# Anomalies Detection in an Internet Firewall Data

Bantu Keerthana
*Master of Applied Computing*
*Wilfrid Laurier University*
Waterloo, Canada
235847180

Sayan Banerjee
*Master of Applied Computing*
*Wilfrid Laurier University*
Waterloo, Canada
235847750

***Abstract*- Data is an integral part of today's world, be it financial, transactional, personal data. With the rise in data, fraudulent activities are also rising at a higher rate. For instance, during a transaction, a hacker could launch a Man-In-The Middle Attacks (MITM), which allows them to gain access into an authorized institutions server, allowing them to eavesdrop on and manipulate the data being transmitted. They often gain access during a transaction phase and steal transactional data as it's being sent or received. In present days, the availability of internet and the wide variety of services like e-commerce and online shopping gained much popularity. On the other side of the coin, customers are facing adverse benefits due to fraudulent activities. Therefore, analyzing, detecting, and preventing such unusual fraudulent activities in- real time is very essential and plays a crucial part. [1] In this paper we are conducting a prediction study using Isolation Forest and Gradient Boosting of Machine Learning and Graph Neural Network of Deep Learning towards determining the anomalies in Firewall Data.**

***Keywords—GNN Graphical Neural Network), Isolation Forest, Detection, Anomalies, Gradient Boosting.***

## I. INTRODUCTION

With the raising of Internet technologies virtualization became mainstream in the deployment process [5], the integrity and security of data exchanged over networks have become paramount. However, the rise of cyber threats, including sophisticated attacks like Man-In-The-Middle (MITM) breaches, poses a constant challenge to maintaining the confidentiality and authenticity of information.

Imagine a scenario where a financial institution processes thousands of transactions daily through its online platform. Unbeknownst to the institution and its clients, a cybercriminal orchestrates a MITM attack during one of these transactions. With this information in hand, the hacker can manipulate or exploit the data for malicious purposes, including identity theft or financial fraud. A similar situation happened in July 2021 when hackers found a way to break into a software called Kaseya's VSA. This software helps businesses manage their computers remotely. The hackers used this to spread ransomware, which locks up computers until you pay money. Lots of businesses got hit, causing big problems in different industries like banking, healthcare, and manufacturing. It shows how one attack can affect lots of other businesses connected to the same software.

Implicit authentication, an approach that uses observations of user behavior for authentication and can also be considered to differentiate legitimate users from fraudsters. Basically, there are two ways to detect fraud. One is Misuse Detection and the second is Anomaly Detection. For detecting misuse transaction classification algorithms can be applied. For anomaly detection, the basic profile of the user and past transactions, along with behavior of the owner are considered during the analysis.[1] The evolving nature of cyber threats underscores the urgency for robust cybersecurity measures. Among these measures, anomaly detection in internet firewall data stands out as a crucial defence mechanism against unauthorized access and data breaches. Anomaly detection complements traditional cybersecurity measures by offering a proactive approach to threat detection that can identify novel and previously unseen threats.

Traditional cybersecurity measures often rely on predefined rules, signatures, or patterns to identify known threats. While effective against known attack vectors, these methods may struggle to detect emerging or sophisticated attacks that deviate from established patterns. This is where anomaly detection fills a crucial gap. In this paper, we focus on leveraging advanced analytics techniques to detect anomalies within firewall data, specifically targeting historical datasets for predictive analysis.In our research, we used three main algorithms: Isolation Forest, Gradient Boosting, and Graph Neural Networks. These algorithms have been selected because of their effectiveness in identifying aberrations within network traffic patterns, indicating the potential of a cyberattack or anomalous behaviour. By training predictive models on historical firewall data, we hope to develop a proactive approach to cybersecurity, empowering organizations to anticipate and mitigate threats before they are ramped up into full-blown attacks.

We comprehensively review historical firewall datasets and perform experiments in predictive models to evaluate the performance and efficacy of each algorithm in spotting anomalies. Through a comprehensive examination of historical firewall datasets and rigorous experimentation with predictive models, we seek to evaluate the performance and efficacy of each algorithm in detecting anomalies. By elucidating the strengths and limitations of these approaches, we aim to provide valuable insights for cybersecurity practitioners and network administrators tasked with safeguarding critical infrastructure and data assets. Deploying anomaly detection models in real-time enables continuous monitoring of network traffic and cybersecurity events.

Ultimately, our goal is to contribute to the ongoing efforts to enhance cybersecurity resilience and fortify defences against evolving cyber threats. By leveraging the power of predictive analytics and anomaly detection, we strive to empower organizations to stay one step ahead of cyber adversaries, ensuring the integrity and security of their digital ecosystems.

## II. Related Works

### A. *Isolation Forest and XG Boosting for Classifying Credit Card Fraudulent Transactions*

Previous works have worked on combining Isolation Forest and Local Outlier Factor algorithms to detect outliers and anomalies in credit card transactions. These algorithms explicitly identify unusual fraudulent activities in real-time. Additionally, the authors utilize Extreme Gradient Boosting to construct and evaluate the predictive model comparison.
This comparative study highlights the advantages and superior performance of the Isolation Forest and XgBoosting approach in detecting fraudulent transactions.
In conclusion, this paper presents a comprehensive study on credit card fraud detection using Isolation Forest, Local Outlier Factor, and Extreme Gradient Boosting algorithms. The proposed model demonstrates high accuracy in capturing fraudulent transactions. But this paper had the problem of dataset overfitting.

### B. *A Network Traffic Anomaly Detection Method based on CNN and XG Boost*

This work discusses on network traffic anomaly detection using combination of convolution neural network (CNN) and eXtreme Gradient Boosting algorithm. The traditional methods of intrusion detection are limited in dealing with complex network environments, so the study proposes a model that can effectively detect abnormal traffic patterns. The CNN algorithm is used to learn traffic characteristics, and the XGBoost algorithm is utilized for classification. The limitation of using a classification approach in the context of network traffic anomaly detection is that it may not provide a clear understanding of the underlying patterns and relationships in the data. Classification methods, such as SVM, Bayesian network, and neural network, are effective in categorizing data into predefined classes based on specific features. However, they may not capture the complex and dynamic nature of network traffic anomalies.

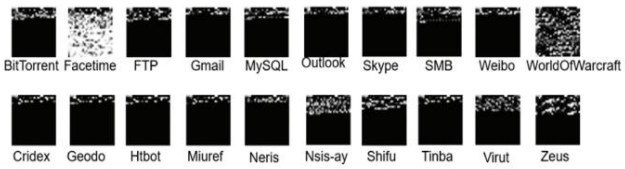


Fig. 1 Flow grayscale image of USTC-TFC2016

By adopting a predictive approach, the authors would have gained insights into the underlying dynamics of network traffic anomalies, enabling them to compare and analyze the anomalies more effectively. This approach would have provided a more holistic understanding of the anomalies and potentially improved the accuracy and reliability of the detection method. The paper does not provide a comprehensive analysis of the performance of the proposed method. It only mentions that the accuracy is improved and the false alarm rate is reduced. However, additional evaluation metrics such as precision, recall, and F1 score could provide a more complete assessment of the method's performance.

### C. *Network Anomaly Detection using a Graph Neural Network*

GNNs have gained popularity in recent years due to their ability to represent both structured and unstructured data in graphical form. Several GNN models have been proposed for anomaly detection purposes in network environments, including GCN, Mixture Model Network (MoNet), GraphSage, and Graph Attention Network (GAT). The AEN model is analyzed to acquire a structural foundation on graphs, which is then applied to construct the GNN model. GNN model, specifically a Graph Convolutional Network (GCN), is utilized as a message passing method for accurately distinguishing between normal and abnormal traffic in a network system.

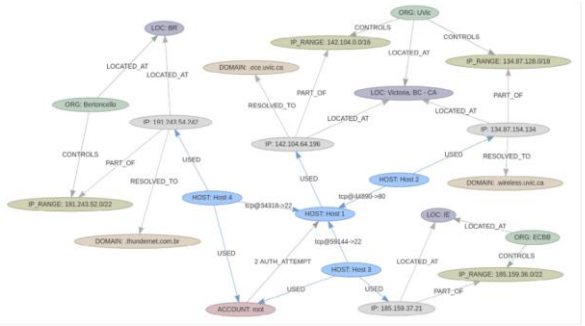


Fig.2 Snapshot of a AEN graph Based On a subset of the ISOC CID dataset

It only used two types of datasets for their experiments: the DDoS dataset and the TOR-nonTOR dataset. While these datasets provide valuable insights into network anomaly detection, they may not cover the full range of possible network threats and anomalies. It would be beneficial to include additional datasets representing different types of attacks or anomalies to further validate the proposed model's effectiveness. While the accuracy scores obtained using the DDoS and TOR-nonTOR datasets are promising (76% and 88% respectively), it is important to evaluate the model's performance on other metrics such as precision, recall, and F1-score.

## III. Proposed Approach

The proposed method is to use the Machine Learning algorithms and Deep Learning Algorithms together to carry out a predictive study in a historical firewall data.

We have employed Machine Learning Algorithms like Isolation Forest and Gradient Boosting. On the other hand, we have used Graph Neural Network (GNN), a deep leaning algorithm.

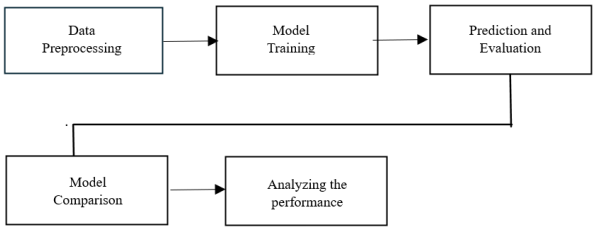


Fig.3 Implementation of the model

### A. *Isolation Forest*

Isolation Forest is a machine learning algorithm specifically designed for anomaly detection tasks. It works by recursively partitioning the data space into smaller subsets, isolating anomalies in fewer steps compared to traditional methods. Anomalies are often isolated into smaller partitions because they are rare instances in the dataset, making them stand out from most normal data points. This approach allows Isolation Forest to efficiently identify outliers or anomalies in firewall data without requiring a comprehensive understanding of normal behavior, making it particularly effective for detecting novel or previously unseen threats. The contamination parameter, set to 0.2847, represents the proportion of outliers in the dataset. Essentially, it specifies the expected percentage of anomalies in the dataset. By setting this parameter, the algorithm aims to identify observations that are significantly different from most of the data.

Unlike some anomaly detection techniques that rely on modeling the entire data distribution, Isolation Forest constructs simple decision trees based on randomly selected features and split points. By focusing on random partitions of the data, Isolation Forest reduces the risk of overfitting to specific patterns present in the training data. This inherent randomness in the tree construction process helps prevent the model from memorizing outliers in the training set, resulting in better generalization performance when applied to unseen data. Isolation Forest has linear time complexity with respect to the size of the dataset, making it highly scalable to large volumes of firewall data.

### B. *Gradient Boosting*

Gradient Boosting is an ensemble learning technique and a tree-based supervised machine learning algorithm, that iteratively builds a strong predictive model by combining multiple weak learners, typically decision trees, in a sequential manner. Like the regular gradient boosting algorithm, XGBoost also uses decision trees as its base estimators. It enables you to adjust the maximum size of the trees to reduce the possibility of overfitting the data. XGBoost also constructs many trees and determines the final prediction using all the trees. Each tree's value is scaled by the learning rate, allowing the algorithm to develop more gradually at each step [5].

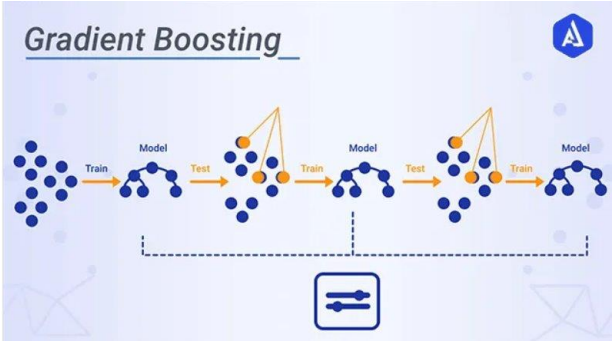


Fig. 4 Implementation of Gradient boosting

The process begins by preprocessing the dataset, where irrelevant or redundant features are removed, and the data is split into training and testing sets. Then, Gradient Boosting is applied to the training data, where a series of decision trees are built sequentially, each one focusing on correcting the errors of its predecessors. During training, each subsequent tree is trained on the residual errors of the previous ones, gradually improving the overall model's predictive capability. The performance of the Gradient Boosting model is evaluated using cross-validation techniques to estimate its accuracy, precision, recall, and F1-score on the training data.

Gradient Boosting trains each weak learner sequentially, with each subsequent learner focusing on the errors made by the ensemble of previous learners. This iterative process allows the model to gradually improve its performance by placing more emphasis on instances that are challenging to classify correctly. During each iteration, Gradient Boosting optimizes the model's parameters by minimizing a loss function, typically the gradient of the loss function with respect to the predictions made by the ensemble. By iteratively adjusting the model's parameters, Gradient Boosting aims to reduce the overall prediction error.

### C. *Graph Neural Network*

Graph Neural Networks (GNNs) are a class of deep learning models specifically designed to operate on graph-structured data, making them well-suited for analyzing network data such as firewall logs. GNNs excel at capturing complex dependencies and relationships within the network, enabling them to detect anomalous patterns that may span multiple nodes or connections. The graphs are represented with both structured and unstructured data, making GNN an important tool to model various real-time data such as data from network systems and text recognition systems.[4]

During the implementation of GNN, the dataset is split into training and testing sets. Then, the data is converted into PyTorch tensors and organized into a custom dataset using the Custom Dataset class. DataLoader is utilized to manage data loading for training and testing. The GNNModel is instantiated with input, hidden, and output dimensions. Cross-entropy loss is chosen as the loss function, and Adam optimizer is used for optimization. The model is trained over multiple epochs, iterating through the training dataset in mini batches, computing the loss, and updating the model parameters using backpropagation.

The key operation in GNNs is message passing, where information is exchanged between neighboring nodes in the graph. At each layer of the GNN, nodes aggregate information from their neighbors, typically through a learnable aggregation function, which allows them to update their own representations based on the information received. GNNs often employ graph convolutional layers to perform message passing efficiently. These layers compute node representations by aggregating features from neighboring nodes and applying a transformation function. By iteratively applying multiple graph convolutional layers, GNNs can capture increasingly complex dependencies within the graph structure.

### D. Equations

Due to the performance metric calculation, we had to use several respective formulae to get the respective values.

- Recall

$$\frac{(TP+TN)}{(TP+TN+FP+FN)}$$

TP + TN being the number of correct predictions which is divided by All predictions.

- Precision

$$\frac{TP}{TP+FP}$$

TP being the number of correct positive predictions divided by total number of positive prediction instances predicted.

- Accuracy

$$\frac{TP}{TP+FN}$$

- F1 Score

$$\frac{TP}{TP+0.5(FP+FN)}$$

### E. Dataset

Our dataset is an Internet Firewall Data featuring the following attributes:

- Source Port – Sender
- Destination Port – Receiver
- NAT Source
- NAT Destination
- Bytes
- Bytes Sent
- Bytes Received
- Packets
- Elapsed Time
- Pkts_sent
- Pkts_received
- Action

Our focus was primarily on the Action attribute, which denotes whether a connection was allowed (“allow”) or denied (“deny”). By examining this attribute, we were able to identify anomalous connections within the dataset. By analyzing this data, we were able to identify irregular connections based on whether they were allowed or denied. This insight into the action taken for each connection helped us to detect anomalous behavior within the firewall data.

## IV. Performance Analysis

The performance analysis of the anomaly detection algorithms, Isolation Forest, Gradient Boosting, and Graph Neural Networks (GNN), reveals distinctive strengths and capabilities in identifying anomalous activities within firewall data. Across key metrics such as F1 score, precision, recall, and accuracy, Gradient Boosting demonstrates remarkable effectiveness, achieving near-perfect scores across all metrics. Its ability to accurately classify anomalies while maintaining high precision underscores its suitability for real-world cybersecurity applications. Isolation Forest also exhibits strong performance, particularly in precision and recall metrics, albeit with slightly lower scores compared to Gradient Boosting. However, its perfect precision indicates its reliability in flagging anomalies with minimal false positives. Meanwhile, Graph Neural Networks (GNN) showcase competitive performance, balancing high accuracy with respectable precision and recall scores. Although not as high-performing as Gradient Boosting, GNN offers a promising alternative, leveraging network structure for anomaly detection.

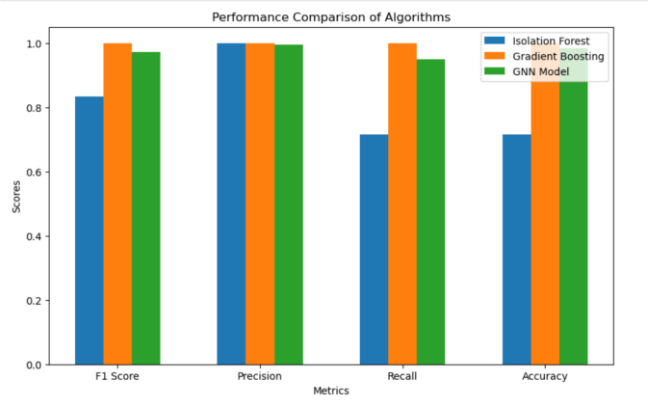


Fig.5 Performance comparison

TABLE 1 : PERFORMANCE TABLE

| *METRIC* | *ISOLATION FOREST* | *GRADIENT BOOSTING* | *GNN* |
|---|---|---|---|
| ***F1 SCORE*** | 0.834022 | 0.99916332 | 0.971197575 |
| ***PRECISION*** | 1 | 0.998996152 | 0.994137931 |
| ***RECALL*** | 0.715298231 | 0.999330544 | 0.949292065 |
| ***ACCURACY*** | 0.715298231 | 0.999524952 | 0.983754513 |

These findings underscore the importance of selecting appropriate algorithms based on specific use cases and desired performance metrics. Ultimately, the comprehensive evaluation presented in the performance table provides valuable insights for cybersecurity practitioners, guiding the selection and deployment of anomaly detection models to bolster network security and resilience against evolving cyber threats.

## V. Applications

Real-time applications leveraging anomaly detection in firewall data offer critical solutions for enhancing cybersecurity across various domains. Financial institutions can deploy these systems to monitor transactions and detect unauthorized access attempts, safeguarding sensitive financial data from potential breaches and fraudulent activities. Similarly, healthcare organizations can utilize such applications to protect patient records and ensure compliance with data privacy regulations, mitigating the risk of unauthorized access to medical information. In the e-commerce sector, real-time anomaly detection can prevent fraudulent transactions and protect customer data during online purchases, bolstering consumer trust and confidence in digital transactions. Moreover, critical infrastructure operators, such as energy utilities and transportation networks, can implement these applications to monitor network traffic and identify potential cyber threats, safeguarding essential services and infrastructure from malicious attacks. By continuously analyzing firewall data in real time and detecting anomalies using advanced algorithms like Isolation Forest, Gradient Boosting, and Graph Neural Networks, these applications enable proactive threat mitigation and strengthen overall cybersecurity posture, thereby ensuring the integrity and security of digital ecosystems in today's interconnected world.

## VI. Future Scope

The results obtained from our study showcase promising performance metrics for anomaly detection using Isolation Forest, Gradient Boosting, and Graph Neural Networks (GNN). Moving forward, there are several ways for expanding and enhancing the application of these anomaly detection models in cybersecurity operations. Integrating these models with Security Information and Event Management (SIEM) systems presents a significant opportunity to bolster cybersecurity defenses. By leveraging the capabilities of SIEM systems to collect, correlate, and analyze security event data from various sources, anomaly detection models can provide real-time insights into network traffic anomalies. This integration enables security analysts to receive timely alerts and investigate potential threats promptly.

To facilitate real-time anomaly detection, it is imperative to establish a scalable infrastructure capable of handling high volumes of data and processing demands. Cloud-based platforms such as Amazon Web Services (AWS), Google Cloud Platform (GCP), or Microsoft Azure offer scalable computing resources and services suitable for deploying and managing real-time analytics applications. By leveraging cloud-based infrastructure, organizations can effectively manage computational resources and ensure the scalability and reliability of their anomaly detection systems.

Furthermore, the integration of automated response mechanisms into the deployment pipeline enhances the proactive nature of cybersecurity operations. Automated response mechanisms can be triggered in response to detected anomalies, initiating immediate actions to mitigate potential security breaches. For instance, suspicious network traffic patterns could trigger automated firewall rules to block or quarantine the source of the anomaly, thereby reducing the risk of a successful cyber attack.

## VII. conclusion

In conclusion, the performance evaluation of anomaly detection algorithms—Isolation Forest, Gradient Boosting, and Graph Neural Networks—reveals valuable insights into their efficacy in safeguarding against cyber threats within firewall data. While each algorithm demonstrates varying degrees of effectiveness in detecting anomalies, Gradient Boosting emerges as the top performer, boasting impressive metrics across F1 score, precision, recall, and accuracy. Its near-perfect precision and high recall underscore its ability to accurately identify anomalies while minimizing false positives. However, Graph Neural Networks also exhibit strong performance, particularly in terms of accuracy, indicating their potential for detecting subtle deviations in network traffic patterns. Isolation Forest, although effective, lags behind in recall compared to Gradient Boosting and Graph Neural Networks. Nevertheless, its performance remains respectable, especially considering its role in identifying outliers within the data. These findings underscore the importance of leveraging advanced analytics techniques in bolstering cybersecurity resilience and proactively mitigating emerging threats. By deploying anomaly detection models trained on historical firewall data, organizations can enhance their ability to anticipate and respond to cyber threats in real-time, thereby safeguarding critical infrastructure and data assets from malicious actors. Continued research and refinement of anomaly detection algorithms will further empower organizations to stay one step ahead of cyber adversaries, ensuring the integrity and security of their digital ecosystems in an ever-evolving threat landscape.


## Acknowledgment

We would like to take this opportunity to thank Prof. Jiashu (Jessie) Zhao for giving us the opportunity and support to work on this project.


## References


[1] D. Niu, J. Zhang, L. Wang, K. Yan, T. Fu and X. Chen, "A Network Traffic anomaly Detection method based on CNN and XGBoost," *2020 Chinese Automation Congress (CAC)*, Shanghai, China, 2020, pp. 5453-5457, doi: 10.1109/CAC51589.2020.9327030.

[2] "Isolation Forest and Xg Boosting for Classifying Credit Card Fraudulent Transactions" Chandra Sekhar Kolli, T.Uma Dev, Volume-8 Issue-8, June 2019, Retrieval Number: E5795038519/19©BEIESP

[3] P. Kisanga, I. Woungang, I. Traore and G. H. S. Carvalho, "Network Anomaly Detection Using a Graph Neural Network," 2023 International Conference on Computing, Networking and Communications (ICNC), Honolulu, HI, USA, 2023, pp. 61-65, doi: 10.1109/ICNC57223.2023.1007411

[4] P. Kisanga, I. Woungang, I. Traore and G. H. S. Carvalho, "Network Anomaly Detection Using a Graph Neural Network," 2023 International Conference on Computing, Networking and Communications (ICNC), Honolulu, HI, USA, 2023, pp. 61-65, doi: 10.1109/ICNC57223.2023.10074111.

[5] Yolchuyev, "Extreme Gradient Boosting based Anomaly detection for Kubernetes Orchestration," 2023 27th International Conference on Information Technology (IT), Zabljak, Montenegro, 2023, pp. 1-4, doi:10.1109/IT57431.2023.10078576.