **UNSW-NB15** and `src_bytes` and `protocol_type` for **KDD Cup 1999.**

The findings demonstrate the potential of data-driven approaches to enhance network security and provide early warnings for cyber threats.

- **Key Findings and Model Performance**
  - Random Forest classifiers achieved **99.98% accuracy on KDD Cup 1999** and **98.31% on UNSW-NB15**, confirming their effectiveness in anomaly detection.
  - Features such as **source bytes (`sbytes`)** and **protocol type (`proto`)** in UNSW-NB15, and **destination host count (`dst_host_count`)** in KDD, were identified as highly predictive of anomalies.
- **Significance of Statistical Analysis**
  - Statistical tests like **ANOVA** and **Chi-Square** highlighted significant associations between features (e.g., protocol types) and anomaly patterns, reinforcing their relevance for classification.
- **Challenges Identified**
  - Redundancy in datasets (e.g., 78% duplicates in KDD Cup 1999) posed risks of overfitting and reduced generalizability.
  - High-dimensional features required preprocessing and feature selection to ensure efficient and accurate modeling.
- **Contribution to Network Security**
  - Demonstrated the viability of **real-time anomaly detection systems** using machine learning, offering early warning capabilities to mitigate cyber threats.
  - Provided actionable insights for optimizing detection systems, including feature prioritization and dataset preparation.
- **Future Directions**
  - Incorporate more diverse datasets and real-time traffic data for improved robustness.
  - Explore advanced ensemble methods and adaptive learning to address evolving cyber threats.



References

1. KDD Cup 1999 Dataset
2. Machine Learning Approaches to Network Anomaly Detection
3. Feature Selection in UNSW-NB15 and KDDCUP'99 Datasets
4. An Analysis of the KDD99 and UNSW-NB15 Datasets for the Intrusion Detection System
5. Anomaly Detection in Network Traffic Using Machine Learning for Early Threat Detection
6. K. Lu, “Network Anomaly Traffic Analysis,” Academic Journal of Science and Technology, vol. 10, no. 3, pp. 65–68, Apr. 2024, doi
7. Anomaly Detection with Various Machine Learning Classification Techniques
8. M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, Network Traffic Anomaly Detection and Prevention : Concepts, Techniques, and Tools. Cham: Springer International Publishing, 2017.
9. D. K. Bhattacharyya, Network anomaly detection : a machine learning perspective. Chapman And Hall/Crc, 2013.