Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Special Issue 4, April : 2024

IDENTIFYING OUTLIERS IN NETWORK TRAFFIC USING IFOREST

Mrs. SK. Rahimunnisa1, Iragavarapu Venkata Sai Pranavi2, Devupalli Reshma3, Amarthaluri Hephsi4, Dundangi Kanishka Divakar5
1Assistant professor, Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India
2-5Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India

ABSTRACT
With the rapid expansion of networked systems and the ever-growing threat landscape, the need for robust anomaly detection mechanisms in network traffic has become paramount. Traditional rule-based and signature-based approaches often struggle to keep pace with the evolving nature of cyber threats. Unsupervised learning techniques offer a promising solution by autonomously identifying patterns of normal behavior and flagging deviations as potential anomalies without the need for labeled training data. This project presents a comprehensive review of unsupervised learning methods for anomaly detection in network traffic. The dimensionality reduction methods such as principal component analysis (PCA) and standard scalar facilitates efficient representation learning from high dimensional network data to generate new features. For models, Iforest is used as it is efficient and can separate training process from testing to predict the results of data. The recent advancements in deep learning-based approaches that is Artificial Neural Networks are used, which demonstrate promising results in capturing complex patterns in network traffic data.

Keywords: Unsupervised Learning, Dimensionality reduction, PCA, Iforest, Artificial Neural Networks.

INTRODUCTION
Nowadays, the number of networking devices is increasing at an exponential rate, and the workplace has a lot of devices that handle sensitive data communication. In recent years, the number of unknown attacks has increased rapidly both from inside and outside the organization. So, it has become imperative to provide customers and users secure access to the network and at the same time keeping the network attack free. In today's interconnected world, the volume of network traffic has skyrocketed, making it increasingly challenging to detect anomalous patterns or outliers amidst the vast amount of data. The identification of outliers in network traffic is crucial for ensuring the security and integrity of networks, as outliers often indicate potentially malicious activities or anomalies that require investigation. Traditional methods of anomaly detection often fall short in handling the complexities and scale of modern network environments, necessitating the development of more sophisticated techniques.

LITERATURE SURVEY

[1] Unsupervised Clustering Approach for Network Anomaly Detection
The paper titled "Unsupervised Clustering Approach for Network Anomaly Detection" likely presents a novel methodology for detecting anomalies within network data utilizing unsupervised clustering techniques. By leveraging clustering algorithms such as K-means, DBSCAN, or hierarchical clustering, the paper likely proposes a method to partition network data into clusters based on similarities, enabling the identification of deviations indicative of anomalies. Through the evaluation of performance metrics

like precision, recall, and F1-score, the authors likely assess the effectiveness and robustness of their approach in detecting anomalies in network traffic. Additionally, the paper may compare the proposed method with existing anomaly detection techniques, providing insights into its relative strengths and potential advantages. Furthermore, the discussion likely extends to real-world applications such as intrusion detection and network monitoring, elucidating the practical implications and relevance of the proposed approach in enhancing network security measures. Overall, this research is expected to contribute significantly to the field of anomaly detection in network security by introducing an innovative unsupervised clustering approach and elucidating its performance and applicability.

[2] An Unsupervised Deep Learning Model for Early Network Traffic Anomaly Detection

The paper "An Unsupervised Deep Learning Model for Early Network Traffic Anomaly Detection" by

Ren-Hung Hwang, Po-Ching Lin, Min-Chun Peng, and Chien-Wei Huang introduces an innovative approach to early detection of network traffic anomalies using unsupervised deep learning techniques. By focusing on deep learning methodologies, the study acknowledges the potential of these advanced techniques to effectively capture intricate patterns inherent in network data. The emphasis on early detection signifies a proactive stance towards cybersecurity, recognizing the critical importance of identifying anomalies at their nascent stages to mitigate potential risks and prevent security breaches. This research contributes to the ongoing efforts to enhance network security by introducing a sophisticated model capable of autonomously detecting anomalies in network traffic, thereby bolstering defenses against emerging cyber threats.

[3] Machine Learning Techniques for Anomaly Detection: An Overview

This paper presents an overview of research directions for applying supervised and unsupervised methods for managing the problem of anomaly detection. The paper "Machine Learning Techniques for Anomaly Detection: An Overview" authored by Salima Omar, Asri Ngadi, and Hamid H. Jebur from University Teknologi Malaysia provides a comprehensive survey of machine learning techniques employed in anomaly detection. By examining a wide range of algorithms and methodologies, the authors offer readers a broad understanding of the landscape of anomaly detection using machine learning. Through evaluating the performance and applicability of various techniques across different domains and datasets, the paper likely sheds light on the strengths, weaknesses, and suitability of different approaches based on factors such as data characteristics and computational requirements. Moreover, the authors likely identify challenges encountered in anomaly detection and propose future research directions, addressing issues such as scalability, interpretability, and adaptability to evolving threats. Overall, this paper serves as a valuable resource for researchers and practitioners in the field, providing insights into the current state, challenges, and future prospects of anomaly detection using machine learning techniques.

[4] IoT Botnet Anomaly Detection Using Unsupervised Deep Learning

This paper proposed an IoT botnet anomaly-detection solution based on a deep autoencoder model for detecting anomalies indicative of IoT botnet activity. Given the rising threat of botnet attacks targeting Internet of Things (IoT) devices, the paper likely addresses the need for robust anomaly detection mechanisms to safeguard IoT ecosystems. By leveraging unsupervised deep learning algorithms, such as autoencoders or generative adversarial networks (GANs), the study likely proposes a method to autonomously identify abnormal patterns within IoT network traffic that may signify botnet activity. The paper likely includes an evaluation of the proposed approach's performance using real-world IoT datasets, assessing its ability to accurately detect botnet-related anomalies while minimizing false positives. Furthermore, the research likely contributes to the advancement of cybersecurity measures for IoT

environments by introducing an innovative approach to detect and mitigate the threat of botnet attacks through unsupervised deep learning techniques.

EXISTING SYSTEM

The existing system for anomaly detection in network traffic using unsupervised machine learning relies on a variety of methods and algorithms.
Density-based anomaly detection relies on the assumption that normal data points tend to cluster together in dense neighborhoods, while anomalies are often isolated or located in sparse regions. This approach typically employs two main algorithms: k-nearest neighbor (k-NN) and local outlier factor (LOF). In k-NN, data points are classified based on similarities in distance metrics like Euclidean, Manhattan, Minkowski, or Hamming distance, making it a straightforward non-parametric technique. On the other hand, LOF assesses the relative density of data points by considering their reachability distance, providing a measure of how much an individual data point differs from its local neighborhood. These density-based methods offer effective ways to identify anomalies by examining the distribution of data points in relation to their surrounding context.

Clustering-based anomaly detection leverages the fundamental assumption that data points exhibiting similarity tend to belong to the same clusters, defined by their proximity to local centroids. K-means, a ubiquitous clustering algorithm, partitions data into 'k' clusters based on similarity. Instances diverging from these established clusters may signal anomalies within the dataset. By identifying data points that lie outside the defined clusters, anomalies can be detected and flagged for further investigation or action. This approach offers a powerful means to uncover irregularities in datasets without requiring labeled examples, making it

particularly valuable in scenarios where labeled data is scarce or unavailable.

Disadvantages

While existing systems for outlier detection in network traffic using unsupervised learning are valuable, they also encounter several challenges and limitations:
High False Positive Rates: Unsupervised learning approaches can often result in high false positive rates, where normal network behavior is misclassified as anomalous. This can lead to alert fatigue for network administrators and make it difficult to differentiate between benign outliers and genuine security threats.
Difficulty in Interpreting Results: Unsupervised learning models often lack transparency and interpretability, making it challenging to understand why certain network traffic patterns are classified as outliers. This makes it difficult for network administrators to take appropriate action and may lead to distrust in the detection system.
Scalability Issues: Analyzing large volumes of network traffic data in real-time can be computationally intensive and may require significant resources. Existing algorithms may struggle to scale effectively to handle the increasing volume, velocity, and variety of network data generated by modern networks.
Imbalance in Data Distribution: Network traffic data is often highly imbalanced, with normal traffic patterns occurring much more frequently than anomalous patterns. Existing models may struggle to identify outliers effectively in such imbalanced datasets, leading to biased detection results.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Special Issue 4, April : 2024

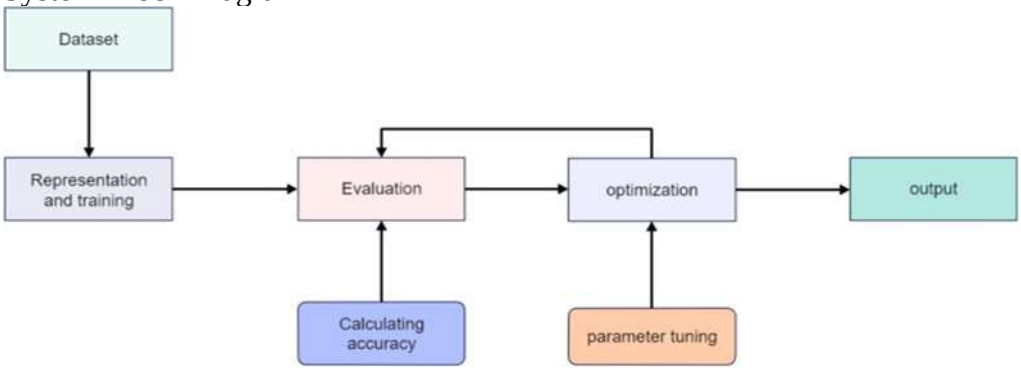
PROPOSED SCHEME

The proposed system aims to improve the accuracy of Isolation Forest algorithm for outlier detection, leveraging its ability to discern outliers in large and imbalanced datasets. Through meticulous preprocessing and model tuning, the system aims to enhance detection accuracy and mitigate false positives, thereby fortifying network security against emerging threats. The proposed system aims to enhance the accuracy of the Isolation Forest algorithm for outlier detection in network traffic analysis. By employing meticulous preprocessing techniques and fine-tuning model parameters, the system seeks to improve detection accuracy while minimizing false positives, thereby bolstering network security against emerging threats. The key innovation lies in the integration of Artificial Neural Networks (ANNs) to augment the learning process. By training ANNs on normal network traffic data, the system leverages their ability to capture complex patterns and relationships within the data, thus enhancing the overall robustness and effectiveness of the outlier detection system.

ADVANTAGES

Unsupervised learning techniques offer the advantage of not requiring labeled data for training, making them particularly well-suited for anomaly detection in network traffic where labeled examples of anomalies may be scarce or unavailable. By leveraging the inherent structure and patterns within the data, unsupervised learning algorithms can autonomously identify deviations from normal behavior, enabling the detection of previously unseen or novel threats. The utilization of unsupervised learning techniques, such as Isolation Forests and ANNs, offers several advantages in the context of anomaly detection in network traffic. Firstly, the absence of a requirement for labeled data during training makes these techniques particularly well-suited for scenarios where labeled examples of anomalies may be scarce or unavailable. This autonomy enables the system to identify deviations from normal behavior autonomously, facilitating the detection of previously unseen or novel threats. Additionally, Isolation Forests provide a scalable and efficient solution for identifying anomalies in high-dimensional data, such as network traffic, without necessitating extensive computational resources. By incorporating ANNs into the system, it becomes adept at learning intricate patterns within the data, thereby improving its ability to discern subtle anomalies that may evade traditional detection methods. This combined approach enhances the system's capacity to detect sophisticated attacks and bolster network security measures effectively.

System Block Diagram



OUTPUT SCREENS:

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Special Issue 4, April : 2024

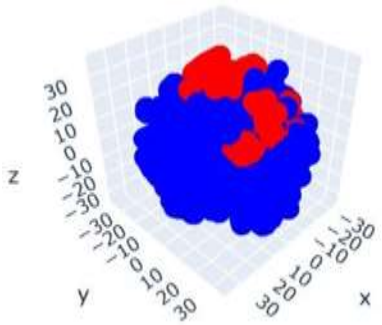


Fig1:3-D visualization

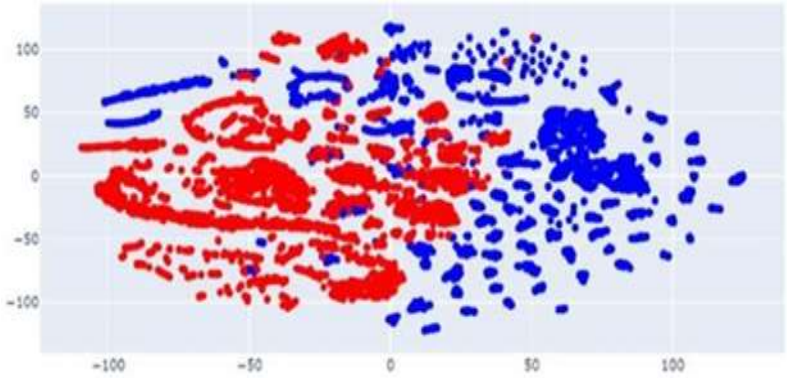


Fig2: 2-D visualization

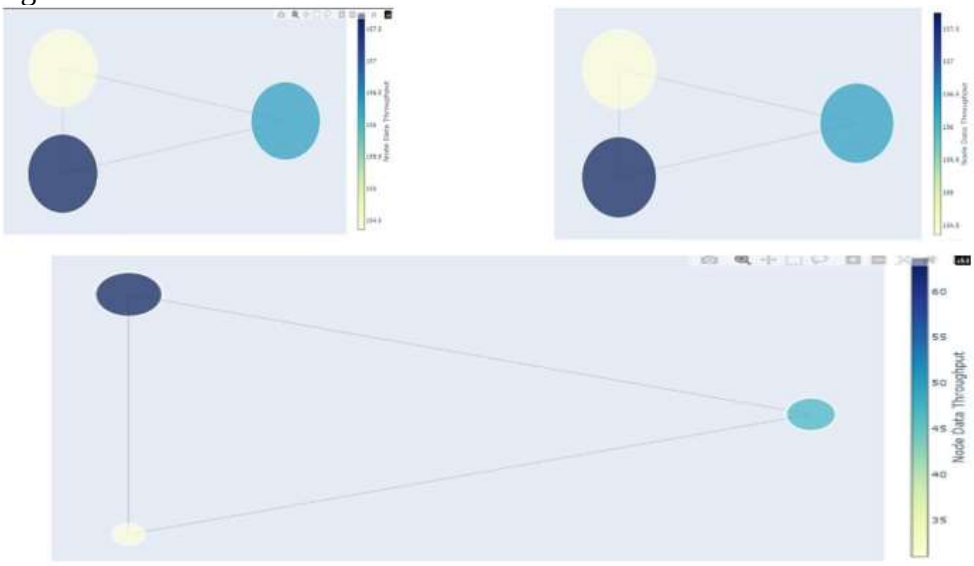


Fig 3: Anomaly Data Nodes

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Special Issue 4, April : 2024

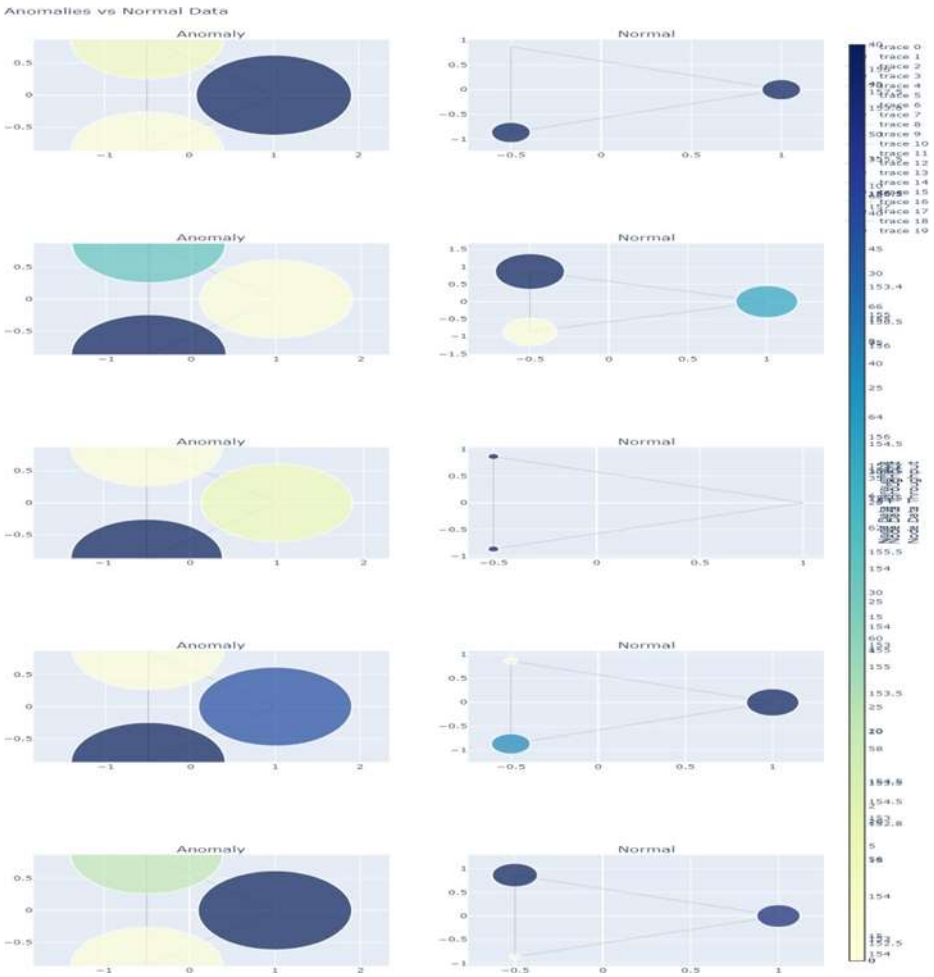


Fig4: Anomaly Vs Normal Data nodes

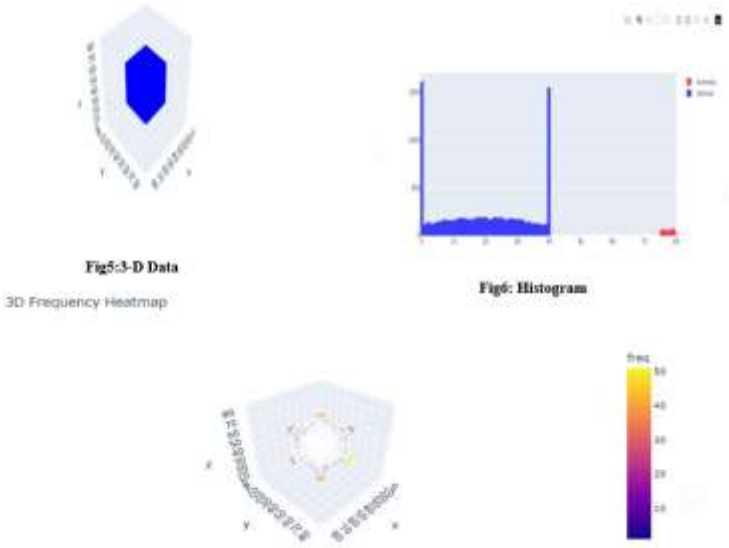


Fig 7: Frequency Heatmap

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Special Issue 4, April : 2024

TRAIN CLASSIFICATION REPORT

	precision	recall	f1-score	support
anomaly	0.38	0.00	0.00	9394
normal	0.53	1.00	0.70	10759
accuracy			0.53	20153
macro avg	0.45	0.50	0.35	20153
weighted avg	0.46	0.53	0.37	20153

TEST CLASSIFICATION REPORT

	precision	recall	f1-score	support
anomaly	0.33	0.00	0.00	2349
normal	0.53	1.00	0.70	2690
accuracy			0.53	5039
macro avg	0.43	0.50	0.35	5039
weighted avg	0.44	0.53	0.37	5039

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Special Issue 4, April : 2024

	precision	recall	f1-score	support
anomaly	1.00	1.00	1.00	2349
normal	1.00	1.00	1.00	2690
accuracy			1.00	5039
macro avg	1.00	1.00	1.00	5039
weighted avg	1.00	1.00	1.00	5039

PCA

	precision	recall	f1-score	support
anomaly	1.00	0.99	0.99	2349
normal	0.99	1.00	1.00	2690
accuracy			0.99	5039
macro avg	1.00	0.99	0.99	5039
weighted avg	0.99	0.99	0.99	5039

AUTOENCODER

	precision	recall	f1-score	support
anomaly	1.00	0.99	0.99	2349
...				
accuracy			0.99	5039
macro avg	0.99	0.99	0.99	5039
weighted avg	0.99	0.99	0.99	5039

CONCLUSION:

An unsupervised machine-learning model was built due to highly imbalanced data. Performance metrics are computed for the both algorithms. There is tremendous growth in the different types of network attacks and thus organizations are developing Intrusion Detection System (IDS) that are not only highly efficient but also capable of detecting threats in real time. In the course of implementation, it has become evident that refining the anomaly detection process involves exploring diverse parameter values across the algorithms employed. Moreover, the quality and completeness of the dataset significantly influence the efficacy of anomaly detection outcomes, with cleaner datasets yielding superior results. Notably, the contamination parameter plays a pivotal role in determining the proportion of anomalies effectively identified. However, it's imperative to acknowledge that while machine learning and deep learning applications hold promise in network security, their integration is still in its nascent stages. Consequently, challenges persist, particularly regarding scalability and efficiency, underscoring the ongoing need for advancements in this domain to meet the evolving demands of cybersecurity in networked environments.

REFERENCE:

[1] IoT Botnet Anomaly Detection Using Unsupervised Deep Learning (2021) by Ioana Apostol, Marius Preda, Constantin Nila

[2] Unsupervised Anomaly Detection Based on Deep Autoencoding and Clustering (2021) by Jiangtao Liu,
Wei Chen Chuanlei Zhang Xiaoning Yan, Jinyuan Shi, Minda Yao Nenghua Xu, and Dufeng Chen

[3] A Hybrid Unsupervised Clustering-Based Anomaly Detection Method (2021) by Guo Pu, Lijuan Wang,
Jun Shen, Fang Dong

[4] Anomaly detection in Network Traffic Using Unsupervised Machine Learning Approach (2020) by Aditya Vikram, Mohana

[5] An Unsupervised Deep Learning Model for Early Network Traffic Anomaly Detection (2020) by Ren-
Hung hwang, Min-Chun Peng, Chien-wei Huang, Po-Ching Lin, Van-Linh Nguyen

[6] Machine Learning Techniques for Anomaly Detection: An Overview (2013) by Salima Omar, Asri Ngadi,
Hamid H. Jebur

[7] Unsupervised Clustering Approach for Network Anomaly Detection (2012) by Iwan Syarif, Adam Prugel- Bennett and Gary Wills