INTRUSION DETECTION SYSTEM USING COHEN'S D AND WILCOXON TEST

Thesis Submitted
in Partial Fulfilment of the Requirements for the
Degree of

MASTER OF SCIENCE(M.SC.)

in

MATHEMATICS

by

KOMAL
(Roll No. 2k22/MSCMAT/22)

Under the supervision of
DR. ANSHUL ARORA
Department of Applied Mathematics

To the
Department of Applied Mathematics

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

June, 2024

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Komal , Roll No. 2K22/MSCMAT/22 student of Master in Science (Mathematics), hereby certify that the work which is being presented in the thesis entitled INTRUSION DETECTION SYSTEM USING COHEN'S D AND WILCOXON TEST in partial fulfilment of the requirement for the award of the Degree of Master of Science in Mathematics, submitted in the Department of Applied Mathematics, Delhi Technological University, Delhi is an authentic record of my work carried out during the period from August 2023 to May 2024 under the supervision of Dr. Anshul Arora**.**

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Candidate's signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor **Signature of External Examiners**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE BY THE SUPERVISIOR

Certified that **Komal** (2K22/MSCMAT/22) has carried out their research work presented in the thesis entitled **"INTRUSION DETECTION SYSTEM USING COHEN'S D AND WILCOXON TEST"** for the award of **Master of Science** from Department of Applied Mathematics, Delhi Technological University, Delhi, under my supervision. The thesis embodied results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other university/ institution.

Signature
Dr. Anshul Arora
Assistant Professor
Delhi Technological University

Date: 6 June 2024

INTRUSION DETECTION SYSTEM USING COHEN’S D AND WILCOXON TEST

Komal

ABSTRACT

In an age where cyber-attacks and malicious activities are at their peak, cybersecurity is crucial for detecting network intrusions and preventing unauthorized access to sensitive data. This thesis underscores the importance of network security, highlighting that network attacks can cause significant financial and operational losses for companies, organizations, and individuals. Traditional defences like antivirus software and firewalls, once sufficient, are now inadequate against the evolving nature of these threats. These conventional tools have proven ineffective in protecting network systems from increasingly sophisticated attacks and malware. This situation necessitates intelligent countermeasures to maintain the security of networks and critical systems. The objective of this work is to create a robust system designed to identify intrusions by analysing network traffic.

Chapter 1 introduces the concept and function of intrusion detection systems (IDS). Intrusion detection entails monitoring network traffic and computer activities to detect any malicious or unauthorized behaviour. An IDS can be either a device or a software application that performs this detection. Unlike firewalls, which focus on protecting the network perimeter, IDSs scrutinize activities within the protected network. IDSs can be categorized into three types: signature-based, anomaly-based, and hybrid systems. For an IDS to be effective, it must be efficient, adaptable, and scalable. The chapter concludes with an exploration of the limitations and challenges associated with current systems.

Chapter 2 reviews the literature on network traffic-based intrusion detection, examining over 50 research papers.

In Chapter 3, we present the 11 features utilized in our research, detailing their names and meanings. We conduct statistical tests to evaluate and rank these features according to their effectiveness in detecting network intrusions. Our objective is to

determine a set of features that together provide higher accuracy than any single feature or any other combination of features .We use two statistical tests: Cohen’s d, which assesses the effect size by comparing the means of various sample data, and the Wilcoxon test, which determines whether there are significant differences between paired samples. We then discuss the purpose of the machine learning classifiers used in our research: SVM, Navies Bayes, Logistic Regression and Decision Tree. Features are ranked from least to most efficient, creating three columns: one for Cohen’s d , one for Wilcoxon test. We also prepare a table where each feature is individually evaluated using all three classifiers. To investigate the potential effectiveness of different feature combinations, we evaluate various combinations using all four classifiers, producing four distinct sets of results. This method is repeated across the four sets of results to identify the final set of features that are most effective for network intrusion detection.

Chapter 4 presents the tables, calculations, and evidence supporting our conclusion that a specific combination of five features—"Floe duration”, “Packets received per second," "Average packet size received," "Average Packet size sent," and "Average packet size"—achieved the highest accuracy of 99.70% when using the Decision Tree classifier, outperforming all other features and the combinations.

Chapter 5 concludes the thesis and outlines potential directions for future research.

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

Acknowledgement

My supervisor, **Dr. Anshul Arora** of the Department of Applied Mathematics at **Delhi Technological University**, has my sincere gratitude for his meticulous guidance, profound expertise, constructive criticism, attentive listening, and amiable demeanour have been invaluable throughout the process of composing this report. I am eternally grateful for his benevolent and supportive approach, as well as his perceptive counsel, which played a pivotal role in the successful culmination of my project. Furthermore, I would like to express my appreciation to all my classmates who have played a pivotal role in aiding me to complete this endeavour by offering assistance and facilitating the exchange of pertinent information.

KOMAL
2K22/MSCMAT/22

List of Tables

List of Figures

TABLE OF CONTENTS

Chapter 1
INTRODUCTION

Today, cybersecurity is probably the most important issue that businesses and organizations have to deal with. The number of attacks between professional organizations or private network organizations continues to increase. While new cases emerge every day, piles of old cases still persist. The effectiveness of a cyber-attack can be significant, whether in a corporate environment or an individual organization. In many cases, cyber-attacks can cause material and physical losses. As the web continues to evolve, the number and variety of computer attacks continues to increase, ransomware is more proliferating than ever, and abuse has never been more important than attracting media attention. Currently, antivirus and firewalls are not enough to keep organizations safe, and organizations need to implement multiple layers of security. The most important system is provided by the Intrusion Detection System (IDS), which is designed to ensure that its target is protected against attacks by consistent monitoring of the frame.

1.1 The Security Problem

In today's threats, networks and computers have become the focus of security issues in the last few years. Cybersecurity consists of a set of tools and procedures designed to protect computers, networks, software and data from attacks, unauthorized access, alteration or destruction. Access is an attempt to maintain confidentiality, integrity, or system availability. Finding access to distribution, restrictions, and changing environments is a difficult task. Given these challenges, most people cannot make definitive decisions on the spot. In addition, the profile, goals and capabilities of intruders have changed significantly. Historically, cyber intruders are generally thought to be young, social individuals motivated by curiosity, criminality, and, more commonly for hackers, familiarity with friends. Although there are many clever intruders, most intruders do not have the financial resources to carry out an attack. But today we face many challenges and motivations. For example, the use of advance persistent threats (APTs) using advanced technology members has become commonplace.

Security risks to network infrastructure can be divided into active risks and passive risks. In the case of an active threat, attackers will try to disrupt network operations. In contrast, in passive threats, the intruder remains hidden and aims to disrupt the message exchange. Traditional security measures such as firewalls, security policies, access control and encryption have proven inadequate in protecting networks and systems against attacks and malware. In addition, security teams face great challenges in managing large amounts of data. The main goal is to reduce the risk of loss as much as possible. This challenge requires smarter protection to secure networks and critical systems. It has practical use in every field. That's why the security group started using machine learning to make it easier to fight against attackers. Automatic detection of malicious activity on the network has great potential to improve security. Additionally, data extraction can reveal new attack strategies, intrusion scenarios, and identify attackers' targets and methods. This improves the ability to distinguish attacks from legitimate claims. Machine learning models continuously learn from data, identifying patterns and anomalies that indicate potential threats, enabling faster and more accurate responses. This is especially important in an environment where threats are constant and new vulnerabilities are discovered every day. Therefore, the integration of machine learning with traditional security measures can provide a stronger and more comprehensive defence strategy, ultimately increasing the strength of networks and systems against cyber threats.

1.2 Intrusion Detection System

Intrusion detection is the systematic process of analysing network traffic and computer activities to identify malicious or unauthorized actions. Devices or software designed for this purpose are known as Intrusion Detection Systems (IDS). IDSs play an important role in monitoring internal network activities, providing a deeper level of security than just perimeter defences. Unlike firewalls, which can actively block suspicious activities, IDSs are passive systems that detect and alert administrators to potential threats, requiring human intervention for response. IDS is divided into three main types based on detection methods: signature-based detection, anomaly-based detection, and hybrid detection.

- Signature-based detection (also known as "misuse detection") relies on a database of known attack signatures to identify threats. This method involves comparing incoming data with known patterns of malicious activity. When a match is found, an alert is triggered. This approach is similar to traditional antivirus programs, which scan for specific sequences of bytes or characters indicative of a threat. To remain effective, the signature database must be regularly updated to include the latest known threats. However, this method can only detect known attacks and may miss new, unknown exploits.
- Anomaly-based detection builds a baseline of normal system behaviour by analysing regular activities and patterns. Any deviation from this established norm is flagged as a potential threat. This method does not rely on a pre-existing database of signatures and is capable of identifying novel attacks. However, it can generate a high number of false positives, as benign activities that deviate from the norm can also be flagged. This makes it challenging to distinguish between genuine threats and harmless anomalies, often requiring further analysis.
- Hybrid detection combines the strengths of both signature-based as well as anomaly-based detection methods to provide a more comprehensive security solution. This approach can either sequentially or simultaneously apply both techniques, using signature-based detection to identify known threats quickly, while anomaly detection monitors for unusual activities that might indicate new types of attacks. By integrating both methods, hybrid detection systems aim to reduce the weaknesses inherent in each approach, such as the false positives of anomaly detection and the inability of signature-based detection to catch unknown threats.

Moreover, advanced technologies like machine learning and artificial intelligence are enhancing IDS capabilities. Machine learning algorithms can analyse large amounts of data to identify patterns and more accurately predict potential threats. These algorithms can be trained to recognize subtle indicators of attacks that might be missed by conventional methods, improving the overall detection rate and reducing false positives.

Additionally, the development of more automated response systems is advancing the evolution of IDS. These systems can take pre-defined actions in response to certain alerts, such as isolating affected parts of the network or initiating data backups, thereby minimizing the time between detection and response. This is particularly important in mitigating the impact of fast-moving threats, such as ransomware, which can spread rapidly through networks.

As a result, IDS is an essential part of a modern network security strategy, providing significant monitoring and alerting capabilities to protect the network from known and unknown threats. By employing a combination of signature-based, anomaly-based, and hybrid detection methods, and incorporating advanced technologies like machine learning, IDSs can offer robust protection against an increasingly complex threat landscape.

1.3 The role of Intrusion Detection

Various detection and protection strategies have been developed over time, but the core component of an effective defence system remains the Intrusion Detection System (IDS). An IDS is essential as it identifies potential threats before they inflict significant damage on the network. Intrusion refers to any unauthorized attempt to access network components with the intention of causing harm or corruption.

For an IDS to be effective, it must have key attributes: efficiency, adaptability, and extensibility. Efficiency involves the system's ability to detect even subtle variations of known attacks. Adaptability means the IDS can adjust to different environments and evolving threats. Extensibility allows the IDS to be easily modified and expanded to address new security challenges.

Most IDSs historically have relied on signature detection due to their effectiveness in identifying known threats. Notable examples include Snort and ClamAV, both of which have extensive signature databases—Snort with over 4,000 rules and ClamAV with more than 800,000 signatures. Despite their effectiveness, these systems face two significant challenges:

- Manual Signature Creation: Creating and updating signatures manually is labour-intensive and costly, requiring continuous human effort to maintain the database. This process introduces delays between encountering new threats and updating the system to counter them.
- Ineffectiveness Against New Threats: Signature-based IDSs are not equipped to detect new, previously unknown threats because they rely on pre-existing signatures.

Automating signature generation using machine learning algorithms offers a promising solution to these issues. Machine learning provides adaptability, fault tolerance, high processing speed, and resilience against noisy data. It enables continuous updates to the signature database in dynamic environments, employing incremental learning to recognize and respond to new threats as they arise.

Integrating machine learning into IDSs also enhances their adaptability to rapidly changing environments, which is crucial for maintaining their extensibility. This allows IDSs to evolve alongside emerging security threats.

Furthermore, machine learning can help reduce possibility of false positives and improve the accuracy of threat detection. By analysing extensive data sets, these systems can identify subtle patterns and anomalies indicative of potential intrusions, providing a more robust defence mechanism.

In conclusion, the labour-intensive process of manual signature creation and the inability of traditional IDSs to adapt to dynamic environments highlight the need for innovative research. Emphasizing adaptability and extensibility in IDS design is crucial for developing systems capable of keeping pace with evolving threats, representing a new direction in IDS research and underscoring the motivation for ongoing advancements in this field.

Contribution of the Thesis:

Considering these constraints, This Thesis introduces an anomaly-based intrusion detection model. Here are the key contributions of our work:

1. We conducted feature extraction from network traffic data, comprising both normal and intrusion scenarios.

2. Employing statistical methods that is Cohen's d and the Wilcoxon test, we ranked these extracted features to discern their significance.

3. Our thesis presents a unique algorithm for intrusion detection by both normal and malware data, leveraging machine learning classifiers on the prioritized features.

Organisation of the Thesis:

The thesis is organized as follows: Chapter 2 reviews the existing literature on intrusion detection. Chapter 3 details the methodology of our proposed model. Chapter 4 analyses the results obtained from our model. Finally, Chapter 5 concludes the thesis and suggests directions for future research.

Chapter 2
Related Work

This chapter focuses on reviewing the existing literature concerning intrusion detection based on network traffic. Section 2.1 reviews various approaches and methodologies proposed in the literature for network traffic-based intrusion detection. Section 2.2 summarizes the findings of the chapter.

2,1 Related works in Intrusion Detection

The authors in [1] provides an overview of web intrusion detection techniques, including signature-based, anomaly-based, and hybrid approaches, amidst the rising cybercrime and increased internet usage. It emphasizes the importance of employing multiple detection methods to enhance accuracy and reduce false positives, while also discussing the challenges and latest trends, such as machine learning-based approaches, and offering recommendations for future improvements. The authors in [2] discussed the critical importance of network malware detection, emphasizing its capability to monitor incoming streams of unknown data and implement a warning-based system as a precautionary measure. Their system, designed to prevent unauthorized access, was based on FCMA and adept at modifying the detection framework and organizing traffic flow. Additionally, they developed a model using various Machine Learning algorithms. Author in [3] provided valuable insights into methods for identifying and selecting features in the domain of network security. This paper introduces IP-filtered multi-channel convolutional neural networks (IP-MCCLSTM) as an innovative solution to enhance network intrusion detection efficiency amid escalating bandwidth challenges. Through filtering traffic by IP and minimizing system loading, IP-MCCLSTM achieves superior performance, with an accuracy of 98.9% and a Macro-Recall rate of 99.7%, as evidenced by tests on the 2017CICIDS dataset. The paper [4] addresses the challenge of intrusion detection in software-defined networks (SDNs) by proposing a hybrid model that integrates CNN and BiLSTM with an attention mechanism. By leveraging the unified vision of SDN and deep learning, the model effectively captures complex network traffic patterns, achieves high accuracy in identifying various intrusion types, and demonstrates superior performance compared to existing models like Alexnet, Lenet5, and CNN-LSTM. Evaluation on real-world SDN data confirms its efficacy, making it a promising solution for enhancing IDS security in SDN environments.

The Author in [5] presents an intrusion detection method using the naive Bayes algorithm to enhance detection in complex multi-dimensional node combination mixed topology networks. By establishing a distributed structure model and conducting traffic analysis, it extracts abnormal traffic features and detects intrusion data clusters. The proposed method demonstrates high accuracy in detecting intrusion data, showcasing robust detection capabilities and potential for bolstering network security against attacks. The paper [6] introduces AHDM, an autoencoder-based hybrid detection model, to address the challenge of detecting small-sample malicious traffic in the context of increasing cyber-attacks and IoT development. AHDM employs a dual classifier framework, utilizing neural networks trained on both encoded and original features to detect various types of traffic. Experimental results demonstrate AHDM's superior performance, particularly in detecting small-sample malicious traffic like DDoS attacks, surpassing traditional models such as DNN and ACID, with a notable advantage in the IOT-23 dataset. The Author in [7] introduces an optimal framework for network intrusion detection, leveraging image processing techniques to enhance security measures amidst evolving hacking methodologies. The framework combines feature selection, image transformation, and deep-learning-based anomaly detection, showcasing efficiency across diverse intrusion detection datasets. Comparative analysis with recent image-processing-based approaches underscores the effectiveness of the proposed framework in bolstering network security. The paper [8] addresses the imperative need for efficient and accurate network intrusion detection amidst the vast volumes of network data in the era of big data. It introduces a hybrid network classifier, integrating improved residual network blocks and bidirectional gated recurrent units, preceded by feature dimensionality reduction via an enhanced autoencoder. The paper [9] addresses the escalating network security challenges posed by IoT proliferation, expanding attack surfaces, and growing network heterogeneity. It proposes a novel approach of analysing 5G slice traffic as images using a CNN model enhanced by NAS for detecting anomalous network behaviour. Leveraging SDS, the paper designs a scalable and efficient 5G traffic defence system, achieving high accuracies of 94.9% and 96.4% in multiclass classification, with further optimization suggestions for enhanced CNN performance. The papers[10-12] collectively address challenges in network intrusion detection systems (IDS) using advanced machine learning techniques. One proposes a deep hierarchical network for packet-level malicious traffic detection, achieving high accuracy across diverse datasets. Another introduces a meta-learning framework-based method for few-shot scenarios, demonstrating universal applicability and robust performance with limited training data. Lastly, a system, MANDA, is proposed to detect adversarial attacks on ML-based IDSs by leveraging manifold evaluation and model uncertainty analysis, ensuring high true-positive rates against various attacks. These innovative approaches contribute to enhancing network security amid evolving cyber threats.

The work in [13-15] is a novel approach to addressing cybersecurity challenges in IoT-based critical infrastructures is proposed, leveraging an intelligent hybrid intrusion detection system (HIDS) architecture. The HIDS model demonstrates significant efficacy with an F1-score of 0.8171, precision of 0.8572, and recall of 0.8183, showcasing its ability to accurately categorize network traffic flows. Additionally, a framework called DGIDS is introduced to enhance network-based intrusion detection systems by generating semi-synthetic datasets, boosting detection quality from 54% to 90.39%.The study in [16-18] explores advanced methods for enhancing network security through intrusion detection systems (IDS). It highlights the efficacy of machine learning models like Decision Trees and Random Forests in anomaly detection, achieving over 99% accuracy. Additionally, it proposes AI-based solutions to address data imbalance issues in network intrusion detection systems, using generative models to improve detection accuracy. Finally, it introduces a complex network theory-based method for mapping communication traffic, demonstrating superior intrusion detection performance. The paper [19-21] addresses the challenges of industrial intrusion detection in Industry 4.0 using one-class broad learning systems (OCBLS) and stacked OCBLS (ST-OCBLS) algorithms, demonstrating efficient training and robust performance against diverse intrusions. Additionally, it introduces the BOTA system for explainable botnet detection and a computer vision-based approach for real-time intrusion detection in 6G networks, both achieving high accuracy and efficiency in their respective evaluations.

The paper [22] proposes a Tree-based BLS (TBLS) intrusion detection method to address cybersecurity risks in the Internet of Vehicles (IoV), exacerbated by the increased reliance on online activities during COVID-19. Tested on NSL-KDD and UNSW-NB15 datasets, TBLS demonstrates superior accuracy and lower false alarm rates compared to 16 existing solutions, effectively mitigating issues related to unbalanced data distribution. The paper [23] conducts a thorough analysis of the NSL-KDD network traffic dataset, employing various machine learning models for intrusion detection. Unlike traditional approaches, a hierarchical strategy is adopted, first classifying intrusion vs. normal behaviour and then identifying specific attack types. The study showcases the effectiveness of unsupervised representation learning and addresses data imbalance issues using SVM-SMOTE oversampling, while also highlighting the advantages and drawbacks of this technique in conjunction with deep neural networks. The research paper [24] investigates the implementation of the Random Forest algorithm in building a robust Intrusion Detection System (IDS) to combat the escalating threats faced by computer networks. By leveraging ensemble learning, the proposed IDS efficiently detects diverse intrusion types in real-time network traffic, addressing the complexities of large-scale datasets. The study underscores the necessity for ongoing advancements in ensemble analytics to enhance IDS resilience against evolving cyber threats. The paper [25] examines the growth of Software Defined Networking (SDN) and its benefits, including simplified network management and high-level programming abstractions. However, the centralized decision-making of SDN can lead to performance issues due to the installation of flow rules against anomaly traffic. To address this, the paper proposes DMTA-SDN, a solution combining a signature-based IDS with the SDN controller, to detect and mitigate traffic anomalies. Extensive simulations demonstrate improved network performance, particularly in terms of Round Trip Time (RTT) and bandwidth. The paper [26] underscores the significance of Network-based Intrusion Detection Systems (NIDS) in safeguarding information security amidst a surge of network attacks.

Leveraging the advancements in neural networks, the study introduces a novel approach—flow-based NIDS utilizing Graph Neural Network (GNN)—to better utilize network traffic data. By representing traffic data in graph structures and employing a discriminator to address data imbalances, the proposed model demonstrates improved performance in detecting attacks across binary and multi-class classification tasks, highlighting the efficacy of graph structures in enhancing intrusion detection capabilities. The paper [27] highlights the importance of intrusion detection systems as a key defence mechanism against cyber-attacks, particularly network-based systems for real-time detection of abnormal behaviours. By employing deep learning techniques, specifically convolutional neural networks, to analyse and process data traffic, the study demonstrates improved system efficiency, as validated by experimental results. The paper [28] addresses the growing concern of cybersecurity threats to Internet of Things (IoT) devices by proposing a lightweight and efficient intrusion detection method based on feature grouping. By designing a fast protocol parsing method and implementing session merging and feature grouping techniques, the proposed method achieves over 99.5% classification accuracy on three public IoT datasets while requiring significantly fewer computational resources compared to baseline methods. Experimental results highlight the method's effectiveness, efficiency, and suitability for IoT intrusion detection. The paper [29] provides an overview of Network Security Monitoring (NSM) and introduces a novel taxonomy outlining the functionalities and modules within NSM systems. By categorizing popular tools based on this taxonomy, the paper offers valuable insights for both researchers and practitioners in assessing NSM deployments. Additionally, the paper addresses challenges in applying NSM to contemporary network architectures, such as Software Defined Network (SDN) and Internet of Things (IoT). The paper [30] delves into the challenges and complexities surrounding anonymity tools, highlighting their pivotal role in safeguarding user anonymity through encryption and obfuscation techniques. It provides a thorough review of existing approaches for categorizing anonymous and encrypted network traffic within the darknet, emphasizing the growing importance of employing machine learning techniques for monitoring and identifying potential traffic attacks within this realm. The research [31] introduces a novel deep learning framework for anomaly detection in the Internet of Everything (IoE), combining decomposition methods, deep neural networks, and evolutionary computation to enhance outlier detection in IoE environments. Through the utilization of clustering algorithms, DL architectures, and evolutionary computational algorithms such as genetic and bee swarm, the proposed framework demonstrates superior performance in two use cases: road traffic outlier detection and network intrusion detection, showcasing its advantages over existing approaches. The author in [32] research addresses the growing concern of cyber threats to in-vehicle networks (IVN) by proposing a novel self-supervised anomaly detection method. Utilizing deep learning models, the approach generates noised pseudo normal data and trains an anomaly detector to distinguish between normal and abnormal traffic. Experimental results showcase significant improvements in detecting unknown attacks and outperforming existing semi-supervised learning-based methods. The paper [33] introduces Manticore, an unsupervised intrusion detection system utilizing contrastive learning for 5G networks. By incorporating statistical and original packet features, Manticore captures comprehensive traffic information and automatically establishes positive and negative pairs without manual labelling. Experimental results on two datasets showcase Manticore's superior performance compared to existing methods, highlighting its efficacy in addressing the challenges of intrusion detection in complex 5G network environments. The paper [34-35] compares the performance of various unsupervised deep and machine learning-based anomaly detection algorithms for real-time detection of anomalies on the Automotive Ethernet-based in-vehicle network, demonstrating the superiority of deep learning models over traditional methods. Additionally, the paper introduces a hybrid algorithm, SDAID, combining signature and deep learning for intrusion detection, achieving exceptional F1-scores on well-known datasets CSE-CIC-IDS2018 and NSL-KDD, outperforming related models. The paper [36] proposes a causal deep learning-based network intrusion detection system (NIDS) to enhance stability and generalization. By optimizing causal weights and addressing feature correlations, the system improves detection accuracy in diverse network environments. Experimental results demonstrate a significant increase in stability, particularly with the use of binary coding features and causal intervention screening. The author [37] presents the ToN_IoT data set, emphasizing its importance for intrusion detection research in the Internet of Things (IoT). The study highlights the need for heterogeneous data sets, diverse data collection methods, and standardized feature descriptions and cyberattack classifications to improve detection performance and practical application in real-world environments. The paper [38] introduces EDA4GNeT, a novel method for early detection of GOOSE poisoning attacks in IEC 61850 substations, focusing on ensuring communication availability in critical infrastructure. Utilizing mathematical modelling and statistical hypothesis testing, EDA4GNeT effectively detects anomalies in network traffic. The method is validated through simulations of Denial of Service (DoS) attacks on a 66/11kV substation, demonstrating its efficacy compared to existing approaches. The research [39] introduces a unique testbed leveraging the panOULU Municipal public network in Oulu, Finland, to enhance AI-driven intrusion detection systems (IDS) in public networks. Utilizing edge-to-cloud infrastructures and Federated Learning (FED-ML), the study automates traffic data collection and model training to improve security. The publicly available dataset and configuration aim to deepen understanding of protecting diverse, interconnected networks. The study [40] presents an Intrusion Detection System (IDS) for detecting anomalies in vehicular CAN bus traffic by analysing message identifier sequences. Using data collected over several months from a heavy-duty truck, the IDS's performance was evaluated against various attack types, demonstrating high sensitivity and specificity, quick decision times, and an efficient memory footprint. The paper [41] presents a novel traffic sampling mechanism for SDN-capable networks using deep deterministic policy gradient (DDPG) to optimize resource allocation and reduce monitoring overheads. The proposed system effectively captures malicious flows and balances the load across multiple traffic analysers, as demonstrated through extensive simulations and testbed experiments. The author in paper [42] introduces a novel intrusion detection method combining feature selection with a Transformer model. By using univariate feature selection and recursive feature elimination, the model enhances its detection capabilities. The Transformer-based architecture encodes traffic features and classifies them with a neural network, utilizing a fusion loss function to improve performance and convergence. This approach outperforms existing methods on both public and self-built datasets. In paper [43] author presents an Intrusion Type Classifier (ITC) that uses machine learning to autonomously detect and classify various network intrusions, such as DoS, DDoS, and malware attacks. Utilizing extensive datasets, the ITC achieves high precision and recall rates, enhancing network security by enabling quick responses to potential threats. The support vector machine algorithm is highlighted for its strong performance in network infrastructure applications. The paper [44] examines the security challenges in edge computing environments, emphasizing the use of Machine Learning (ML) and Deep Learning (DL) techniques for real-time anomaly detection. It highlights recent advancements in Network Intrusion Detection Systems (NIDS), particularly using Graph Neural Networks (GNNs), and investigates the impact of adversarial attacks on NIDS, along with available defence mechanisms. The author [45] reviews the challenges and advancements in intrusion detection systems (IDS), focusing on machine learning (ML) and deep learning (DL) approaches to enhance detection accuracy and reduce false alarms. It provides a taxonomy of ML strategies used in network-based IDS (NIDS), analysing recent publications to highlight the strengths and weaknesses of various proposed solutions. In [46] author states that the most cost-effective approach to cybersecurity is prevention, and Network Intrusion Detection Systems (NIDS) play a crucial role in this regard by scrutinizing network flow for potential intrusions. However, despite advancements, Deep Learning-based NIDS still grapple with high false alarm rates and detecting novel attacks. Hence, this paper introduces a novel NIDS framework cantered on generating images from feature vectors and employing Unsupervised Deep Learning. Evaluation across four publicly available datasets showcases an accuracy enhancement of up to 8.25% compared to Deep Learning models applied directly to the original feature vectors. In [47] the integration of edge computing in fifth-generation mobile networks (5G) brings core network components closer to end users, necessitating the detection of malicious signalling traffic to prevent potential attacks. Leveraging a 5G Core Network (5GC) simulator, this study generates a dataset of normalized service interactions from captured network traffic, enabling the identification and classification of normal traffic profiles using machine learning techniques. Results demonstrate the efficacy of supervised learning in classifying normal services based on traffic metadata alone, offering insights for resource allocation and anomaly detection in the dynamic 5G Service Based Architecture (SBA). In [48] the Wireless sensor networks (WSNs) are critical in various applications but are vulnerable to attacks like Denial of Service (DoS), hindering their functionality. To mitigate such threats, a Hybrid Machine

Learning model incorporating algorithms like Support Vector Machine, k-Nearest Neighbour, Naïve Bayes, and Random Forest is proposed. This model aims to pre-emptively identify potential targets for DoS attacks, enhancing the security of WSNs and ensuring uninterrupted data gathering and transmission. The paper [49] introduces an AI-driven intrusion detection method tailored for the ITRI AI BOX information security application. By analysing captured packets, AI algorithms discern potential network attacks or abnormal traffic, facilitating the adjustment or isolation of suspicious data transmissions. The method aims to enhance information security by employing AI models to detect anomalies, with future iterations anticipated to extend to IT and OT fields, ensuring protection against system threats and vulnerabilities. Experimental tests have demonstrated promising accuracy levels, reaching 99%, validating the effectiveness of the approach within the AI BOX environment. In [50] response to the escalating complexity of cyber threats, this study introduces multi-class deep learning models for network intrusion detection, emphasizing the importance of detecting out-of-distribution inputs to address evolving threats effectively. The research compares deterministic and Bayesian neural networks, presenting new uncertainty quantification scoring measures for performance evaluation. Results demonstrate the efficacy of the proposed Bayesian deep learning model in detecting OOD packets, achieving significant detection rates at various significance levels.

2.2 Summary

This chapter provides an extensive review of the existing literature on intrusion detection systems that leverage network traffic behaviour. Over 50 papers spanning various methodologies and approaches have been thoroughly examined and synthesized in this chapter.

Chapter 3

METHODOLOGY

This chapter provides essential information to comprehend this paper. Before explaining our methods in detail, it covers the processes used for access, the tests performed and the learning machines used.

3.1 Features Name

Table 3.1: This table summarizes the names and notations of the features utilized throughout the paper.

S.no	Features	Notation
1	Average Packet Size	F1
2	Average Packet Size Sent	F2
3	Average Packet Size Received	F3
4	Packet sent/Packet received	F4
5	Packets sent per second	F5
6	Packets received per second	F6
7	Bits sent per second	F7
8	Bits received per second	F8
9	Floe duration	F9
10	Time interval between packets sent	F10
11	Time interval between packets received	F11

3.2 About Feature Ranking Method

3.2.1 Cohen’s d Method

Cohen's d is a statistical method used to assess the impact on size between two populations or data sets. It measures the standardized difference between the means of the two groups. This test is especially useful to compare the means of two groups with continuous data, allowing researchers to determine the magnitude of the difference between them.

Cohen's d test calculates the effect size by dividing the difference between the means of two groups by the standard deviation of the mean. A larger value of Cohen's d indicates a large difference between the means of the two groups, while a smaller value indicates a small difference. Similar to other statistical methods, Cohen's d test is based on the principle of hypothesis testing, where one hypothesis is negative and the other hypothesis is formed. The null hypothesis states that there is no significant difference between the two methods, while the other hypothesis states that there is a significant difference. In essence, Cohen's d test allows researchers to determine the practical significance of the difference between two groups, providing valuable insights into the effect size beyond mere statistical significance.

$$Cohen's\ d = \frac{Group_A\ mean - Group_B\ mean}{\text{Pooled Standard deviation (PSD)}}$$

$$\text{PSD} = \sqrt{\frac{(\text{SD}_\text{A})^2 + (\text{SD}_\text{B})^2}{2}}$$

Figure 3.2.1.1: Cohen’s d formulas

Where, SD_A is standard deviation of group A and SD_B is standard deviation of group B.

When applying Cohen's d method, the following assumptions are generally made:

1. Normal Distribution : Data in both groups should follow a normal distribution.

2. Independent Observations: Observations within each group should be independent.

3. Equal Variances: The variances of the two groups should be similar.

4. Measurement Scale: Data should be on an interval or ratio scale.

Meeting these assumptions is essential for the accuracy of Cohen's d calculations.

Working of Cohen’s d :

Cohen's d serves as a principle of hypothesis testing, including null hypotheses and alternative hypotheses.

1. Null Hypothesis: This hypothesis states that there is no difference in means between the two groups being compared. A noticeable difference is due to a difference or a different sample.

2. Alternative Hypothesis: This hypothesis states that there is a difference between group means. This means that the observed difference is due to a true effect rather than a causal effect. It's better than a wrong impression.

3.2.2 Wilcoxon Signed Rank Test

The Wilcoxon signed-rank test is an important nonparametric statistical method used to test for differences between paired samples. Unlike parametric tests, the Wilcoxon test does not assume that the data follow a specific distribution, so it is suitable for analysing data with skewed distributions or data with outliers.

This test evaluates whether the median difference between paired observations is significantly different from the number zero. It does so by ranking the absolute differences between paired observations, disregarding the signs, and summing the ranks of the positive differences. The test statistic is then compared to critical values from a reference distribution to determine statistical significance.

Similar to other statistical tests, the Wilcoxon signed-rank test is based on the formulation of null and alternative hypotheses. The null hypothesis asserts that there is no significant difference between the paired observations, while the alternative hypothesis suggests otherwise.

In summary, Wilcoxon signed-rank test provides a robust method for assessing differences between paired samples, particularly when data may not meet the assumptions of parametric tests. It offers researchers a reliable means of determining the significance of differences in non-normally distributed data.

$$mean = \frac{\mathrm{n(n+1)}}{4}$$

$$standard\ deviation\ (SD) = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

$$z = \frac{|\mathrm{R} - \mathrm{mean}|}{\mathrm{SD}}$$

Figure 3.2.2.1 Wilcoxon formulas

Where, n is the number of samples and R represents the sum of Ranks.

Assumptions of Wilcoxon Signed Rank Test:

1. Paired Observations: The data consists of pairs where each data point in one sample has a corresponding data point in the other sample.

2. Ordinal or Continuous Differences: The differences between paired observations are either ordinal or continuous, suitable for ranking.

3. Symmetric Distribution: The distribution of differences between pairs is symmetric around the median.

4. Independence: Each pair of observations is independent from all other pairs.

The hypotheses for the Wilcoxon Signed Rank Test:

1. Null hypothesis: The mean difference between observations is zero. This means there is no significant difference between the two pairs.
2. Alternative hypothesis: The mean difference between the combined observations is not zero. This shows the main difference between the two pairs.

3.3 Classification Algorithms in Machine Learning

Classification in machine learning can be viewed as a two-phase process: the learning phase and the prediction phase. During the learning phase, a model is constructed using the provided training data. Decision trees are often regarded as one of the simplest and most widely used classification algorithms due to their ease of understanding and interpretation.

Let's illustrate this with a simple example. Detecting spam in email services is a typical classification problem, where an email classifier scans messages to label them as either Spam or Non-Spam. Classification falls under supervised learning, where the input data is accompanied by corresponding target labels. This method has a wide range of applications, including medical diagnosis and email filtering.

For example, in the medical field, classifiers can assist in diagnosing diseases by categorizing patient data based on symptoms and test results. In finance, they help identify fraudulent transactions by distinguishing between normal and suspicious activities. Additionally, in customer service, classification algorithms can sort support tickets into different categories, ensuring they reach the right departments.

Types of Learners in Classification

1) Lazy Learners: These learners defer the majority of their processing until they receive the testing data. Instead of building a model in advance, they store all the training data and perform classification by comparing the new, unseen data to the most relevant examples in the training set. This results in longer prediction times compared to eager learners because the system must analyse the entire dataset to make a classification.

2) Eager Learners: These learners construct a classification model in advance using the training data. They generate a single hypothesis that can be applied universally across the dataset for classification. This means that while the training phase is more time-intensive, the prediction phase is considerably faster than with lazy learners, as the model has already been built.

In machine learning, there are many classification algorithms, each with its own strengths and weaknesses. It's challenging to declare one algorithm as the best overall, as the optimal choice depends on the specific application and characteristics of the data being used.

3.3.1 Support Vector Machine (SVM)

Support Vector Machines (SVMs) are powerful supervised learning models applicable to both classification and regression tasks. The key objective of SVMs is to identify the optimal hyperplane that divides the data into distinct classes most effectively. This is accomplished by maximizing the margin, defined as the distance between the hyperplane and the nearest data points from each class, referred to as support vectors. By maximizing this margin, the model enhances its ability to generalize to new and unseen data.

SVMs transform the input data into a higher-dimensional space, making it easier to achieve linear separation between classes. This transformation is done using kernel functions, which implicitly map the data into a new space. Common kernel functions include the linear kernel, polynomial kernel, radial basis function (RBF) kernel, and sigmoid kernel. The selection of a kernel function depends on the specific characteristics of the data and the problem being solved.

Training an SVM involves solving a convex optimization problem to identify the hyperplane that maximizes the margin. The objective function balances margin maximization and classification error minimization, controlled by a regularization parameter (C). This parameter determines the trade-off between achieving a low error on the training data and maximizing the margin. A high C value aims for fewer misclassifications, while a lower C value allows more misclassifications but seeks a broader margin.

For data that is not linearly separable, SVMs use a soft margin approach, which allows some misclassifications. This method helps in handling noisy data and outliers, making the model more flexible and robust.

Support Vector Machines (SVMs) perform exceptionally well in high-dimensional spaces and are effective even when the number of dimensions exceeds the number of samples. They are memory-efficient because they utilize only the support vectors—a subset of the training points—in the decision function. However, SVMs can be computationally intensive with large datasets, which makes them less suitable for very large-scale problems.

In regression tasks, SVMs use a variation called Support Vector Regression (SVR). The goal in SVR is to find a function that deviates from the actual observed values by a specified margin (ε) while remaining as flat as possible. Like SVM classification, SVR uses kernels to handle non-linear relationships.

Key assumptions and characteristics of SVMs include:

1. Linearity: SVMs aim for a linear separation in a transformed feature space, even if the original space is non-linear.

2. Margin Maximization: The main objective is to determine the hyperplane that maximizes the margin between classes.

3. Kernel Trick: SVMs can manage non-linear data using kernel functions that transform the input space into higher dimensions.

4. Support Vectors: Only a subset of training data points, the support vectors, are crucial in defining the hyperplane.

5. Regularization: The regularization parameter balances the trade-off between maximizing the margin and minimizing classification error.

6. Scalability: SVMs are effective for high-dimensional data but can be computationally demanding for very large datasets.

In summary, SVMs are versatile and powerful tools for classification and regression tasks. Their ability to find an optimal hyperplane and manage non-linear relationships through kernel functions makes them a popular choice in diverse applications, from image recognition to bioinformatics.

3.3.2 Naive Bayes

Naive Bayes (NB) is a group of straightforward yet powerful probabilistic classifiers that use Bayes' theorem with the assumption that features are independent. It calculates the likelihood of each class given the features and selects the class with the highest probability. Despite the unrealistic assumption of feature independence, Naive Bayes often yields good performance, particularly with high-dimensional data. The classifier has variations such as Gaussian Naive Bayes for continuous data, Multinomial Naive Bayes for discrete data like text, and Bernoulli Naive Bayes for binary features. Its simplicity, computational efficiency, and effectiveness in applications like text classification, medical diagnosis, and recommendation systems make it widely used. However, the independence assumption can affect accuracy, and techniques like Laplace smoothing are sometimes necessary to handle unseen feature values.

3.3.3 Logistic Regression

Logistic regression is a method used for binary classification that predicts the likelihood of an input belonging to a specific class. It employs the sigmoid function to convert input features into a probability value between 0 and 1. The model parameters are estimated using maximum likelihood estimation and often optimized via gradient descent. Logistic regression is known for its interpretability, with coefficients indicating the impact of each feature on the predicted outcome. It is commonly applied in areas like medical diagnosis, credit scoring, and marketing due to its efficiency and capability to provide probability scores for class membership. However, it assumes a linear relationship between the features and the log-odds of the outcome, which might not always be suitable.

3.3.4 Decision Tree

Decision trees are powerful models for classification and regression functions that form a tree-like structure where each node represents a decision based on an attribute and each leaf A represents a group of letters or numbers. The decision-making process follows an if-then process as a single classification process. The decision tree is constructed iteratively; each rule is learned sequentially from the training data and the dataset is iteratively partitioned according to the main attribute until a decision is made. The better attitudes are categorical or continuous if they are predetermined. The accuracy of the decision tree depends on several segmentation strategies determined by statistical methods such as data augmentation or Gini impurity. The goal is to create a division that creates the most homogeneous subsets with different goals. Decision trees are versatile and can handle both classification and regression, prioritizing features that have the greatest impact on classification. However, they can lead to overfitting, especially when dealing with noisy data or outliers. Techniques such as tree pruning or limiting tree depth can be used to reduce this. In general, decision trees provide a means to capture complex decision-making processes, making them useful in machine learning.

3.4 Proposed Approach

In the aforementioned section 3.2, we have organized the results of our statistical tests in descending order of feature importance. The feature with the highest rank based on the statistical test appears at the top, followed by features with progressively lower ranks, ending with the least important feature at the bottom of the column. We have created two separate columns: one for the Cohen’s d test, another for the Wilcoxon signed Rank test. Subsequently, we compiled a table showing the accuracy of different machine learning algorithms when applied to each feature individually. The machine learning algorithms used in this analysis, as discussed in section 3.3, includes SVM, Naïve Bayes, Logistic Regression and Decision Tree.

After obtaining the accuracy of individual features for all four algorithms, the next step involved combining features based on their ranks for each statistical test (column). We begin by pairing the feature of highest-ranked with the next highest-ranked feature and calculated the new accuracy of these combined features using This process was repeated for each combination to evaluate the performance improvement.

In this process, we considered two scenarios for combining the features:

1. If the new accuracy obtained from the combined features is less than or equal to the accuracy of the highest-ranked feature alone, we do not combine those features.

2. If the new accuracy from the combined features is greater than the accuracy of the highest-ranked feature alone, we retain the combination.

We repeated this process until we reached the lowest-ranked feature. This resulted in four distinct sets of features, one for each column. Next, we combined these four sets, following the same method and removing any duplicate features. The resulting final set of features represents the most effective ones for detecting network intrusions.

3.5 Summary

In this chapter, our methodology is divided into four distinct sections:

- Section 3.1: This section provides a detailed description of the features utilized in our detection model.
- Section 3.2: Here, we explain the statistical tests employed to rank the features discussed in Section 3.1.
- Section 3.3: We discuss the machine learning classifiers used and how we calculated the accuracy of the features.
- Section 3.4: This section outlines the approach we used in this paper to achieve optimal results.

Chapter 4

RESULTS

In this chapter, we present the results derived from our proposed methodology. Section 4.1 focuses on the feature rankings obtained through the Cohen’s d test and Wilcoxon signed Rank test analysis. Section 4.2 examines the detection performance using individual features. Lastly, Section 4.3 provides a detailed discussion of our approach and its corresponding results.

4.1 Features Ranking

The Cohen’s d tests yield d-values, as shown in Table 4.1.1. When the Wilcoxon signed rank test is applied to data and we obtain the respective z-values for each feature, also presented in Table 4.1.1.

Table 4.1.1: Test Values

Features	Cohen's d value	Wilcoxon Z value
F1	0.623394054	234.9005599
F2	0.75651002	200.1084524
F3	0.620476263	236.4566549
F4	0.298461342	134.2707826
F5	0.739296278	192.0875542
F6	0.746493957	189.7819607
F7	0.056406391	231.414196
F8	0.057689004	233.2640394
F9	0.201068836	247.106314
F10	0.193658033	246.7603721
F11	0.180449414	246.7206267

The d-value is directly proportional to the feature rank, with lower d-values indicating higher ranks. Consequently, features are ordered from top to bottom, with the topmost feature being the highest ranked according to the Cohen’s d test, and the bottom feature being the lowest ranked. For the Wilcoxon signed rank test , features with higher z-values receive higher ranks. Thus, the features are organized accordingly, as shown in Table 4.1.2.

Table 4.1.2: Feature Ranking

Cohen's d	Wilcoxon
F2	F9
F6	F10
F5	F11
F1	F3
F3	F1
F4	F8
F9	F7
F10	F2
F11	F5
F8	F6
F7	F4

4.2 Test results with individual characteristics

Here, we calculate the accuracy of different features using four different algorithms (separate learning machines), SVM, Naïve Bayes, Logistic Regression and Decision Tree, respectively. All these classifiers give different results for highest accuracy. Results for Decision trees and Random Forests

Similarly, F6 is the most accurate, reaching 84.59%; For SVM, F1 is still the most accurate, reaching 97.75%. Now if we compare all the algorithms, we see that F6 is the most accurate among all the algorithms for individual features.

Table 4.2: Single Feature

Features	SVM	NB	LR	DT
F1	82.4	73.52	73.62	**97.75**
F2	80.03	76.95	80.62	96.95
F3	79.52	78.04	80.33	91.71
F4	64.52	73.58	64.96	78.73
F5	84.12	74.49	83.84	83.5
F6	**84.59**	**81.86**	**83.92**	83.31
F7	53.59	46.51	53.59	81.66
F8	53.39	53.53	71.86	73.28
F9	61.45	66.12	46.62	82.04
F10	57.63	57.91	56.23	83.51
F11	63.17	62.52	63.16	83.32

4.3. Detection Results with Proposed Approach

Our objective is to identify a set of features, based on the rankings in Table 4.1, that yield more accurate results than individual features. We applied the method described in Section 3.4 to each test column to achieve this.

Table 4.3.1: Cohen's d test Result

Cohen's d Feature Ranking	SVM	NB	LR	DT
F2	77.32071	76.88098	81.09133	96.95419
F2F6	81.95413	**84.98418**	81.19676	97.69106
F2F6F5	**82.08688**	79.56901	81.43607	98.86745
F2F6F5F1	79.38392	83.50333	79.32971	99.47993
F2F6F5F1F3	81.05687	81.3022	80.29167	99.52494
F2F6F5F1F3F4	79.91331	82.59824	80.59968	99.57233
F2F6F5F1F3F4F9	78.62901	66.30415	80.16372	**99.68038**
F2F6F5F1F3F4F9F10	78.59222	66.1691	80.39118	99.62327
F2F6F5F1F3F4F9F10F11	78.60821	66.39419	**83.82557**	99.61616
F2F6F5F1F3F4F9F10F11F8	79.02725	66.65482	80.11041	99.6209
F2F6F5F1F3F4F9F10F11F8F7	79.03205	66.46883	80.18742	99.67184

Starting with the Cohen's d test results, the highest-ranked feature is F2. We began by using Support Vector Machine(SVM) with feature F2. For Cohen's d, three features (F2, F6, F5) share the first rank with 82.08% accuracy. We combined Features further as per Cohen's d ranking , but there was no improvement in accuracy. Next, we tested feature combination (F2,F6,F5,F1), but it resulted in lower accuracy than the feature combination (F2,F6,F5), so we skipped it and continued in this manner until the last feature. None of the combinations surpassed the accuracy of (F2, F6, F5) , making it the most accurate feature combination in this category.

Similarly, for NB, we started with F2 and combined it with F6, F5, F1, and F3. Again, the combinations did not yield better results, and testing feature (F2,F6) produced better accuracy than F2 alone. We continued this process and found that for NB (F2,F6) was the most accurate, achieving 84.98% accuracy, Similarly in LR highest accuracy of 83.825% is given by the combination (F2,F6,F5,F1,F3,F4,F9,F10,F11).

However, when applying Decision Tree, the combination F2F6F5F1F3F4F9 emerged as the most accurate set, with an accuracy of 99.68%, following the same steps as earlier. Comparing the results across all algorithms, the best set of features is F2F6F5F1F3F4F9, with an accuracy of 99.68%.

Table 4.3.2: Wilcoxon Test Results

Wilcoxon Feature Ranking	SVM	NB	LR	DT
F9	66.0514	66.4534	46.7629	82.0391
F9F10	66.1202	66.2876	**81.2963**	91.802
F9F10F11	66.1234	66.5245	81.1423	92.2013
F9F10F11F3	82.5411	66.2994	47.2995	93.027
F9F10F11F3F1	83.0625	66.0104	46.8257	99.583
F9F10F11F3F1F8	83.0481	66.1691	47.9665	99.5723
F9F10F11F3F1F8F7	83.0337	66.5269	47.8243	99.6493
F9F10F11F3F1F8F7F2	**83.1057**	**66.5944**	48.887	**99.6718**
F9F10F11F3F1F8F7F2F5	75.0432	66.5186	79.9967	99.6612
F9F10F11F3F1F8F7F2F5F6	78.4195	66.5518	79.9422	99.6612
F9F10F11F3F1F8F7F2F5F6F4	79.0321	66.5849	79.8581	99.6612

Going on with Wilcoxon signed rank test results, starting with the highest-ranked feature F9. We began by using SVM with feature F9. For Wilcoxon test, eight features F9F10F11F3F1F8F7F2 share the first rank with 83.1% accuracy. We combined Features further as per Wilcoxon signed rank test ranking , but there was no improvement in accuracy. Next, we tested feature further, but it resulted in lower accuracy than F9F10F11F3F1F8F7F2, so we skipped it and continued in this manner until the last. None of the combinations surpassed the accuracy of F9F10F11F3F1F8F7F2, making it the most accurate feature combination in this category.

Similarly, for NB and DT, we started with F9 and combined it with further as per Wilcoxon signed rank test ranking. Again, the combinations did not yield better results. We continued this process and found that for both NB and Decision Tree, the feature combination F9F10F11F3F1F8F7F2 was the most accurate, achieving 66.59% and 99.67% accuracy, respectively.

However, when applying Logistic Regression, the combination F9F10 emerged as the most accurate set, with an accuracy of 81.29%, following the same steps as earlier. Comparing the results across all algorithms, the best set of features is F9F10F11F3F1F8F7F2, with an accuracy of 99.67%.

Now combining the above results, we get a feature F9F6F3F2F1 as the best feature.

Calculating accuracy of new set, we get following results.

Table 4.3.3: Final set Accuracy (F9F6F3F2F1)

SVM	83.25
NB	81.56
LR	92.66
DT	**99.70**

Ultimately, from the proposed model, we achieved the highest accuracy of 99.7% by combining five features: F9, F6, F3, F2, and F1. This accuracy surpasses that of any individual feature or any other set of features obtained from the three tests.

4.4 Summary

To identify the optimal set of features for detecting intrusions, we employed various tests and machine learning algorithms. By applying the method outlined in Section 3.4, we discovered that the results in Chapter 4 revealed a set of five features that achieved the highest accuracy of 99.70%.

Chapter 5
Conclusion

In this thesis, we emphasize the critical need for network security in the face of increasing network attacks. Such attacks can lead to significant financial and operational losses for companies, organizations, and individuals. Traditional defences like antivirus software and firewalls, which once provided sufficient security, are no longer adequate against these evolving threats. This situation necessitates intelligent countermeasures to maintain the security of networks and critical systems.

Chapter 1 delves into the concept and importance of Intrusion Detection Systems (IDS). An IDS can be defined as a device or software application designed to detect unauthorized access. We explored various types of IDS, including signature-based, anomaly-based, and hybrid systems. The primary goal of an IDS is to identify intrusions before they can cause substantial damage to the network.

Chapter 2 reviews existing literature on network traffic-based intrusion detection, summarizing over 50 research papers. In Chapter 3, we introduced the 11 features used in our study, detailing their names and functions. We then employed statistical tests to rank these features based on their effectiveness in detecting network intrusions, the core objective of our research. We used two statistical tests: the Cohen’s d test and the Wilcoxon Signed Rank test. After explaining the purpose, formulas, and functioning of these tests, we introduced the machine learning classifiers used in our study: SVM, Naive Bayes, Logistic Regression, Decision Tree.

Chapter 3 primarily focused on our methodology for ranking the features. The features were ordered so that the least efficient feature was at the bottom, and the most efficient at the top. This created two columns: one for Cohen’s d test, one for Wilcoxon signed rank test. We then created another table where each feature was individually evaluated using all four machine learning classifiers. To explore the possibility that feature combinations might yield better results, we tested various feature combinations with all four classifiers, resulting in four different sets of columns.

When combining features, we observed two outcomes: either the new combination improved accuracy, or it did not. Combinations that resulted in higher accuracy were ranked higher, while those that did not were discarded. This process was repeated for the four columns, ultimately yielding a set of features that were most effective for network intrusion detection.

Chapter 4 presents the tables, calculations, and evidence supporting our findings. Our research concluded that the combination of five features—"Floe duration", "Packets received per second," "Average packet size received," "Average Packet size sent," and "Average packet size"—achieved the highest accuracy of 99.70% when using the Decision Tree classifier, outperforming all other features and combinations.

References

[1] S. A. Ali, B. J, H. N and A. S, "Web Intrusion Detection with Machine Learning Algorithm," *2023 Global Conference on Information Technologies and Communications (GCITC)*, Bangalore, India, 2023, pp. 1-5

[2] H.Li , “Research on network intrusion detection technology based on improved FCMA algorithm,” Journal of Ambient Intelligence and Humanized Computing, 2021.

[3] Q. Feng, Z. Lin and L. Bing, "IP-MCCLSTM: A Network Intrusion Detection Model Based on IP Filtering," *2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, Chengdu, China, 2023, pp. 1-6

[4] R. B. Said and I. Askerzade, "Attention-Based CNN-BiLSTM Deep Learning Approach for Network Intrusion Detection System in Software Defined Networks," *2023 5th International Conference on Problems of Cybernetics and Informatics (PCI)*, Baku, Azerbaijan, 2023, pp. 1-5

[5] Y. Huang, "Network Intrusion Detection Method Based on Naive Bayes Algorithm," *2022 6th Asian Conference on Artificial Intelligence Technology (ACAIT)*, Changzhou, China, 2022, pp. 1-10

[6] N. Wei *et al.*, "An Autoencoder-Based Hybrid Detection Model for Intrusion Detection with Small-Sample Problem," in *IEEE Transactions on Network and Service Management*, vol. 21, no. 2, pp. 2402-2412, April 2024

[7] M. A. Siddiqi and W. Pak, "Tier-Based Optimization for Synthesized Network Intrusion Detection System," in *IEEE Access*, vol. 10, pp. 108530-108544, 2022

[8] H. Yu, C. Kang, Y. Xiao and Y. Yang, "Network Intrusion Detection Method Based on Hybrid Improved Residual Network Blocks and Bidirectional Gated Recurrent Units," in *IEEE Access*, vol. 11, pp. 68961-68971, 2023

[9] C. Lu, Z. Tao and M. Yuanyuan, "A 5G Slice Traffic Anomaly Detection Method Based on Convolution Neural Network," *2023 8th International Conference on Signal and Image Processing (ICSIP)*, Wuxi, China, 2023, pp. 963-967

[10] B. Wang, Y. Su, M. Zhang and J. Nie, "A Deep Hierarchical Network for Packet-Level Malicious Traffic Detection," in *IEEE Access*, vol. 8, pp. 201728-201740, 2020

[11] C. Xu, J. Shen and X. Du, "A Method of Few-Shot Network Intrusion Detection Based on Meta-Learning Framework," in *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3540-3552, 2020

[12] N. Wang, Y. Chen, Y. Xiao, Y. Hu, W. Lou and Y. T. Hou, "MANDA: On Adversarial Example Detection for Network Intrusion Detection System," in *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 2, pp. 1139-1153, 1 March-April 2023

[13] S. Bebortta, T. Panda and S. K. Singh, "An Intelligent Hybrid Intrusion Detection System for Internet of Things-based Applications," *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, Bengaluru, India, 2023, pp. 01-06

[14] N. -T. Nguyen, T. -N. Le, K. -H. Le-Minh and K. -H. Le, "Towards Generating Semi-Synthetic Datasets for Network Intrusion Detection System," *2023 International Conference on Information Networking (ICOIN)*, Bangkok, Thailand, 2023, pp. 62-66

[15] Y. A. Rani, K. Deepthi Reddy and R. U. Rani, "A Novel Network Intrusion Detection Model using Residual Recurrent Neural Network with Improved Garter Snake-based Optimization Strategy," *2023 Global Conference on Information Technologies and Communications (GCITC)*, Bangalore, India, 2023, pp. 1-8

[16] M. Wang, K. Shi, A. Jiang and Z. Zhang, "Constructing Weighted Host Communication Networks for Intrusion Detection based on Complex Network Theory," *2023 4th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, Nanjing, China, 2023, pp. 424-428

[17] A. K. Sah and V. K, "Anomaly-Based Intrusion Detection in Network Traffic using Machine Learning: A Comparative Study of Decision Trees and Random Forests," *2024 2nd International Conference on Networking and Communications (ICNWC)*, Chennai, India, 2024, pp. 1-7

[18] C. Park, J. Lee, Y. Kim, J. -G. Park, H. Kim and D. Hong, "An Enhanced AI-Based Network Intrusion Detection System Using Generative Adversarial Networks," in *IEEE Internet of Things Journal*, vol. 10, no. 3, pp. 2330-2345, 1 Feb.1, 2023

[19] K. Yang, Y. Shi, Z. Yu, Q. Yang, A. K. Sangaiah and H. Zeng, "Stacked One-Class Broad Learning System for Intrusion Detection in Industry 4.0," in *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 251-260, Jan. 2023

[20] D. Uhříček, K. Hynek, T. Čejka and D. Kolář, "BOTA: Explainable IoT Malware Detection in Large Networks," in *IEEE Internet of Things Journal*, vol. 10, no. 10, pp. 8416-8431, 15 May15, 2023

[21] E. Paolini, L. Valcarenghi, L. Maggiani and N. Andriolli, "Real-Time Network Packet Classification Exploiting Computer Vision Architectures," in *IEEE Open Journal of the Communications Society*, vol. 5, pp. 1155-1166, 2024

[22] Y. Gao, H. Miao, J. Chen, B. Song, X. Hu and W. Wang, "Explosive Cyber Security Threats During COVID-19 Pandemic and a Novel Tree-Based Broad Learning System to Overcome," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 1, pp. 786-795, Jan. 2024

[23] Z. Tauscher, Y. Jiang, K. Zhang, J. Wang and H. Song, "Learning to Detect: A Data-driven Approach for Network Intrusion Detection," *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, Austin, TX, USA, 2021, pp. 1-6

[24] N. Bhattacharya, A. Subudhi, S. Mishra, V. Sharma, A. P. Aderemi and C. Iwendi, "A Novel Ensemble based Model for Intrusion Detection System," *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Greater Noida, India, 2024, pp. 620-624

[25] N. Kausar, Z. Latif, C. Lee and U. Iqbal, "Towards Detection and Mitigation of Traffic Anomalies in SDN," *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, Republic of, 2021, pp. 728-731

[26] H. Zhu and J. Lu, "Graph-based Intrusion Detection System Using General Behavior Learning," *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, Rio de Janeiro, Brazil, 2022, pp. 2621-2626

[27] S. Wang, Z. Chen, J. Chen and P. Zhu, "Design and Implementation of Intrusion Detection System Based on Deep Learning," *2023 IEEE 3rd International Conference on Electronic*

Technology, Communication and Information (ICETCI), Changchun, China, 2023, pp. 1495-1497

[28] M. He, Y. Huang, X. Wang, P. Wei and X. Wang, "A Lightweight and Efficient IoT Intrusion Detection Method Based on Feature Grouping," in *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 2935-2949, 15 Jan.15, 2024

[29] M. Fuentes-García, J. Camacho and G. Maciá-Fernández, "Present and Future of Network Security Monitoring," in *IEEE Access*, vol. 9, pp. 112744-112760, 2021

[30] J. Saleem, R. Islam and M. Z. Islam, "Darknet Traffic Analysis: A Systematic Literature Review," in *IEEE Access*, vol. 12, pp. 42423-42452, 2024,

[31] Y. Djenouri, D. Djenouri, A. Belhadi, G. Srivastava and J. C. -W. Lin, "Emergent Deep Learning for Anomaly Detection in Internet of Everything," in *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3206-3214, 15 Feb.15, 2023

[32] H. M. Song and H. K. Kim, "Self-Supervised Anomaly Detection for In-Vehicle Network Using Noised Pseudo Normal Data," in *IEEE Transactions on Vehicular Technology*, vol. 70, no. 2, pp. 1098-1108, Feb. 2021

[33] L. Yuan *et al*., "Manticore: An Unsupervised Intrusion Detection System Based on Contrastive Learning in 5G Networks," *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, 2024, pp. 4705-4709

[34] N. Alkhatib, M. Mushtaq, H. Ghauch and J. -L. Danger, "Unsupervised Network Intrusion Detection System for AVTP in Automotive Ethernet Networks," *2022 IEEE Intelligent Vehicles Symposium (IV)*, Aachen, Germany, 2022, pp. 1731-1738

[35] H. V. Vo, H. N. Nguyen, T. N. Nguyen and H. P. Du, "SDAID: Towards a Hybrid Signature and Deep Analysis-based Intrusion Detection Method," *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, Rio de Janeiro, Brazil, 2022, pp. 2615-2620

[36] Z. Zeng, W. Peng and D. Zeng, "Improving the Stability of Intrusion Detection With Causal Deep Learning," in *IEEE Transactions on Network and Service Management*, vol. 19, no. 4, pp. 4750-4763, Dec. 2022

[37] T. M. Booij, I. Chiscop, E. Meeuwissen, N. Moustafa and F. T. H. d. Hartog, "ToN_IoT: The Role of Heterogeneity and the Need for Standardization of Features and Attack Types in IoT Network Intrusion Data Sets," in *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 485-496, 1 Jan.1, 2022

[38] G. Elbez, K. Nahrstedt and V. Hagenmeyer, "Early Attack Detection for Securing GOOSE Network Traffic," in *IEEE Transactions on Smart Grid*, vol. 15, no. 1, pp. 899-910, Jan. 2024

[39] A. B. Z. Mahmoodi, S. Sheikhi, E. Peltonen and P. Kostakos, "Autonomous Federated Learning for Distributed Intrusion Detection Systems in Public Networks," in *IEEE Access*, vol. 11, pp. 121325-121339, 2023

[40] T. C. M. Dönmez, "Anomaly Detection in Vehicular CAN Bus Using Message Identifier Sequences," in *IEEE Access*, vol. 9, pp. 136243-136252, 2021

[41] S. Kim, S. Yoon and H. Lim, "Deep Reinforcement Learning-Based Traffic Sampling for Multiple Traffic Analyzers on Software-Defined Networks," in *IEEE Access*, vol. 9, pp. 47815-47827, 2021

[42] D. Ke, "Network Intrusion Detection Based on Feature Selection and Transformer," *2023 International Conference on Intelligent Communication and Computer Engineering (ICICCE)*, Changsha, China, 2023, pp. 23-28

[43] A. Kumar and I. Sharma, "Intrusion Type Classifier: A Machine Learning Based Approach for Automated Detection and Classification of Network Intrusions," *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, Bangalore, India, 2023, pp. 1-6

[44] A. Wath, T. Kurian and J. P. Martin, "Security Landscape of Anomaly based Defence Mechanisms in Edge Environments," *2023 9th International Conference on Smart Computing and Communications (ICSCC)*, Kochi, Kerala, India, 2023, pp. 395-400

[45] A. Dandaras, J. Borges, B. Igor and N. Doukas, "Machine Learning Applications for Network Intrusion Detection Systems," *2023 13th International Conference on Dependable Systems, Services and Technologies (DESSERT)*, Athens, Greece, 2023, pp. 1-7

[46] R. Hosler, A. Sundar, X. Zou, F. Li and T. Gao, "Unsupervised Deep Learning for an Image Based Network Intrusion Detection System," *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, Kuala Lumpur, Malaysia, 2023, pp. 6825-6831

[47] R. Pell, M. Shojafar, D. Kosmanos and S. Moschoyiannis, "Service Classification of Network Traffic in 5G Core Networks using Machine Learning," *2023 IEEE International Conference on Edge Computing and Communications (EDGE)*, Chicago, IL, USA, 2023, pp. 309-318

[48] G. S. Rao, M. Harshitha, V. R. Joshitha, S. S. Sravya and M. V. Priya, "DoS Attack Detection in Wireless Sensor Networks (WSN) Using Hybrid Machine Learning Model," *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India, 2023, pp. 384-388

[49] J. -L. Chen *et al.*, "AI-Based Intrusion Detection System for Secure AI BOX Applications," *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Bali, Indonesia, 2023, pp. 360-364

[50] J. A. Wong, A. M. Berenbeim, D. A. Bierbrauer and N. D. Bastian, "Uncertainty-Quantified, Robust Deep Learning for Network Intrusion Detection," *2023 Winter Simulation Conference (WSC)*, San Antonio, TX, USA, 2023, pp. 2470-2481