Student Name: Carlos Nakagomi

Instructor: Quynh Nguyen

Langara College

Cyber Security Case Study

CodeMax

*Maximizing Security. Powering Intelligence.*

A blue circle with a lock and a key

Description automatically generated

Vision

To revolutionize cybersecurity by leveraging AI-driven threat detection, ensuring a safer digital world where organizations can operate without fear of cyber threats.

Mission

To empower businesses and institutions with cutting-edge AI-powered Intrusion Detection Systems (IDS) that provide real-time threat prediction, anomaly detection, and predictive security insights—enabling proactive cybersecurity defenses against evolving threats.

Focus

* Proactive Threat Detection – Delivering real-time security solutions that identify and mitigate cyber threats before they cause harm.
* AI & Machine Learning Innovation – Utilizing advanced data analytics and ML algorithms to continuously enhance intrusion detection capabilities.
* Enterprise & Institutional Security – Providing scalable and adaptable security solutions tailored for universities, enterprises, and government institutions.
* Compliance & Risk Mitigation – Ensuring organizations meet cybersecurity standards while reducing operational risks associated with cyberattacks.

Value Proposition

We offer a next-generation AI-powered Intrusion Detection System (AI-IDS) that delivers:

* Real-Time Threat Prediction – Detect and stop cyberattacks before they escalate.
* Advanced Anomaly Detection – Uncover hidden threats through behavioral analysis.
* Predictive Security Insights – Identify and mitigate vulnerabilities proactively.
* Scalability & Efficiency – AI-driven automation reduces security team workloads and enhances response time.
* Proven Accuracy & Adaptability – Built using industry-leading datasets (CICIDS2017) to stay ahead of evolving threats.

Our AI-driven cybersecurity solutions ensure that organizations remain secure, compliant, and resilient against modern cyber threats.

1. Institutional Need

Queensland University faces increasing cybersecurity threats, including Denial-of-Service (DoS), phishing, and malware attacks, which jeopardize sensitive research data and disrupt critical operations. A traditional rule-based Intrusion Detection System (IDS) is insufficient to combat sophisticated and evolving cyber threats. To address this issue, we propose a machine learning-driven IDS that leverages advanced algorithms to detect and respond to network intrusions in real time.

Key Reasons for Implementation:

* Enhancing Cybersecurity Resilience: Machine learning-based IDS can proactively detect new and evolving cyber threats, reducing the risk of data breaches and operational disruptions.
* Real-Time Threat Detection: Unlike traditional systems, which rely on predefined signatures, an AI-powered IDS continuously learns from network traffic patterns to identify anomalies and suspicious activities.
* Automated Incident Response: The system minimizes the burden on cybersecurity teams by automating threat detection and providing actionable insights, enabling faster response times.
* Scalability & Adaptability: Machine learning models can adapt to changing network environments, making the IDS more effective in protecting research data and university infrastructure.
* Regulatory Compliance: Implementing an advanced IDS ensures compliance with cybersecurity regulations and best practices, reinforcing the university’s commitment to data protection.
* Cost-Effective Security Enhancement: By reducing manual intervention and improving detection accuracy, a machine learning-based IDS lowers operational costs associated with security breaches and forensic investigations.

1. Description of Service

Product Name: AI-Powered Intrusion Detection System (AI-IDS)

The proposed solution is an application-specific machine learning-based Intrusion Detection System (IDS) designed to safeguard Queensland University's network from cyber threats. It is built using the CICIDS2017 dataset, ensuring comprehensive coverage of modern attack patterns. The system enhances cybersecurity by detecting anomalies and mitigating threats in real-time.

Core Features of AI-Powered Intrusion Detection System (AI-IDS)

* Real-time Threat Prediction: The system continuously monitors network traffic, user behavior, and system logs to detect potential attacks before they escalate. By identifying threats in their early stages, it enables immediate response to prevent damage.
* Anomaly Detection: AI-IDS identifies unusual activity patterns that differ from normal behavior, such as unauthorized file access, sudden network spikes, or suspicious login attempts. These anomalies are flagged for further investigation to prevent security breaches.
* Predictive Security Insights: By analyzing past attack data and current security settings, the system assesses overall cybersecurity readiness and predicts potential vulnerabilities before they can be exploited.

1. Context and Background

Queensland University is one of the leading educational and research institutions, and thus, it controls sensitive data and critical services, which makes it a potential target for various cyberattacks, including DoS, phishing, and malware. Traditional rule-based IDS faces inefficiencies due to evolving threats, such as high false positive rates.

It will, therefore, develop a machine learning-based IDS tuned for the university's network based on publicly available datasets and anonymized internal traffic. It shall detect malicious activities with high accuracy, adapt to unique network patterns, assure data protection, continuity of service, and conformance to cybersecurity standards.

1. In-Depth Review of Machine Learning Solutions for Intrusion Detection Systems (IDS)

This section explores three ML solutions that address the institutional need for detecting cyberattacks at Queensland University. These solutions focus on handling imbalanced datasets and optimizing performance, drawing on academic literature, research studies, and expert analyses.

* 1. Intrusion Detection by RF Overview

Random Forest is a decision-tree-based ensemble learning model that will give robust and high-performance results in intrusion detection. In RF, multiple trees are built, and all the trees vote through mechanisms for target class final predictions. Hence, it proves to be efficient in both normal and malicious network traffic detection.

In order to handle class imbalance in the datasets, RF uses methods such as class weighting to give more importance to minority attack classes. Preprocessing methods such as SMOTE for oversampling further balance the data. Hyperparameter tuning and feature importance analysis enhance RF's performance and efficiency, enabling it to handle large, high-dimensional datasets with ease.

Various research, such as that by Yaseen and Shamsuddin (2018), presents the high accuracy of RF for detecting attacks like DoS and U2R; the performance regarding computational intensity may challenge real-world applications, especially in wide-scale environments, hence optimization or hybrid methods are required.

* 1. Gradient Boosting Machines overview

Gradient Boosting Machines (GBM) is a tree-based ensemble learning model that builds trees sequentially, with each one correcting the errors of the previous one. Unlike models that construct trees independently, GBM’s boosting approach enhances predictive accuracy, making it highly effective for intrusion detection.

Research by Li et al. (2020) highlights GBM’s high detection rates for attacks like DoS and R2L, often outperforming traditional models. However, given the dataset's size, optimized versions such as XGBoost or LightGBM can be applied to accelerate computation and improve efficiency.

In cybersecurity, GBM is particularly useful for **detecting zero-day attacks and evolving threat patterns**, as its iterative learning process adapts to complex attack behaviors. It is widely used in **network intrusion detection systems (NIDS), malware classification, and fraud detection**, offering robust performance in distinguishing between normal and malicious traffic.

* 1. XGBOOST overview

XGBoost (Extreme Gradient Boosting) is a powerful, tree-based ensemble learning algorithm known for its efficiency and high performance in machine learning tasks, including cybersecurity and intrusion detection. It is an optimized implementation of gradient boosting that enhances speed and accuracy through techniques such as regularization, parallel processing, and handling of missing values.

XGBoost is particularly effective in detecting malicious network traffic due to its ability to handle high-dimensional data, manage imbalanced classes using weighted loss functions, and prevent overfitting through L1 (Lasso) and L2 (Ridge) regularization. The model’s feature importance analysis helps in identifying key network attributes that influence attack predictions, making it a valuable tool for security analysts.

Various studies, including those by Chen and Guestrin (2016), demonstrate XGBoost’s superior accuracy in detecting sophisticated cyberattacks like DDoS and probing. However, its computational complexity can be a challenge, especially when applied to large-scale intrusion detection systems. Techniques such as hyperparameter tuning and distributed training can optimize its performance, ensuring its applicability in real-world cybersecurity environments.

1. Model selected

This project proposes a Random Forest (RF)-based IDS, an ensemble learning approach known for its high accuracy, scalability, and robustness against overfitting. RF builds multiple decision trees and aggregates their votes, effectively distinguishing between normal and malicious network traffic.

To address class imbalance, SMOTE oversampling and class weighting will ensure rare attack types are properly detected. RF also provides feature importance analysis, optimizing model performance by identifying key network traffic indicators. Its efficiency in handling large datasets makes it ideal for university-wide security applications.

The implementation involves data preprocessing, feature selection, handling missing values, and outlier management. The model will be trained on CICIDS2017, optimized using GridSearchCV, and validated through cross-validation. Performance will be evaluated using precision, recall, F1-score, and AUC-ROC to minimize false positives and negatives.

By deploying a Random Forest-based IDS, Queensland University will achieve real-time cyber threat detection, improved security automation, compliance with cybersecurity standards, and a scalable security framework. This system will enhance resilience against evolving threats, ensuring a proactive and intelligent defense mechanism.

1. Table of feature descriptions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Description | Data Type | Units | Possible Values/Range |
| flow\_id | Unique identifier for each network flow | String | N/A | N/A |
| timestamp | Exact time when the flow was recorded | Datetime | Seconds | N/A |
| src\_ip | Source IP address | String | N/A | N/A |
| src\_port | Source port number | Integer | Port Number | 1-65535 |
| dst\_ip | Destination IP address | String | N/A | N/A |
| dst\_port | Destination port number | Integer | Port Number | 1-65535 |
| protocol | Protocol used in the flow (e.g., TCP, UDP) | String | N/A | TCP, UDP |
| duration | Total duration of the flow in seconds | Float | Seconds | N/A |
| packets\_count | Total number of packets transmitted in the flow | Integer | Packets | N/A |
| fwd\_packets\_count | Number of packets sent in the forward direction | Integer | Packets | N/A |
| bwd\_packets\_count | Number of packets sent in the backward direction | Integer | Packets | N/A |
| total\_payload\_bytes | Total payload bytes transmitted in the flow | Integer | Bytes | N/A |
| fwd\_total\_payload\_bytes | Total payload bytes sent in the forward direction | Integer | Bytes | N/A |
| bwd\_total\_payload\_bytes | Total payload bytes sent in the backward direction | Integer | Bytes | N/A |
| payload\_bytes\_max | Maximum payload size in bytes | Integer | Bytes | N/A |
| payload\_bytes\_min | Minimum payload size in bytes | Integer | Bytes | N/A |
| payload\_bytes\_mean | Mean payload size in bytes | Float | Bytes | N/A |
| payload\_bytes\_std | Standard deviation of payload size in bytes | Float | Bytes | N/A |
| payload\_bytes\_variance | Variance of payload size in bytes | Float | Bytes | N/A |
| fwd\_payload\_bytes\_max | Maximum payload size in the forward direction | Integer | Bytes | N/A |
| fwd\_payload\_bytes\_min | Minimum payload size in the forward direction | Integer | Bytes | N/A |
| fwd\_payload\_bytes\_mean | Mean payload size in the forward direction | Float | Bytes | N/A |
| fwd\_payload\_bytes\_std | Standard deviation of payload size in the forward direction | Float | Bytes | N/A |
| fwd\_payload\_bytes\_variance | Variance of payload size in the forward direction | Float | Bytes | N/A |
| bwd\_payload\_bytes\_max | Maximum payload size in the backward direction | Integer | Bytes | N/A |
| bwd\_payload\_bytes\_min | Minimum payload size in the backward direction | Integer | Bytes | N/A |
| bwd\_payload\_bytes\_mean | Mean payload size in the backward direction | Float | Bytes | N/A |
| bwd\_payload\_bytes\_std | Standard deviation of payload size in the backward direction | Float | Bytes | N/A |
| bwd\_payload\_bytes\_variance | Variance of payload size in the backward direction | Float | Bytes | N/A |
| total\_header\_bytes | Total header bytes transmitted in the flow | Integer | Bytes | N/A |
| max\_header\_bytes | Maximum size of header bytes in a single packet | Integer | Bytes | N/A |
| min\_header\_bytes | Minimum size of header bytes in a single packet | Integer | Bytes | N/A |
| mean\_header\_bytes | Mean size of header bytes per packet | Float | Bytes | N/A |
| std\_header\_bytes | Standard deviation of header bytes per packet | Float | Bytes | N/A |
| fwd\_total\_header\_bytes | Total header bytes transmitted in the forward direction | Integer | Bytes | N/A |
| fwd\_max\_header\_bytes | Maximum header size in the forward direction | Integer | Bytes | N/A |
| fwd\_min\_header\_bytes | Minimum header size in the forward direction | Integer | Bytes | N/A |
| fwd\_mean\_header\_bytes | Mean header size in the forward direction | Float | Bytes | N/A |
| fwd\_std\_header\_bytes | Standard deviation of header size in the forward direction | Float | Bytes | N/A |
| bwd\_total\_header\_bytes | Total header bytes transmitted in the backward direction | Integer | Bytes | N/A |
| bwd\_max\_header\_bytes | Maximum header size in the backward direction | Integer | Bytes | N/A |
| bwd\_min\_header\_bytes | Minimum header size in the backward direction | Integer | Bytes | N/A |
| bwd\_mean\_header\_bytes | Mean header size in the backward direction | Float | Bytes | N/A |
| bwd\_std\_header\_bytes | Standard deviation of header size in the backward direction | Float | Bytes | N/A |
| fwd\_avg\_segment\_size | Average size of TCP segments in the forward direction | Float | Bytes | N/A |
| bwd\_avg\_segment\_size | Average size of TCP segments in the backward direction | Float | Bytes | N/A |
| avg\_segment\_size | Average size of TCP segments in both directions | Float | Bytes | N/A |
| fwd\_init\_win\_bytes | Initial window size in the forward direction | Integer | Bytes | N/A |
| bwd\_init\_win\_bytes | Initial window size in the backward direction | Integer | Bytes | N/A |
| active\_min | Minimum time a flow is active | Float | Seconds | N/A |
| active\_max | Maximum time a flow is active | Float | Seconds | N/A |
| active\_mean | Mean time a flow is active | Float | Seconds | N/A |
| active\_std | Standard deviation of active flow times | Float | Seconds | N/A |
| idle\_min | Minimum idle time between packets | Float | Seconds | N/A |
| idle\_max | Maximum idle time between packets | Float | Seconds | N/A |
| idle\_mean | Mean idle time between packets | Float | Seconds | N/A |
| idle\_std | Standard deviation of idle time between packets | Float | Seconds | N/A |
| bytes\_rate | Rate of bytes transmitted per second | Float | Bytes per second | N/A |
| fwd\_bytes\_rate | Rate of forward bytes transmitted per second | Float | Bytes per second | N/A |
| bwd\_bytes\_rate | Rate of backward bytes transmitted per second | Float | Bytes per second | N/A |
| packets\_rate | Rate of packets transmitted per second | Float | Packets per second | N/A |
| fwd\_packets\_rate | Rate of forward packets transmitted per second | Float | Packets per second | N/A |
| bwd\_packets\_rate | Rate of backward packets transmitted per second | Float | Packets per second | N/A |
| down\_up\_rate | Ratio of downloaded to uploaded packets | Float | Ratio | N/A |
| avg\_fwd\_bytes\_per\_bulk | Average forward bulk bytes per flow | Float | Bytes | N/A |
| avg\_fwd\_packets\_per\_bulk | Average forward bulk packets per flow | Float | Packets | N/A |
| avg\_fwd\_bulk\_rate | Average forward bulk rate | Float | Bytes per second | N/A |
| avg\_bwd\_bytes\_per\_bulk | Average backward bulk bytes per flow | Float | Bytes | N/A |
| avg\_bwd\_packets\_bulk\_rate | Average backward bulk packets rate | Float | Packets per second | N/A |
| avg\_bwd\_bulk\_rate | Average backward bulk rate | Float | Bytes per second | N/A |
| fwd\_bulk\_state\_count | Number of forward bulk states | Integer | States | N/A |
| fwd\_bulk\_total\_size | Total forward bulk size | Integer | Bytes | N/A |
| fwd\_bulk\_per\_packet | Forward bulk size per packet | Float | Bytes | N/A |
| fwd\_bulk\_duration | Duration of forward bulk | Float | Seconds | N/A |
| bwd\_bulk\_state\_count | Number of backward bulk states | Integer | States | N/A |
| bwd\_bulk\_total\_size | Total backward bulk size | Integer | Bytes | N/A |
| bwd\_bulk\_per\_packet | Backward bulk size per packet | Float | Bytes | N/A |
| bwd\_bulk\_duration | Duration of backward bulk | Float | Seconds | N/A |
| fin\_flag\_counts | Number of FIN flags in the flow | Integer | Flags | 0-255 |
| psh\_flag\_counts | Number of PSH flags in the flow | Integer | Flags | 0-255 |
| urg\_flag\_counts | Number of URG flags in the flow | Integer | Flags | 0-255 |
| ece\_flag\_counts | Number of ECE flags in the flow | Integer | Flags | 0-255 |
| syn\_flag\_counts | Number of SYN flags in the flow | Integer | Flags | 0-255 |
| ack\_flag\_counts | Number of ACK flags in the flow | Integer | Flags | 0-255 |
| cwr\_flag\_counts | Number of CWR flags in the flow | Integer | Flags | 0-255 |
| rst\_flag\_counts | Number of RST flags in the flow | Integer | Flags | 0-255 |
| fwd\_fin\_flag\_counts | Number of forward FIN flags in the flow | Integer | Flags | 0-255 |
| fwd\_psh\_flag\_counts | Number of forward PSH flags in the flow | Integer | Flags | 0-255 |
| fwd\_urg\_flag\_counts | Number of forward URG flags in the flow | Integer | Flags | 0-255 |
| fwd\_ece\_flag\_counts | Number of forward ECE flags in the flow | Integer | Flags | 0-255 |
| fwd\_syn\_flag\_counts | Number of forward SYN flags in the flow | Integer | Flags | 0-255 |
| fwd\_ack\_flag\_counts | Number of forward ACK flags in the flow | Integer | Flags | 0-255 |
| fwd\_cwr\_flag\_counts | Number of forward CWR flags in the flow | Integer | Flags | 0-255 |
| fwd\_rst\_flag\_counts | Number of forward RST flags in the flow | Integer | Flags | 0-255 |
| bwd\_fin\_flag\_counts | Number of backward FIN flags in the flow | Integer | Flags | 0-255 |
| bwd\_psh\_flag\_counts | Number of backward PSH flags in the flow | Integer | Flags | 0-255 |
| bwd\_urg\_flag\_counts | Number of backward URG flags in the flow | Integer | Flags | 0-255 |
| bwd\_ece\_flag\_counts | Number of backward ECE flags in the flow | Integer | Flags | 0-255 |
| bwd\_syn\_flag\_counts | Number of backward SYN flags in the flow | Integer | Flags | 0-255 |
| bwd\_ack\_flag\_counts | Number of backward ACK flags in the flow | Integer | Flags | 0-255 |
| bwd\_cwr\_flag\_counts | Number of backward CWR flags in the flow | Integer | Flags | 0-255 |
| bwd\_rst\_flag\_counts | Number of backward RST flags in the flow | Integer | Flags | 0-255 |
| packets\_IAT\_mean | Mean inter-arrival time of packets | Float | Seconds | N/A |
| packets\_IAT\_std | Standard deviation of inter-arrival time of packets | Float | Seconds | N/A |
| packets\_IAT\_max | Maximum inter-arrival time of packets | Float | Seconds | N/A |
| packets\_IAT\_min | Minimum inter-arrival time of packets | Float | Seconds | N/A |
| packets\_IAT\_total | Total inter-arrival time of packets | Float | Seconds | N/A |
| fwd\_packets\_IAT\_mean | Mean inter-arrival time of forward packets | Float | Seconds | N/A |
| fwd\_packets\_IAT\_std | Standard deviation of inter-arrival time of forward packets | Float | Seconds | N/A |
| fwd\_packets\_IAT\_max | Maximum inter-arrival time of forward packets | Float | Seconds | N/A |
| fwd\_packets\_IAT\_min | Minimum inter-arrival time of forward packets | Float | Seconds | N/A |
| fwd\_packets\_IAT\_total | Total inter-arrival time of forward packets | Float | Seconds | N/A |
| bwd\_packets\_IAT\_mean | Mean inter-arrival time of backward packets | Float | Seconds | N/A |
| bwd\_packets\_IAT\_std | Standard deviation of inter-arrival time of backward packets | Float | Seconds | N/A |
| bwd\_packets\_IAT\_max | Maximum inter-arrival time of backward packets | Float | Seconds | N/A |
| bwd\_packets\_IAT\_min | Minimum inter-arrival time of backward packets | Float | Seconds | N/A |
| bwd\_packets\_IAT\_total | Total inter-arrival time of backward packets | Float | Seconds | N/A |
| subflow\_fwd\_packets | Number of forward packets in a subflow | Integer | Packets | N/A |
| subflow\_bwd\_packets | Number of backward packets in a subflow | Integer | Packets | N/A |
| subflow\_fwd\_bytes | Number of forward bytes in a subflow | Integer | Bytes | N/A |
| subflow\_bwd\_bytes | Number of backward bytes in a subflow | Integer | Bytes | N/A |
| label | Flow label (e.g., Botnet, Benign) | String | N/A | Botnet, Benign |

* 1. Data Accuracy Check: Missing values, duplications, outliers (using univariate method)
  2. Missing Values Analysis

A comprehensive check for missing values was conducted across all 122 columns of the dataset. The results confirmed that there are no missing values in the dataset. This finding eliminates the need for imputation, ensuring that the data is complete and ready for further analysis without additional preprocessing steps related to null values. This completeness streamlines the data preparation phase, allowing more focus on other aspects of data quality.

* 1. Duplicate Records Analysis

An analysis of duplicate rows identified 2,360 duplicates within the dataset. These duplicates are likely caused by repeated observations of network traffic or artifacts from data collection processes. Duplicate rows can bias the analysis by over-representing certain patterns and skewing model training results. Removing these rows using .drop\_duplicates() is recommended to ensure that only unique records are retained, thereby improving the reliability of the dataset and preventing unnecessary distortions during model training.

* 1. Outliers Detection Analysis

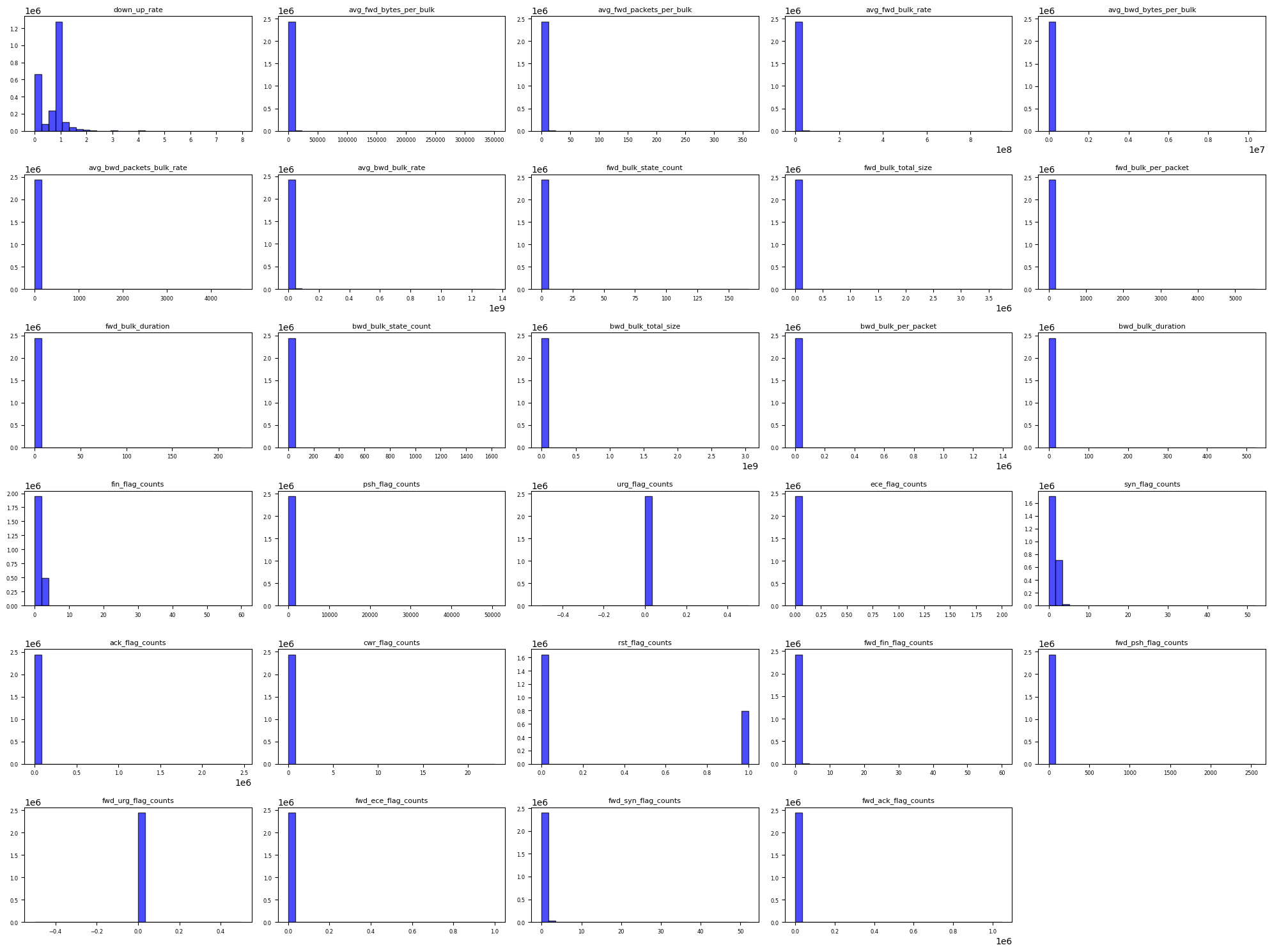
Outlier detection using the interquartile range (IQR) method revealed the presence of numerous extreme values across key numerical features. For example, src\_port had 313,719 outliers, dst\_port had 436,635, and duration contained 552,896. These outliers are significant as they may represent malicious activities or abnormal network behavior, which are vital for detecting intrusions. However, they can also introduce noise into the dataset, potentially impacting model accuracy. Handling these outliers requires a balanced approach. Relevant outliers should be retained to maintain the integrity of anomaly detection, while extreme values that do not contribute meaningfully to the analysis should be normalized or scaled. Employing robust algorithms that are less sensitive to outliers, such as Random Forest or Autoencoders, is also recommended to improve model performance.

Histogram of the numerical variablesA collage of graphs

Description automatically generated

Histogram of the numerical variables A collage of graphs

Description automatically generated

Histogram of the numerical variables  Histogram of the numerical variables A screenshot of a graph

Description automatically generated

Boxplot of the numerical variables A collage of a computer graphics

Description automatically generated

Boxplot of the numerical variables A collage of a diagram

Description automatically generated

Boxplot of the numerical variables A collage of a screenshot of a graph

Description automatically generated

Boxplot of the numerical variables A collage of a screenshot of a computer screen

Description automatically generated

Correlation Heatmap of the numerical variablesA screen shot of a graph

Description automatically generated

1. Data statistical summary

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | count | mean | std | min | 25% | 50% | 75% | max |
| src\_port | 1828409 | 43428.5 | 20624.9 | 7 | 36208 | 51795 | 58581 | 65535 |
| dst\_port | 1828409 | 5027.9 | 14066.8 | 1 | 53 | 80 | 443 | 55404 |
| duration | 1828409 | 8.8 | 30.9 | 0 | 0 | 0 | 0.1 | 384.3 |
| packets\_count | 1828409 | 7.8 | 17 | 1 | 1 | 4 | 4 | 4674 |
| fwd\_packets\_count | 1828409 | 4 | 7.7 | 0 | 1 | 2 | 2 | 1615 |
| bwd\_packets\_count | 1828409 | 3.9 | 9.4 | 0 | 1 | 2 | 2 | 3059 |
| total\_payload\_bytes | 1828409 | 2400.1 | 12629.8 | 0 | 0 | 192 | 414 | 4834698 |
| fwd\_total\_payload\_bytes | 1828409 | 238.4 | 702.6 | 0 | 0 | 60 | 94 | 20849 |
| bwd\_total\_payload\_bytes | 1828409 | 2161.6 | 12360.1 | 0 | 0 | 122 | 300 | 4831507 |
| payload\_bytes\_max | 1828409 | 479.8 | 1104.3 | 0 | 0 | 76 | 172 | 7300 |
| payload\_bytes\_min | 1828409 | 17.4 | 21.7 | 0 | 0 | 0 | 38 | 112 |
| payload\_bytes\_mean | 1828409 | 101.7 | 183.6 | 0 | 0 | 56 | 96 | 1121.8 |
| payload\_bytes\_std | 1828409 | 140.9 | 325.4 | 0 | 0 | 16 | 64 | 2163.7 |
| payload\_bytes\_variance | 1828409 | 125703 | 479867 | 0 | 0 | 256 | 4096 | 4681805 |
| fwd\_payload\_bytes\_max | 1828409 | 479.8 | 1104.3 | 0 | 0 | 76 | 172 | 7300 |
| fwd\_payload\_bytes\_min | 1828409 | 17.4 | 21.7 | 0 | 0 | 0 | 38 | 112 |
| fwd\_payload\_bytes\_mean | 1828409 | 101.7 | 183.6 | 0 | 0 | 56 | 96 | 1121.8 |
| fwd\_payload\_bytes\_std | 1828409 | 140.9 | 325.4 | 0 | 0 | 16 | 64 | 2163.7 |
| fwd\_payload\_bytes\_variance | 1828409 | 9071 | 52877.8 | 0 | 0 | 0 | 0 | 1248354 |
| bwd\_payload\_bytes\_max | 1828409 | 479.8 | 1104.3 | 0 | 0 | 76 | 172 | 7300 |
| bwd\_payload\_bytes\_min | 1828409 | 17.4 | 21.7 | 0 | 0 | 0 | 38 | 112 |
| bwd\_payload\_bytes\_mean | 1828409 | 101.7 | 183.6 | 0 | 0 | 56 | 96 | 1121.8 |
| bwd\_payload\_bytes\_std | 1828409 | 140.9 | 325.4 | 0 | 0 | 16 | 64 | 2163.7 |
| bwd\_payload\_bytes\_variance | 1828409 | 179817 | 721081 | 0 | 0 | 0 | 0 | 8208420 |
| total\_header\_bytes | 1828409 | 177.6 | 463.8 | 8 | 20 | 32 | 52 | 93504 |
| max\_header\_bytes | 1828409 | 19.8 | 11.5 | 8 | 8 | 20 | 32 | 52 |
| min\_header\_bytes | 1828409 | 16.9 | 8.5 | 8 | 8 | 20 | 20 | 44 |
| mean\_header\_bytes | 1828409 | 17.6 | 9.1 | 8 | 8 | 20 | 22 | 44 |
| std\_header\_bytes | 1828409 | 1.1 | 1.9 | 0 | 0 | 0 | 2 | 8.5 |
| fwd\_total\_header\_bytes | 1828409 | 88.3 | 212.7 | 0 | 16 | 20 | 32 | 32312 |
| fwd\_max\_header\_bytes | 1828409 | 16.9 | 12.6 | 0 | 8 | 8 | 32 | 52 |
| fwd\_min\_header\_bytes | 1828409 | 14.7 | 9.8 | 0 | 8 | 8 | 20 | 44 |
| fwd\_mean\_header\_bytes | 1828409 | 15 | 10.1 | 0 | 8 | 8 | 21.3 | 44 |
| fwd\_std\_header\_bytes | 1828409 | 0.7 | 1.4 | 0 | 0 | 0 | 0 | 5.9 |
| bwd\_total\_header\_bytes | 1828409 | 89.3 | 255 | 0 | 8 | 16 | 40 | 61192 |
| bwd\_max\_header\_bytes | 1828409 | 15 | 13.2 | 0 | 8 | 8 | 32 | 52 |
| bwd\_min\_header\_bytes | 1828409 | 12.5 | 10 | 0 | 8 | 8 | 20 | 44 |
| bwd\_mean\_header\_bytes | 1828409 | 13.1 | 10.7 | 0 | 8 | 8 | 20.9 | 44 |
| bwd\_std\_header\_bytes | 1828409 | 0.9 | 1.8 | 0 | 0 | 0 | 0 | 7.1 |
| fwd\_avg\_segment\_size | 1828409 | 32.3 | 42.3 | 0 | 0 | 33 | 45 | 603.6 |
| bwd\_avg\_segment\_size | 1828409 | 174.7 | 360.2 | 0 | 0 | 68 | 142.9 | 2245.6 |
| avg\_segment\_size | 1828409 | 101.7 | 183.6 | 0 | 0 | 56 | 96 | 1121.8 |
| fwd\_init\_win\_bytes | 1828409 | 4345.7 | 9253 | 0 | 0 | 0 | 1024 | 52744 |
| bwd\_init\_win\_bytes | 1828409 | 5839.8 | 11924.9 | 0 | 0 | 0 | 235 | 51452 |
| active\_min | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| active\_max | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| active\_mean | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| active\_std | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| idle\_min | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| idle\_max | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| idle\_mean | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| idle\_std | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| bytes\_rate | 1828409 | 248222 | 643422 | 0 | 0 | 657.2 | 10428.6 | 3.10E+07 |
| fwd\_bytes\_rate | 1828409 | 69364 | 197916 | 0 | 0 | 65.8 | 2459.9 | 2.60E+07 |
| bwd\_bytes\_rate | 1828409 | 178858 | 510270 | 0 | 0 | 482 | 7574.5 | 2.10E+07 |
| packets\_rate | 1828409 | 6393.1 | 15356.5 | 0 | 0 | 32.8 | 1412.5 | 266305 |
| bwd\_packets\_rate | 1828409 | 3476.7 | 8929.6 | 0 | 0 | 16.2 | 1016.3 | 159277 |
| fwd\_packets\_rate | 1828409 | 2916.4 | 7298.7 | 0 | 0 | 8.9 | 84.8 | 144631 |
| down\_up\_rate | 1828409 | 0.7 | 0.5 | 0 | 0 | 1 | 1 | 2.1 |
| avg\_fwd\_bytes\_per\_bulk | 1828409 | 5.3 | 55.5 | 0 | 0 | 0 | 0 | 4768 |
| avg\_fwd\_packets\_per\_bulk | 1828409 | 0.1 | 0.6 | 0 | 0 | 0 | 0 | 4 |
| avg\_fwd\_bulk\_rate | 1828409 | 1175.8 | 56833.8 | 0 | 0 | 0 | 0 | 9998024 |
| avg\_bwd\_bytes\_per\_bulk | 1828409 | 910.4 | 5206.2 | 0 | 0 | 0 | 0 | 114202 |
| avg\_bwd\_packets\_bulk\_rate | 1828409 | 0.6 | 3 | 0 | 0 | 0 | 0 | 53.8 |
| avg\_bwd\_bulk\_rate | 1828409 | 586803 | 3308720 | 0 | 0 | 0 | 0 | 2.80E+07 |
| fwd\_bulk\_state\_count | 1828409 | 0 | 0.2 | 0 | 0 | 0 | 0 | 1 |
| fwd\_bulk\_total\_size | 1828409 | 5.3 | 55.5 | 0 | 0 | 0 | 0 | 4768 |
| fwd\_bulk\_per\_packet | 1828409 | 0.1 | 0.6 | 0 | 0 | 0 | 0 | 4 |
| fwd\_bulk\_duration | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0.9 |
| bwd\_bulk\_state\_count | 1828409 | 0.1 | 0.4 | 0 | 0 | 0 | 0 | 6 |
| bwd\_bulk\_total\_size | 1828409 | 1408.2 | 10644.7 | 0 | 0 | 0 | 0 | 646004 |
| bwd\_bulk\_per\_packet | 1828409 | 0.9 | 6.2 | 0 | 0 | 0 | 0 | 323 |
| bwd\_bulk\_duration | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 2.4 |
| fin\_flag\_counts | 1828409 | 0.4 | 0.7 | 0 | 0 | 0 | 0 | 2 |
| psh\_flag\_counts | 1828409 | 1.5 | 4.4 | 0 | 0 | 0 | 0 | 166 |
| urg\_flag\_counts | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ece\_flag\_counts | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| syn\_flag\_counts | 1828409 | 0.5 | 0.8 | 0 | 0 | 0 | 1 | 3 |
| ack\_flag\_counts | 1828409 | 5.8 | 17.3 | 0 | 0 | 0 | 1 | 4673 |
| cwr\_flag\_counts | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| rst\_flag\_counts | 1828409 | 0.3 | 0.4 | 0 | 0 | 0 | 1 | 1 |
| fwd\_fin\_flag\_counts | 1828409 | 0.2 | 0.4 | 0 | 0 | 0 | 0 | 1 |
| fwd\_psh\_flag\_counts | 1828409 | 0.7 | 1.9 | 0 | 0 | 0 | 0 | 20 |
| fwd\_urg\_flag\_counts | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| fwd\_ece\_flag\_counts | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| fwd\_syn\_flag\_counts | 1828409 | 0.3 | 0.4 | 0 | 0 | 0 | 1 | 2 |
| fwd\_ack\_flag\_counts | 1828409 | 2.8 | 7.8 | 0 | 0 | 0 | 1 | 1614 |
| fwd\_cwr\_flag\_counts | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| fwd\_rst\_flag\_counts | 1828409 | 0.1 | 0.4 | 0 | 0 | 0 | 0 | 1 |
| bwd\_fin\_flag\_counts | 1828409 | 0.2 | 0.4 | 0 | 0 | 0 | 0 | 1 |
| bwd\_psh\_flag\_counts | 1828409 | 0.8 | 2.7 | 0 | 0 | 0 | 0 | 155 |
| bwd\_urg\_flag\_counts | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| bwd\_ece\_flag\_counts | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| bwd\_syn\_flag\_counts | 1828409 | 0.2 | 0.4 | 0 | 0 | 0 | 0 | 1 |
| bwd\_ack\_flag\_counts | 1828409 | 3 | 9.6 | 0 | 0 | 0 | 1 | 3059 |
| bwd\_cwr\_flag\_counts | 1828409 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| bwd\_rst\_flag\_counts | 1828409 | 0.1 | 0.3 | 0 | 0 | 0 | 0 | 1 |
| packets\_IAT\_mean | 1828409 | 4.00E+08 | 6.70E+08 | 0 | 0 | 0 | 1.50E+09 | 1.50E+09 |
| packet\_IAT\_std | 1828409 | 0.7 | 2.7 | 0 | 0 | 0 | 0 | 41.1 |
| packet\_IAT\_max | 1828409 | 4.00E+08 | 6.70E+08 | 0 | 0 | 0.1 | 1.50E+09 | 1.50E+09 |
| packet\_IAT\_min | 1828409 | 4.00E+08 | 6.70E+08 | 0 | 0 | 0 | 1.50E+09 | 1.50E+09 |
| packet\_IAT\_total | 1828409 | 4.00E+08 | 6.70E+08 | 0 | 0 | 0.2 | 1.50E+09 | 1.50E+09 |
| fwd\_packets\_IAT\_mean | 1828409 | 5.20E+08 | 7.10E+08 | 0 | 0 | 0.5 | 1.50E+09 | 1.50E+09 |
| fwd\_packets\_IAT\_std | 1828409 | 0.7 | 2.4 | 0 | 0 | 0 | 0 | 25.7 |
| fwd\_packets\_IAT\_max | 1828409 | 5.20E+08 | 7.10E+08 | 0 | 0 | 2.6 | 1.50E+09 | 1.50E+09 |
| fwd\_packets\_IAT\_min | 1828409 | 5.20E+08 | 7.10E+08 | 0 | 0 | 0 | 1.50E+09 | 1.50E+09 |
| fwd\_packets\_IAT\_total | 1828409 | 5.20E+08 | 7.10E+08 | 0 | 0 | 3.8 | 1.50E+09 | 1.50E+09 |
| bwd\_packets\_IAT\_mean | 1828409 | 3.30E+08 | 6.20E+08 | 0 | 0 | 0 | 7.2 | 1.50E+09 |
| bwd\_packets\_IAT\_std | 1828409 | 0.6 | 2.4 | 0 | 0 | 0 | 0 | 25 |
| bwd\_packets\_IAT\_max | 1828409 | 3.30E+08 | 6.20E+08 | 0 | 0 | 0 | 10.2 | 1.50E+09 |
| bwd\_packets\_IAT\_min | 1828409 | 3.30E+08 | 6.20E+08 | 0 | 0 | 0 | 0 | 1.50E+09 |
| bwd\_packets\_IAT\_total | 1828409 | 3.30E+08 | 6.20E+08 | 0 | 0 | 0 | 116 | 1.50E+09 |
| subflow\_fwd\_packets | 1828409 | 0.8 | 2.8 | 0 | 0 | 0 | 0 | 538.3 |
| subflow\_bwd\_packets | 1828409 | 0.8 | 3.3 | 0 | 0 | 0 | 0 | 1019.7 |
| subflow\_fwd\_bytes | 1828409 | 57.5 | 249.6 | 0 | 0 | 0 | 0 | 4910 |
| subflow\_bwd\_bytes | 1828409 | 57.5 | 249.6 | 0 | 0 | 0 | 0 | 4910 |
| flow\_id | 1828409 |  |  |  |  |  |  |  |
| timestamp | 1828409 |  |  |  |  |  |  |  |
| src\_ip | 1828409 |  |  |  |  |  |  |  |
| dst\_ip | 1828409 |  |  |  |  |  |  |  |
| protocol | 1828409 |  |  |  |  |  |  |  |
| label | 1828409 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

* 1. Data statical report

Summary of Key Variables

* + 1. The source port (src\_port) has a mean of 43,428.5, while the destination port (dst\_port) averages 5,027.9, reflecting common service ports. Flow duration (duration) averages 8.8 seconds, with peaks up to 384.3 seconds, indicating varying connection lengths.
    2. Packet counts (packets\_count) average 7.8 per flow, but some flows reach 4,674 packets, suggesting large transmissions. Payload data (total\_payload\_bytes) shows a mean of 2,400.1 bytes, but extreme cases exceed 4.8 million bytes, indicating high-traffic flows.
    3. The bytes transfer rate (bytes\_rate) averages 248,222 bytes/sec, but spikes to 31 million bytes/sec, reflecting bursts of activity. Packet inter-arrival times (packet\_IAT\_max) vary widely, reaching 1.5 billion seconds, suggesting long idle periods in some flows.
    4. TCP flag usage is extensive, with acknowledgment packets (ack\_flag\_counts) averaging 5.8 per flow, but reaching 4,673 in some cases. These high counts may indicate persistent communication patterns.

The dataset shows high variability in traffic volume, speed, and timing. Some flows involve large data transfers, potentially indicating bulk downloads or attacks. Asymmetric traffic patterns and extreme values in bytes\_rate and packets\_count suggest the presence of anomalies, possibly linked to malicious activity.

1. Proposed Solutions to Address Data Accuracy Issues
   1. Identified Issues
      1. The dataset contains 2,360 duplicate rows, leading to redundancy. Additionally, several numerical features exhibit extreme outliers detected using the Interquartile Range (IQR) method, which can negatively impact model performance.
   2. Proposed Solutions
      1. Removing Duplicate Rows: Duplicate records will be removed.
      2. Handling Outliers: IQR-Based Clipping: Replacing values beyond the IQR-defined thresholds with the closest limit to control distortions.

These solutions will enhance data integrity by eliminating redundancy and limiting the impact of extreme values. The refined dataset will improve model robustness and ensure reliable intrusion detection.

1. Conclusion

Cyber threats are becoming increasingly sophisticated, necessitating advanced security measures. This report highlights the need for an AI-powered Intrusion Detection System (AI-IDS) at Queensland University, as traditional rule-based systems are insufficient against evolving cyber threats.

A review of machine learning-based IDS solutions identified Random Forest (RF), Deep Neural Networks (DNNs), and Autoencoders as effective approaches. RF offers high accuracy but requires optimization for efficiency. DNNs excel in detecting complex attack patterns but demand significant computational resources, while Autoencoders provide robust anomaly detection with careful threshold tuning.

To enhance detection performance, data preprocessing steps such as removing duplicate records and handling outliers were recommended. These measures ensure data integrity and improve model reliability. The AI-IDS implementation will strengthen cybersecurity, ensure regulatory compliance, and provide a scalable security framework.

In conclusion, integrating AI-driven IDS will fortify Queensland University’s defenses, offering proactive threat mitigation against evolving cyber risks. Future efforts should focus on optimizing detection accuracy and maintaining up-to-date threat intelligence for continuous protection.

1. Reference

Javaid et al. (2016):  
Javaid, A., Niyaz, Q., Sun, W., and Alam, M. (2016). "A Deep Learning Approach for Network Intrusion Detection System." *EAI Endorsed Transactions on Security and Safety*, 3(9), e2.

Kim et al. (2018):  
Kim, M., Kang, M., and Kang, J. (2018). "Anomaly Detection for HTTP Using Convolutional Autoencoders." *IEEE Access*, 6, pp. 70884-70891.

Yaseen and Shamsuddin (2018):  
Yaseen, M. and Shamsuddin, S.M. (2018). "A Survey of Random Forest Based Methods for Intrusion Detection Systems." International Journal of Computer Applications, 113(7), pp. 1-7.