Vivekanand Education Society's
Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai, Approved by AICTE & Recognised by Govt. of Maharashtra)
NAAC accredited with 'A' grade

Semester : VI **Subject : AIDS - 1**

Title of the Project :
Analysis of the Website Traffic Anomalies

Domain: Unsupervised Learning

Member 1: Mohit Kerkar (23) D15C
Member 2: Bhumisha Parchani (38) D15C
Member 3: Bhavisha Khotwani (25) D15C

Professor Name : Dr. Ravita Mishra

Content

- Overview
- Introduction
- Problem Statement
- Objectives
- Dataset info
- Model making approach
- Implementation
- Conclusion
- References

Introduction to Project

What is Anomaly Detection?

- Anomaly detection is the process of identifying data points that significantly deviate from the norm.
- It is used in various fields such as cybersecurity, fraud detection, network traffic monitoring, and predictive maintenance.

Why is it Important?

- Helps detect fraudulent transactions, network intrusions, and sensor failures.
- Enhances security, improves efficiency, and prevents system failures.

Problem Statement

As network traffic continues to grow in complexity and volume, detecting anomalies that indicate cybersecurity threats, fraudulent activities, or operational inefficiencies has become a critical challenge. Traditional rule-based detection systems struggle to adapt to evolving attack patterns and unexpected behaviors. This study explores the use of **machine learning-based anomaly detection** to identify unusual network traffic patterns, aiming to develop a scalable and effective model that enhances network security and performance monitoring.

Objectives of the project

- **To enhance network security** by identifying unusual traffic patterns that may indicate cyber threats, intrusions, or fraudulent activities.
- **To compare different machine learning approaches** for anomaly detection and evaluate their effectiveness in real-world scenarios.
- **To develop a scalable anomaly detection model** that can be applied to large datasets for proactive monitoring and predictive analytics.
- **To develop an Interface** which can detect anomalies within the network traffic by taking Website log files as input

Dataset Info

The dataset ‘Website Anamolies’ includes packet-level network traffic logs, likely from a network monitoring tool such as Wireshark.

It contains the following attributes:

- Time – Timestamp of when the packet was captured.
- Source – The sender of the packet (IP address or MAC address).
- No. – A unique sequence number for each packet.
- Destination – The recipient of the packet (IP address, MAC address, or Broadcast).
- Protocol – The communication protocol used (e.g., ARP, NBNS, ICMPv6, BROWSER).
- Length – The size of the packet in bytes.
- Info – Additional packet metadata (e.g., ARP requests, Host Announcements, RARP requests).

Dataset Info

Based on this dataset structure, anomalies could be:

- **Unusual traffic patterns** (e.g., sudden spikes in packet size).
- **Suspicious repeated requests** (e.g., continuous ARP or RARP queries).
- **Unexpected source or destination addresses** (e.g., an unknown IP requesting multiple connections).
- **Protocol misuse** (e.g., protocols used in an abnormal manner).

Literature Survey

Sr. No.	Title	Name of Author	Name of Journal & Year of Publication	Methodology	Results/ conclusions	Drawbacks/ limitations
1.	Network Anomaly Traffic Analysis	Kaibin Lu	Academic Journal of Science and Technology - 2024	**Statistical Method** – Z-score analysis. **Machine Learning Approaches** – Clustering (DBSCAN), Support Vector Machine (SVM). **Rule-Based and Threshold-Based Detection** – Uses predefined rules for anomaly detection.	**Statistical methods** detected simple attacks but had a **high false positive rate**. **SVM performed well** in separating normal and anomalous traffic but required careful tuning. **Rule-based methods were rigid**, unable to detect zero-day attacks effectively.	**High computational cost** for ML models, requiring more processing power. **Statistical methods had too many false positives**, limiting real-world reliability. **Rule-based detection** lacked flexibility.

Literature Survey

Sr. No.	Title	Name of Author	Name of Journal & Year of Publication	Methodology	Results/ conclusions	Drawbacks/ limitations
2.	Machine Learning in Network Anomaly Detection: A Survey	Song Wang, Juan Fernando Balarezo Serrano, Kandeepan Sithamparanathan, Akram Al-Hourani	IEEE Access, 2021	- The paper reviews various machine learning (ML) approaches for network anomaly detection. - It discusses various algorithms like decision trees, support vector machines (SVM), neural networks, and deep learning models	- ML-based models improve **detection accuracy** and **reduce false positives** comparatively. - **Hybrid approaches** (combining multiple ML models) yield better performance in real-world scenarios.	**Data dependency**: The performance of ML models depends on the quality and quantity of training data. **False Positives**: Some methods still produce a high rate of false alarms, making real-time deployment challenging.

Model making approach

Steps in Anomaly Detection Model Development

Data Preparation – Collect, clean, and preprocess network traffic data.
Feature Engineering – Select key attributes like Time, Length, and Protocol.

Apply Anomaly Detection Methods:

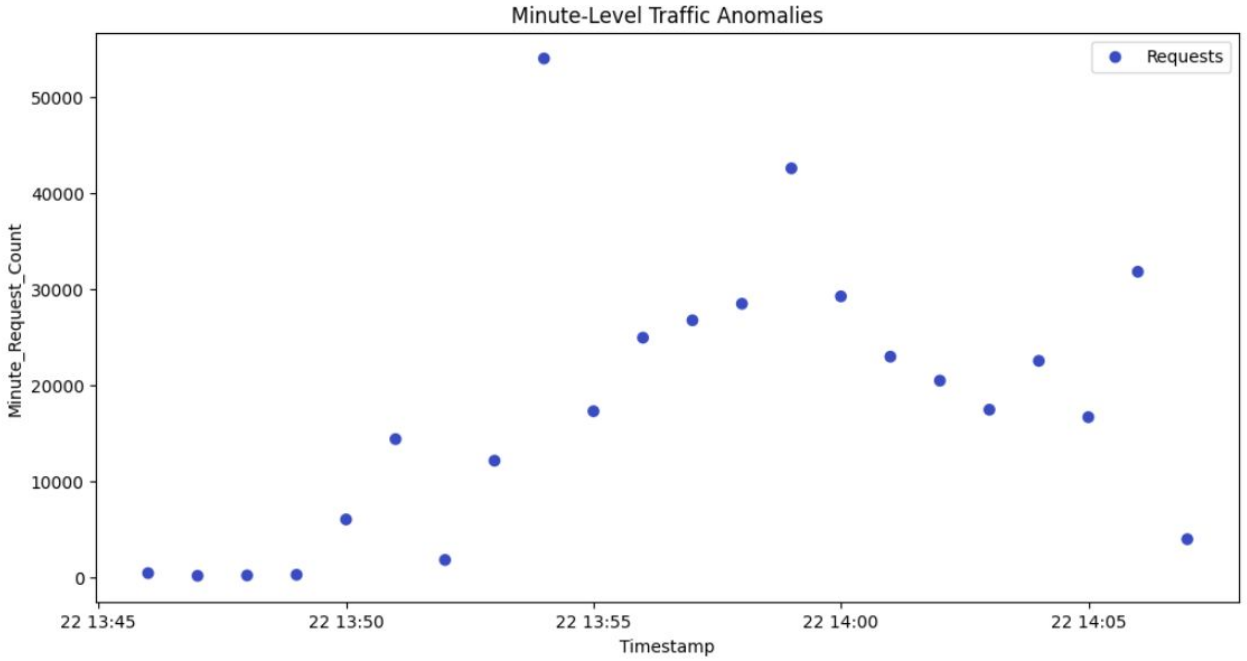
- **Z-Score** (Statistical approach)
- **One-Class SVM** (Machine Learning-based)

Visualization & Analysis – Use plots to interpret and compare anomalies.
Evaluation – Validate detection effectiveness with expert feedback.
Deployment – Integrate into real-time monitoring and alerting systems.

Implementation

Z-Score Anomaly Detection

- Statistical threshold-based detection
- Flags points deviating beyond a set threshold (e.g., |Z| > 2.5); labels outliers as -1.



Key Findings
Total Anomalies Detected: **Varies based on Z-score threshold**

Trend:
Anomalies occur at extreme request spikes, deviating from the mean traffic volume.

Observation:
Higher Z-score values indicate significant deviation from normal traffic patterns.

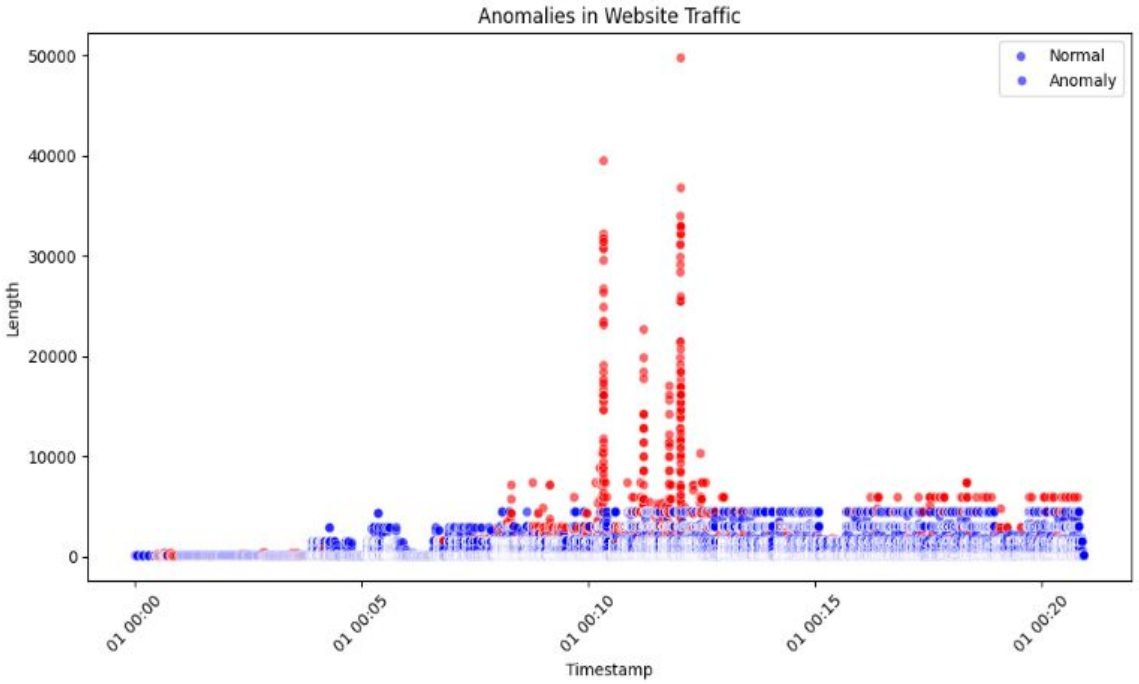
Pros:
Simple, interpretable, effective for normally distributed data.

Cons:
Sensitive to scale; assumes normal distribution, may misclassify skewed data.

Implementation

One Class SVM

- Boundary-based
- **Binary Output**: Labels anomalies as -1.



Key Findings
Total Anomalies Detected: 10,044

Anomalies Labels:
-1 → Anomalies (Outliers)
-0 → Normal Data Points

Trend: Anomalies often occur with extreme spikes in packet length.

Observation: Higher deviation from the normal range leads to anomaly classification.

Pros: Works well with high-dimensional data.

Cons: May misclassify normal fluctuations as anomalies.

Result

Website Traffic Anomaly Detection

This application analyzes website traffic data to detect potential anomalies and security threats. Upload your traffic data CSV file to get started.

Upload Traffic Data

Choose a CSV file

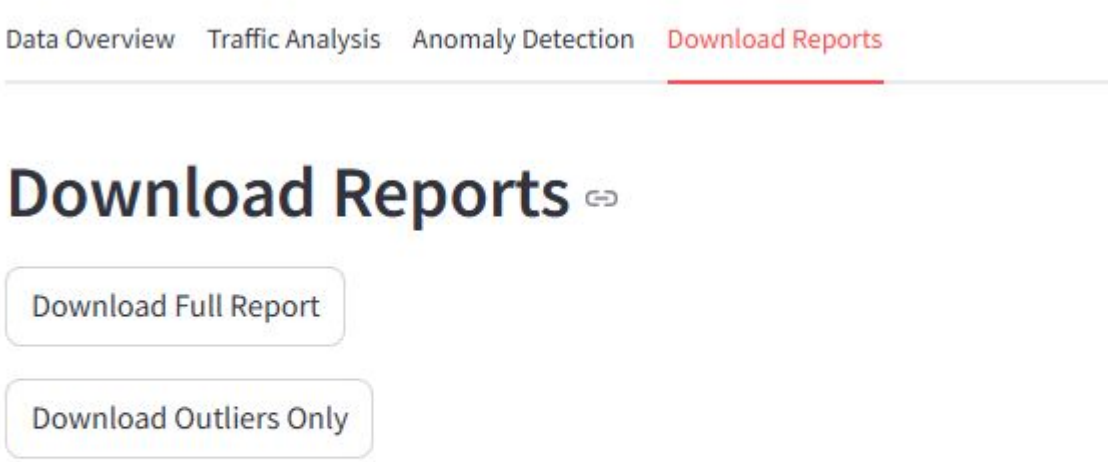
Drag and drop file here
Limit 200MB per file • CSV

Browse files

Please upload a CSV file with the following columns: Time, Length, Source, Destination, Protocol

	Time	Length	Source	Destination	Protocol
0	0	128	192.168.1.1	10.0.0.1	6
1	1	256	192.168.1.2	10.0.0.2	17
2	2	512	192.168.1.3	10.0.0.1	6
3	3	128	192.168.1.4	10.0.0.3	1
4	4	1024	192.168.1.5	10.0.0.2	6

Result


Data Overview
Traffic Analysis
Anomaly Detection
Download Reports
Download Reports
Download Full Report
Download Outliers Only

Conclusion

In this study, we applied various anomaly detection techniques to identify bot-generated traffic within website logs. By comparing statistical (Z-score), clustering (DBSCAN, K-Means), and machine learning-based (One-Class SVM, KNN) approaches, we found:

- **Z-Score** effectively flags sudden spikes but sometimes mistakes real user traffic for bots.
- **One-Class SVM** is useful for identifying unseen bot behavior but may misclassify human-driven traffic variations.

Our findings highlight the importance of anomaly detection in mitigating bot-related threats, such as ad fraud, DDoS attacks, and data scraping. Future enhancements could integrate real-time bot detection using deep learning and behavioral analysis for improved accuracy.

References

1. S. Wang, J. F. B. Serrano, K. Sithamparanathan, and A. Al-Hourani, “Machine Learning in Network Anomaly Detection: A Survey,” *IEEE Access*, vol. 9, pp. 120610-120630, Nov. 2021. doi: 10.1109/ACCESS.2021.3126834.

2. K. Lu, “Network Anomaly Traffic Analysis,” *Academic Journal of Science and Technology*, vol. 10, no. 3, pp. 65-68, Apr. 2024. doi: 10.54097/8as0rg31.