Improving Performance of Intrusion Detection Systems using Machine Learning Techniques

By

Amna Zahid

Fall-2020-MSIT-00000329206

Supervisor

Dr. Syed Imran Ali

Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Information Technology (MS IT)

In

School of Electrical Engineering & Computer Science (SEECS) ,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(August 2024)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Improving Performance of Intrusion Detection Systems using Machine Learning Techniques" written by Amna Zahid, (Registration No 00000329206), of SEECS has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: S. I. Ali

Name of Advisor: Dr. Syed Imran Ali

Date: 09-Jul-2024

HoD/Associate Dean:

Date: 09-Jul-2024

Signature (Dean/Principal):

Date: 09-Jul-2024

Approval

It is certified that the contents and form of the thesis entitled "Improving Performance of Intrusion Detection Systems using Machine Learning Techniques" submitted by Amna Zahid have been found satisfactory for the requirement of the degree

Advisor : Dr. Syed Imran Ali

Signature: S.I. Ali

Date: 09-Jul-2024

Committee Member 1:Mr Bilal Ali

Signature:

04-Jul-2024

Committee Member 2:Mr Fahad Ahmed Satti

Signature: ________________________

Date: 05-Jul-2024

Signature: ________________________

Date: ___________________________

FORM TH-4

National University of Sciences & Technology

MASTER THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: (Student Name & Reg. #) Amna Zahid [00000329206]

Titled: Improving Performance of Intrusion Detection Systems using Machine Learning Techniques

be accepted in partial fulfillment of the requirements for the award of Master of Science (Information Technology) degree.

Examination Committee Members

1. Name: Bilal Ali Signature: 19-Aug-2024 3:48 PM

2. Name: Fahad Ahmed Satti Signature: 19-Aug-2024 3:48 PM

Supervisor's name: Syed Imran Ali Signature: S. I. Ali 19-Aug-2024 6:02 PM

Arham Muslim
HoD / Associate Dean

21-August-2024
Date

COUNTERSINGED

22-August-2024
Date

Muhammad Ajmal Khan
Principal

Certificate of Originality

I hereby declare that this submission titled "Improving Performance of Intrusion Detection Systems using Machine Learning Techniques" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name:Amna Zahid

Student Signature: Amna Zahid.

Certificate for Plagiarism

It is certified that PhD/M.Phil/MS Thesis Titled "Improving Performance of Intrusion Detection Systems using Machine Learning Techniques" by Amna Zahid has been examined by us. We undertake the follows:

a. Thesis has significant new work/knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.

b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.

c. There is no fabrication of data or results which have been compiled/analyzed.

d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.

e. The thesis has been checked using TURNITIN (copy of originality report attached) and found within limits as per HEC plagiarism Policy and instructions issued from time to time.

Name & Signature of Supervisor

Dr. Syed Imran Ali

Signature : S. I. Ali

Dedication

To my beloved grandparents, My Nana Abu **Mukhtar Ahmad** and My Nani Ami **Bashira Bibi**.

Your love, wisdom, and unwavering support have been a constant source of strength and inspiration throughout my journey. I'm forever grateful for the values you have instilled in me and the foundation you have provided.

Acknowledgment

I am deeply grateful to Allah for His countless blessings and guidance throughout this journey. His divine intervention illuminated my path and opened doors of opportunity, making this achievement possible.

I extend my heartfelt gratitude to my supervisor, **Dr. Sayed Imran Ali**, for his unwavering support and invaluable guidance.

I am forever thankful to my parents, **Zahid Mehmood** and **Khushnood Akhtar**, for being my pillars of strength, and to my siblings, **Fatima Zahid** and **Muhammad Umar Farooq**, for their endless encouragement.

A special thanks to my maternal uncles, **Zafar Mehmood** and **Tahir Mehmood**, for their wisdom and support, your words of wisdom and constant encouragement have been a great source of strength. You have been like second fathers to me, and I am truly grateful for your presence in my life.

To my dear friends, **Samra Zafar** and **Novera Pervaiz**, for their unwavering friendship and support. You have been my sounding boards, my cheerleaders, and my sources of laughter during the most challenging times.

This thesis is a testament to the love and support of all these incredible people in my life. I dedicate this work to each of you.

Thank you from the bottom of my heart.

Amna Zahid

List of Figures

List of Tables

Table of Contents

Abstract

The rapid upsurge in network intrusions has driven research into AI techniques for intrusion detection systems (IDS). A major challenge is ensuring AI models are understandable to security analysts, leading to the adoption of explainable AI (XAI) methods. This study presents a framework to evaluate black-box XAI methods for IDS, focusing on global and local interpretability, tested on three well-known intrusion datasets and AI methods. This research enhances IDS using XAI techniques, specifically LIME and SHAP, applied to datasets UNSW-NB15, NSL-KDD, CICIDS2019, and a merged dataset. Preprocessing steps like normalization and feature alignment were used to standardize the data. The findings show that integrating XAI improves IDS interpretability and trustworthiness, aiding analysts in understanding system decisions. This research advances more interpretable and resilient IDS, capable of countering evolving cyber threats, and provides a foundational XAI evaluation tool for the network security community.

Chapter 1

Introduction

1.1 Overview

Due to the constant growth of cybersecurity threats, it is necessary to continuously improve Intrusion Detection Systems (IDS) in order to accurately detect and reduce potential dangers [1, 2]. Machine Learning (ML) is a strong tool that can greatly enhance the performance of Intrusion Detection Systems (IDS). By integrating ML techniques, IDS can transcend traditional static rule-based approaches, becoming more adaptive, accurate, and capable of discerning subtle patterns indicative of intrusions [3].

1.1.1 Types of Intrusion Detection

In order to detect malicious activity, signature-based intrusion detection systems compare incoming network traffic with predetermined patterns or signatures of known assaults. There is a low false-positive rate and it is effective against known threats, but it is susceptible to zero-day attacks and has difficulties keeping its signature database updated [4].

Anomaly-Based Intrusion Detection: Anomaly-based Intrusion Detection Systems (IDS) create a reference point for typical behaviour and identify any divergence as a possible intrusion. It focuses on identifying unusual patterns that may indicate an attack. While effective in detecting novel threats, it may face challenges in defining a clear baseline and adaptability to changing network dynamics [5].

Behavior-Based Intrusion Detection: Behavior-based IDS monitors user and system behavior, identifying deviations from expected patterns. Integrating anomaly detection, it recognizes both known and unknown attack patterns, providing a comprehensive approach. Challenges include defining normal behavior and potential for false positives [6].

1.1.2 Improving IDS Performance Using ML

Ensemble Learning: Combining multiple ML models into an ensemble can enhance overall accuracy and robustness. Ensemble learning mitigates the weaknesses of individual models, leading to a more effective IDS with improved detection rates and resilience against diverse attack strategies [7].

Deep Learning: By utilising deep learning techniques like neural networks, IDS is able to autonomously acquire complex features from data. This enhances the system's ability to identify complex patterns associated with cyber threats, providing increased accuracy, adaptability to evolving threats, and the capability to handle large and complex datasets [2, 8].

Transfer Learning: Through the process of transfer learning, IDS is able to enhance its performance in one domain by applying what it has learned in another. This is particularly useful in enhancing detection capabilities for zero-day attacks, leading to improved detection of novel threats and reduced dependence on extensive labeled datasets.

Explainable AI Techniques: Integrating explainable AI methodologies, such as LIME and SHAP, enhances the interpretability of IDS decisions. This contributes to increased transparency, aiding in the analysis and validation of detected intrusions. Cybersecurity analysts can better understand and trust the system's alerts.

Dynamic Learning Parameters: Implementing dynamic learning parameters allows the IDS to adapt in real-time to changing network conditions and emerging threats. This ensures agility and responsiveness, leading to improved real-time adaptability and increased efficiency in detecting evolving attack patterns.

The integration of ML techniques holds significant promise in advancing the capabilities of IDS, improving detection accuracy, reducing false positives, and enhancing overall performance in the face of evolving cyber threats.

This Table 3.1 presents a comparative analysis of machine learning methods used to enhance Intrusion Detection Systems (IDS). The analysis focuses on important factors including detection accuracy, resilience, effectiveness against new threats, computational requirements, interpretability, real-time adaptability, ability to handle large datasets, and reliance on labelled data.

Table 1.1: The improvements in Intrusion Detection Systems (IDS) through the utilisation of Machine Learning methodologies

Improvement Aspect	**Ensemble Learning**	**Deep Learning**	**Transfer Learning**	**Explainable AI Techniques**	**Dynamic Learning Parameters**
Detection Accuracy	High	Very High	High	Moderate to High	High
Robustness	Resilient against diverse attacks	Enhanced resilience	Improved adaptability	Improved interpretability	Responsive to changing dynamics
Capability Against Threats	Effective	Effective	Effective	Moderate	Effective
Resource Intensity	Moderate	High	Moderate to High	Moderate	Low to Moderate
Interpretability	Moderate	Low	Moderate	High	High
Real-Time Adaptability	Yes	Yes	Yes	No	Yes
Handling Large Datasets	Yes	Yes	Yes	Yes	Yes
Dependence on Labeled Data	Moderate	High	Moderate	Moderate	Low to Moderate

1.2 Background Studies

Continual developments in defence mechanisms, particularly in Intrusion Detection Systems (IDS), are necessary in response to the ever-evolving cyber threats in the field of cybersecurity. The origins of intrusion detection may be traced back to the early stages of computer networks, during which the main objective was to safeguard systems from external intrusions. Over the years, the field has witnessed significant developments, and contemporary challenges demand innovative approaches, leading to the integration of Machine Learning (ML) techniques [9].

When the area of computer security was just starting to take off in the 1970s, researchers recognised the need for systems that could detect and react to intrusions and other forms of harmful activity. Intruder detection systems used to depend on hand-crafted rule-based signatures to identify particular attack patterns. These systems, known as signature-based IDS, were effective to some extent, but they struggled to adapt to the evolving tactics of cyber adversaries [8, 10].

The paradigm shifted with the introduction of anomaly-based intrusion detection in the late 1980s. The objective of this technique was to establish a standard level of normal behaviour and detect any deviations that could indicate potential intrusions. While offering improved adaptability to new threats, anomaly-based systems faced challenges in distinguishing between malicious activities and legitimate variations in network behavior [**11-13**].

Challenges in Traditional IDS: Despite the historical advancements, traditional IDS faced limitations in handling the intricacies of modern cyber threats. Signature-based systems were vulnerable to zero-day attacks, as they relied on known patterns. Anomaly-based systems, while effective in identifying novel threats, struggled with high false positive rates and challenges in defining a clear baseline for normal behavior. The increasing volume and complexity of network data further exacerbated these challenges. Traditional IDS often struggled to keep pace with the sheer volume of information, leading to delays in detecting and responding to intrusions. Additionally, the static nature of rule-based systems hindered their ability to adapt to rapidly changing attack tactics [**14-17**].

Machine Learning in Intrusion Detection: The integration of Machine Learning (ML) techniques into intrusion detection marked a pivotal shift in the field. ML offered the promise of adaptive and intelligent systems capable of learning from data patterns, thereby overcoming the limitations of traditional rule-based approaches. One of the early applications of ML in intrusion detection involved the use of decision trees and clustering algorithms. These models aimed to learn decision boundaries from labeled data, distinguishing between normal and malicious network behavior. However, their success was limited by the need for extensive labeled datasets and challenges in handling dynamic and evolving attack strategies [3, 11, 12].

1.2.1 Types of Machine Learning in Intrusion Detection

Supervised Learning: An example of supervised learning is the process of training a model with labelled data, where each instance is given a class (normal, invasive, etc.). Intrusion detection has been done using algorithms like Random Forests and Support Vector Machines (SVM). Although these models are useful, they are very dependent on properly labelled data being readily available.

Unsupervised Learning: Without labels to guide the process, unsupervised learning attempts to discover patterns in the data. Anomaly detection has made use of clustering algorithms such as hierarchical clustering and k-means. Although it might have trouble with large false positive rates, unsupervised learning works wonders when it comes to discovering new dangers.

Semi-Supervised Learning: To create semi-supervised learning, supervised and unsupervised methods are combined. It makes use of a smaller dataset with labels and a larger dataset without labels. By combining the best features of the two paradigms, this method hopes to overcome the difficulties caused by the lack of labelled data.

Deep Learning: Deep Learning, particularly neural networks, has become increasingly prominent in recent years because of its capacity to autonomously acquire complex properties from data. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have proven to be effective at identifying intricate patterns in network traffic, resulting in enhanced accuracy in detecting intrusions.

1.2.2 Challenges and Opportunities in Machine Learning-Based IDS

There are still certain problems with machine learning, even if it has greatly improved intrusion detection. In particular, supervised learning models continue to face the formidable challenge of insufficient labelled datasets. Understanding the reasoning behind detection decisions is also made more difficult by the interpretability of complicated ML models, like deep neural networks. A number of interpretable AI approaches have evolved to tackle this problem and shed light on decision-making, such as SHapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). In order for cybersecurity analysts to understand and rely on the alarms produced by intrusion detection systems (IDS), these methods attempt to make ML models more transparent. Another exciting new development is transfer learning, which enables models to improve their performance in one domain by using what they've learned in another. The capacity to adjust to new and developing dangers is of the utmost importance in intrusion detection, where this is especially pertinent.

As technology continues to advance, the future of intrusion detection lies in the continuous refinement of ML models, the development of more interpretable AI methodologies, and the exploration of innovative approaches, such as reinforcement learning and meta-learning. Furthermore, the collaboration between academia and industry is crucial for creating comprehensive datasets, validating models in real-world scenarios, and ensuring the practical applicability of machine learning-based IDS. The history of intrusion detection reflects a journey from early rule-based systems to the current era of machine learning integration. The challenges posed by evolving cyber threats necessitate intelligent and adaptive solutions, and machine learning offers a transformative approach. The diverse landscape of ML techniques, from supervised and unsupervised learning to deep learning and transfer learning, presents a rich tapestry of opportunities and challenges. As research continues to push the boundaries, the future holds the promise of more robust, adaptive, and interpretable Intrusion Detection Systems capable of safeguarding digital environments against the ever-evolving threat landscape.

1.3 Research Motivations

The motivations underlying this research stem from the critical need to enhance cybersecurity measures amidst rapid technological advancements and an increasingly complex cyber threat landscape. Traditional Intrusion Detection Systems (IDS) have encountered significant limitations in adapting to the sophisticated nature of contemporary cyber threats, necessitating a shift towards more innovative approaches. One primary motivation is the increasing complexity and diversity of cyber threats. Malicious actors continuously refine their techniques, making conventional signature-based and rule-driven IDS less effective. This research aims to develop dynamic intrusion detection mechanisms that can learn and adapt to new attack vectors, thereby improving the resilience of cybersecurity defenses. Another key motivation is the need for transparency and interpretability in IDS decision-making. As cyber threats become more sophisticated, it is essential for cybersecurity analysts to understand and trust the decisions made by IDS. This research integrates Explainable AI (XAI) methodologies to clarify decision processes and empower analysts with insights into the rationale behind intrusion alerts. Additionally, the rise of zero-day attacks presents a significant challenge. This research leverages transfer learning techniques to utilize knowledge from known threats to detect novel and unseen attacks, contributing to more proactive defense mechanisms. Finally, the motivation extends to optimizing IDS for real-time adaptability. In a rapidly evolving cyber threat environment, IDS must dynamically adjust its learning parameters to remain effective. This research aims to develop adaptive learning mechanisms to ensure IDS can respond to changing network dynamics, ultimately creating a more resilient and responsive cybersecurity defense framework.

1.4 Problem Statement

The creation of strong security measures is crucial in today's digital ecosystem due to the increasing frequency and sophistication of cyber threats. When it comes to protecting networked systems, intrusion detection systems (IDS) are crucial. They help discover and stop harmful actions. The ever-changing nature of cyber threats, however, is putting standard IDS to the test. Existing systems often struggle to keep pace with novel attack vectors, resulting in a heightened risk of successful intrusions. The crux of the problem lies in the limited adaptability and accuracy of current IDS, which rely on static rule sets or signature-based approaches that are inherently reactive. As attackers employ sophisticated techniques to obfuscate their activities, IDS must evolve to exhibit a proactive and dynamic response. Moreover, the sheer volume and complexity of network data necessitate an intelligent approach to discerning normal behavior from anomalies. This research aims to address these challenges by integrating Explainable AI (XAI) methodologies within the IDS framework, enhancing transparency and decision-making processes, and employing advanced machine learning techniques to improve detection capabilities against emerging threats.

This research aims to address these challenges by investigating and implementing advanced Machine Learning (ML) techniques within the IDS framework. The lack of robustness in current IDS is a critical concern, demanding a paradigm shift towards adaptive ensemble-based models that can learn and adapt to emerging threats. The goal of this research is to shed light on the decision-making processes of IDS and increase transparency by incorporating Explainable AI (XAI) approaches. This involves deciphering the reasoning behind intrusion alerts using methods like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). To further aid in the creation of proactive defence measures, this study investigates the use of transfer learning to detect new and unforeseen attacks by leveraging knowledge from existing threats. The end goal is to strengthen cybersecurity safeguards so that networked systems can withstand and adapt to new threats as they emerge.

Explainable AI (XAI) is a developing area of study that focuses on enhancing the clarity and comprehensibility of machine learning models. The research aims to enhance cybersecurity analysts' understanding of the decision-making process of the system by integrating XAI into IDS, thereby providing them with practical insights. This not only improves trust and dependability, but also enables more effective comprehension and control of intrusion alerts. Methods such as LIME and SHAP are utilised to elucidate individual model predictions, so providing a lucid and comprehensible understanding of the decision-making process. This technique guarantees that Intrusion Detection Systems (IDS) are not only efficient in identifying potential dangers, but also clear and comprehensible, resulting in more knowledgeable and prompt reactions to cyber threats.

1.5 Problem Formulations

1.5.1 Dynamic Threat Detection in Network Traffic

The existing Intrusion Detection Systems (IDS) are hindered by their limited adaptability and dynamic response capabilities, particularly in the face of emerging cyber threats. The challenge is to formulate a model that can effectively detect and classify intrusions in real-time, adapting its decision boundaries to evolving attack patterns.

Mathematically, Let D represent the dataset comprising network traffic features over time, and L denote the corresponding labels indicating the presence or absence of an intrusion. The problem can be mathematically expressed as a dynamic classification task:

$Find\ f : D \times T \to L\ \ \ldots\ \ (1)$

where T denotes the time dimension, capturing the evolving nature of network traffic. The model f is expected to adapt its decision boundaries θ over time:

$L = f\left(D, \Theta_t\right), for\ t = 1, 2, ..., T\ \ldots\ \ (2)$

The objective is to minimize the misclassification rate over time:

$min_{\Theta t}\frac{1}{T}\ I\sum_{t=1}^{T}\left(L_t = f\left(D_t, \Theta_t\right)\right)\ \ldots\ \ (3)$

where I is the indicator function.

This problem addresses the need for a dynamic IDS that can adapt its decision-making process over time. The model should be capable of learning from the changing characteristics of network traffic to provide accurate and timely intrusion detection.

1.5.2 Ensemble-Based Anomaly Detection

Enhancing the robustness of Intrusion Detection Systems (IDS) by leveraging ensemble learning techniques to fuse multiple base models and improve detection accuracy.

Mathematically, Let M be the set of base models, x represents input features, and y denote the binary output indicating intrusion (y=1) or normal behavior (y=0). The ensemble model E can be defined as a weighted combination of base models:

$E(x) = \sum_{i=1}^{n} w_i M_i(x) \ldots$ (4)

where w_i are the weights assigned to each base model.

The objective is to optimize the weights w_i to minimize the ensemble's error rate:

$min_{w1,\ldots,N} \frac{1}{N} (y \neq E(x)) \ldots$ (5)

This problem focuses on constructing an ensemble model that exploits the diversity among base models to enhance overall detection accuracy, providing a more robust defense against intrusion attempts.

1.5.3 Transfer Learning for Zero-Day Attacks

Mitigating the impact of zero-day attacks on Intrusion Detection Systems (IDS) by formulating a transfer learning framework that leverages knowledge gained from known threats to improve the detection of novel and unseen attacks.

Mathematically, Let Ds and D_t be the source and target domains, respectively. The goal is to learn a mapping f that transfers knowledge from the source to the target domain:

$f : D_s \to D_t \ldots$ (6)

The objective is to minimize the divergence between the source and target domains while optimizing the model f:

$min_f \ Divergence(Ds, Dt) + Loss(f) \ldots$ (7)

This formulation uses transfer learning to increase detection performance in a new, perhaps undiscovered area, taking on the challenge of zero-day attacks. The domain in question is well-established. The goal of the model is to find a middle ground between optimising for the target domain and minimising the model's loss throughout the adaptation process.

1.6 Research Objectives

This project aims to transform Intrusion Detection Systems (IDS) in the field of cybersecurity by implementing cutting-edge machine learning techniques. It specifically focuses on overcoming the constraints of traditional IDS in order to enhance their adaptability and durability. The purpose of this research is to accomplish the following goals:

- To Develop Ensemble-Based IDS Framework: Create a collaborative ensemble model, uniting diverse machine learning algorithms to enhance intrusion detection accuracy and robustness.

- To Investigate Explainable AI for Intrusion Alerts: Explore LIME and SHAP techniques for transparent insights into IDS decision-making, empowering analysts to comprehend and act upon intrusion alerts effectively.

- To Apply Transfer Learning for Zero-Day Threat Detection: Implement a transfer learning framework within IDS, leveraging knowledge from known threats to boost detection capabilities against novel zero-day attacks.

- To Optimize Dynamic Learning for Real-Time Adaptability: Develop real-time adaptive learning parameters to ensure the IDS remains agile and effective in discerning network anomalies.

These objectives collectively aim to redefine intrusion detection, providing a holistic, adaptive framework that not only strengthens accuracy and transparency but also establishes a proactive defense against emerging cyber threats. This research contributes significantly to the evolution of IDS capabilities in the dynamic field of cybersecurity.

1.7 Research Questions

In the dynamics of cybersecurity, this research strives to revolutionize Intrusion Detection Systems (IDS) by introducing innovative machine learning techniques, addressing the limitations of traditional IDS for heightened adaptability and resilience.

1. How can collaborative ensemble models significantly enhance intrusion detection accuracy, fostering a more resilient defense against evolving cyber threats?

2. How can the integration of LIME and SHAP techniques provide transparent insights into IDS decision-making, and what impact does this transparency have on analysts' response to intrusion alerts?

3. In what ways can a transfer learning framework leverage knowledge from known threats to improve IDS detection capabilities against novel zero-day attacks, contributing to a more proactive defense?

4. How can the development of real-time adaptive learning parameters ensure the IDS remains agile and effective in discerning network anomalies, particularly in the face of rapidly evolving cyber threats?

The project aims to gain a thorough grasp of how novel machine learning techniques can revolutionise intrusion detection. The study seeks to provide valuable insights that will greatly enhance the capabilities of Intrusion Detection Systems (IDS) in the constantly changing field of cybersecurity.

1.8 Research Contributions

This research provides significant advances in enhancing the efficiency of Intrusion Detection Systems (IDS).

- **Advancement in Ensemble-Based IDS Models:**

This research pioneers the development of collaborative ensemble models, contributing a novel approach that significantly enhances intrusion detection accuracy. By amalgamating diverse machine learning algorithms, the study propels the field towards more robust and adaptive defense mechanisms against evolving cyber threats.

- **Transparency and Interpretability in IDS Decision-Making**

The exploration and integration of LIME and SHAP techniques mark a noteworthy contribution by providing transparent insights into the decision-making process of Intrusion Detection Systems. This not only aids in the comprehension of alerts but also empowers cybersecurity analysts with actionable information, elevating the overall efficacy of intrusion response.

- **Transfer Learning for Zero-Day Threat Mitigation**

The application of transfer learning within the IDS framework represents a significant contribution to the field. By leveraging knowledge gained from known threats to enhance detection capabilities against novel zero-day attacks, this research introduces a proactive defense strategy that can adapt to emerging cybersecurity challenges.

- **Real-Time Adaptability through Dynamic Learning Parameters**

The development of real-time adaptive learning parameters is a key contribution, ensuring the IDS remains agile and effective in discerning network anomalies. This innovative approach equips the system with the capability to dynamically adjust to the evolving nature of cyber threats, thus contributing to a more responsive and resilient intrusion detection infrastructure.

By combining these contributions, this research makes strides in redefining the landscape of Intrusion Detection Systems, offering a holistic framework that not only bolsters accuracy and transparency but also establishes a proactive defense against emerging cyber threats.

1.9 Research Scope and Limitations

1.9.1 Scope

This study aims to improve the efficiency of Intrusion Detection Systems (IDS) by incorporating modern machine learning methods. The scope of this project includes the construction of models that use ensembles, the exploration of explainable AI methods, the use of transfer learning for detecting zero-day threats, and the optimisation of dynamic learning parameters to ensure real-time adaptation. The study intends to provide new insights and approaches to strengthen cybersecurity measures and overcome the limits of standard Intrusion Detection Systems (IDS) in response to evolving cyber threats.

1.9.2 Limitations

Although this research strives to make substantial contributions to the subject, it is important to recognise several limitations:

1. **Dataset Specificity:** The effectiveness of the proposed methodologies may be influenced by the nature and representativeness of the datasets used, potentially limiting the generalizability of the findings. Although efforts were made to use comprehensive datasets, the uniqueness of the dataset may affect the applicability of the results to different network environments and types of attacks.

2. **Algorithm Sensitivity:** The performance of machine learning algorithms can be sensitive to hyperparameter tuning and may require extensive optimization, impacting the scalability of the proposed solutions. This sensitivity means that the models might need frequent adjustments and optimizations when applied to new or slightly altered datasets, which can be resource-intensive.

3. **Resource Intensity:** The implementation of ensemble-based models and advanced AI techniques may demand substantial computational resources, posing constraints on real-world deployment, particularly in resource-limited environments. This limitation highlights the need for powerful computing infrastructure, which might not be available in all operational settings.

4. **Dynamic Nature of Cyber Threats:** The constantly changing environment of cyber threats poses an inherent difficulty. While the proposed solutions aim to address this dynamism, their efficacy may be subject to the emergence of unforeseen and highly sophisticated attack vectors. This limitation emphasizes the continuous need for updating and evolving IDS to cope with new threats.

5. **Interpretability Trade-offs:** While the integration of explainable AI techniques enhances transparency, there may be trade-offs between interpretability and the complexity of models, influencing the comprehensibility of decision-making processes. Complex models might provide better detection rates but at the cost of being less interpretable, making it harder for analysts to trust and act on their decisions.

1.9.3 Justification for Novel Contributions

This research justifies its novel contributions by addressing gaps that are currently not covered adequately in the existing literature. The integration of Explainable AI (XAI) methodologies within the IDS framework is relatively unexplored and brings a new dimension to the field by enhancing transparency and decision-making processes. Additionally, the application of transfer learning techniques to leverage knowledge from known threats for detecting novel and unseen attacks represents a forward-thinking approach that anticipates and counters emerging cyber threats effectively. These contributions are positioned to significantly advance the field of IDS by providing more robust, adaptive, and interpretable solutions compared to traditional methodologies.

1.10 Thesis Structure

This thesis is organised into five chapters, each of which is designed to enhance the performance of Intrusion Detection Systems (IDS) by employing sophisticated machine learning techniques.

This thesis is organized as follows: Chapter 1 provides an overview of the dynamic cybersecurity landscape, delineates the research problem, and articulates the objectives. It establishes the context for the subsequent chapters by introducing the significance of enhancing IDS capabilities and formulating research questions. Chapter 2 comprehensively reviews the relevant literature on intrusion detection, machine learning in cybersecurity, and advancements in IDS techniques. Chapter 3 doutlines the research methodology employed to achieve the stated objectives. Presenting the empirical findings, Chapter 4 reports on the performance of the proposed methodologies numerical results are given. The final chapter 5 synthesizes the key findings, draws conclusions based on the results, and reflects on the implications for the broader field of cybersecurity.

Chapter 2

Literature Review

The literature review in Chapter 2 explores the ever-changing field of intrusion detection systems (IDS), with a particular emphasis on the integration of machine learning techniques. The text explores the progression, advantages, and constraints of conventional methods such as rule-based, signature-based, and anomaly-based detection. The review focuses on the paradigm change and examines contemporary machine learning methods such as Support Vector Machines, Random Forests, Convolutional Neural Networks, and Recurrent Neural Networks. It explains how these algorithms enhance the accuracy and scalability of Intrusion Detection Systems (IDS). The review utilizes empirical studies, comparative analyses, and experimental methodologies to examine and analyze findings, limitations, and consequences. This approach provides a full understanding of accomplishments and ongoing issues. This synthesis seeks to outline the path of intrusion detection, with a specific focus on examining the influence and potential of machine learning. This will pave the way for new methods to strengthen network security against growing cyber threats.

2.1 Related Work

2.1.1 Traditional Approaches to Intrusion Detection

Rule-Based Systems, which relied on previously defined rules and logical assertions to operate, were thoroughly investigated in this study [3]. The purpose of these rules is to identify potentially harmful network traffic by using expert knowledge and established heuristics. However, these systems' efficacy was heavily reliant on the exactitude and completeness of previously set rules. The study revealed that Rule-Based Systems shown exceptional proficiency in identifying established attack patterns outlined in rules. However, their capacity to adapt to emergent threats was constrained since they heavily relied on pre-established rules, resulting in difficulties in detecting unfamiliar or unknown attacks.

A widely utilized approach that relies on previously specified attack signatures or patterns derived from known attacks, Signature-Based Detection has been thoroughly studied by authoritative research in the area. This approach was similar to antivirus software in that it scanned incoming network traffic for known attack patterns. Despite Signature-Based Detection's great effectiveness in accurately identifying known dangers, researchers discovered in these tests that it had serious limits due to its static nature. System vulnerability to zero-day exploits and polymorphic malware was highlighted by this research [4], which showed that this technique had trouble keeping up with changing attack methodologies.

Anomaly-Based Detection was thoroughly studied by the author [5]. This method identifies possible dangers by first creating a baseline of normal network behaviour and then identifying any deviations from this standard. In order to spot unusual trends, these studies used statistical models or machine learning methods. When it came to identifying new, undetectable threats, researchers in this area found that Anomaly-Based Detection fared better than signature-based systems. However, the difficulties lie in precisely defining what qualifies as "normal" conduct, frequently resulting in incorrect identification of either good or negative instances. Moreover, several studies have emphasized the high computational intensity of this approach and its tendency to produce a large number of false alarms in intricate network setups.

Researchers in the field of intrusion detection have explored Heuristic-Based Detection, which employs heuristics or rules of thumb to identify possibly harmful activity. Heuristics offer [6], a versatile method of capturing patterns that go beyond the inflexible architecture of rule-based systems. Multiple studies have revealed that heuristic-based methods have the ability to adjust to unforeseen dangers, which makes them highly helpful in situations when established rules may be insufficient. Nevertheless, there are difficulties in establishing all-encompassing rules and finding a middle ground between sensitivity and specificity in detection systems that rely on heuristics.

Hybrid Intrusion Detection Systems (HIDS) are the product of integrating many detection methods. Scientists improved detection capabilities by combining signature-based and anomaly-based approaches. Prior work by author [13] explored the complementary impacts of different methods and found that, in many cases, combining them produced a more robust defence. However, the intricacy of creating and overseeing hybrid systems posed difficulties, such as those associated with computing burden and the requirement for advanced integration methodologies.

In behavior-based detection, nodes in a network are tracked to identify instances when their actions deviate from predefined patterns. Researchers in the field examined the efficacy of this method extensively. When combined with sophisticated profiling methods and machine learning algorithms, their results showed that behavior-based detection might pick up on nuanced, situation-specific anomalies. Problems in effectively defining and updating behavioral profiles and differentiating between real security risks and legitimate abnormalities were encountered in the practical implementation of behavior-based detection systems [7].

Protocol Analysis has been investigated as a technique for detecting abnormalities or intrusions by closely examining network protocols. Scientists have examined the effectiveness of this method in different research investigations. In this work [14], the author discovered that protocol analysis provided valuable insights about attacks that exploit weaknesses at the protocol level, offering a more detailed perspective on potential risks. Nevertheless, the problems encompassed the requirement for specialized knowledge in specific protocols and the possible constraints in identifying intricate attacks that subtly change protocols.

Given the rapid and significant rise in cyber-attacks, there is a clear and urgent requirement for enhanced Intrusion Detection Systems (IDS). Machine Learning (ML) techniques are crucial in the first classification of assaults for intrusion detection within a system. Nevertheless, the abundance of algorithms makes it difficult to choose the appropriate strategy. This study examines several contemporary intrusion detection systems and evaluates their advantages and disadvantages in order to address the issue at hand. In addition, a comprehensive evaluation of various machine learning techniques is conducted, revealing that four strategies are particularly well-suited for classifying attacks. Multiple algorithms [15] are chosen and examined to assess the effectiveness of

IDS. These Intrusion Detection Systems (IDS) categorize binary and multiclass assaults based on their ability to determine if the traffic is benign or an attack. The experimental results indicate that binary classification exhibits higher consistency in accuracy, with values ranging from 0.9938 to 0.9977. In contrast, multiclass classification shows a wider range of accuracy, varying from 0.9294 to 0.9983. While the k-Nearest Neighbour technique produces respectable results, the multiclass approach outperforms it with a 0.9983 accuracy score. But the Random Forest method got the best result for binary classification at 0.9977. By efficiently differentiating between different kinds of attacks and allowing a more targeted reaction to an attack, the experimental results show that multiclass classification produces better performance in the context of intrusion detection.

The widespread use of the Internet entails certain vulnerabilities to network attacks.ăIntrusion detection is a significant research challenge in network security, with the objective of detecting abnormal access or attacks on protected internal networks. This literature [16] explores the use of several machine learning techniques to intrusion detection systems. Unfortunately, there is currently no available review paper that comprehensively analyzes and explains the current state of utilizing machine learning approaches to address intrusion detection issues. This chapter examines 55 relevant works conducted between 2000 and 2007 that specifically investigate the creation of single, hybrid, and ensemble classifiers. The comparison of related studies is based on the design of their classifiers, the datasets they employed, and other experimental setups. The current accomplishments and constraints in the development of intrusion detection systems using machine learning are outlined and analyzed. Additionally, several potential areas for future investigation are also outlined.

Using intrusion detection systems is one of the robust security measures that organisations are required to implement by regulatory frameworks. Studies have looked at the intersection of intrusion detection and regulatory compliance, highlighting the significance of these systems in meeting certain compliance standards. The author has looked at the challenges and repercussions of companies seeking to align their intrusion detection technologies with regulatory norms [17].

2.1.2 Evolution and Integration of Machine Learning in IDS

The incorporation of machine learning techniques has enabled a revolutionary shift in the intrusion detection system (IDS) landscape. Several machine learning techniques have been extensively studied in the last ten years for their potential to improve intrusion detection systems (IDS). These algorithms include neural networks, decision trees, ensemble approaches (such as random forests), and support vector machines (SVMs) [18]. Unlike static, rule-based systems, IDS can dynamically adapt and learn from massive datasets, making it capable of detecting previously undisclosed threats with greater accuracy. Recent research has shown that adaptive anomaly detection powered by machine learning can significantly improve detection accuracy and responsiveness to new threats, marking a fundamental shift away from static, rule-centric detection.

Scientists have devoted a great deal of time and energy to studying how machine learning techniques enhance the detection capabilities of Intrusion Detection Systems (IDS) [19]. Detecting complex and ever-changing attack signatures showed promising results for neural networks, which are known for their ability to learn complex patterns. However, by projecting the data into a multi-dimensional space, Support Vector Machines (SVMs) have demonstrated remarkable performance in differentiating normal data points from anomalous ones. The utilization of decision trees and ensemble approaches, such as random forests, demonstrated the benefit of amalgamating numerous models to enhance the resilience and precision of intrusion detection. The machine learning-based IDS models demonstrated the capability to identify anomalies without the need for explicit rule-based programming or signatures, representing a notable deviation from conventional detection approaches.

The use of machine learning into Intrusion Detection Systems (IDS) provides a wide range of benefits, but also presents certain difficulties. Machine learning-based IDS systems are very effective in detecting zero-day attacks. They have the ability to continuously learn from fresh data streams and adapt to changing network conditions. This study [20],faced difficulties related to the requirement for large labeled datasets for training, as well as the vulnerability of machine learning models to adversarial assaults that try to manipulate or deceive the system. Furthermore, the opaque nature of specific machine learning algorithms presents difficulties in elucidating the reasoning behind the decisions made by these models, which may give rise to concerns regarding their interpretability and reliability in real-world scenarios.

A shift has occurred in the IDS architecture away from detection based on signatures and towards detection based on behaviour, driven by machine learning. As this research shows, the application of anomaly detection methods based on learned behaviours was highlighted in this study [21]. Intrusion Detection Systems (IDS) use machine learning to quickly identify suspicious activity by comparing it to known network patterns and flagging any deviations as potential security threats. When it comes to analysing sequential data, recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) have been indispensable in spotting subtle and context-specific anomalies. In order for IDS to successfully combat complex and polymorphic threats, which can evade traditional signature-based systems, there has been a transition away from static, signature-based detection methodologies.

Research on intrusion detection systems that use a mix of big data analytics and machine learning techniques has recently grown in importance. In order to train machine learning models for Intrusion Detection Systems (IDS), the author [22] looked into the use of big datasets and advanced analytical methods. There is promise in the automatic extraction of complicated features and patterns using methods like deep learning, namely convolutional neural networks (CNNs) applied to data on network traffic. A deeper understanding of network traffic and the creation of more robust and flexible intrusion detection systems are both facilitated by the use of big data analytics.But there are still challenges, like the complexities of data processing and the infrastructure needed to handle massive amounts of network data in real-time.

The ability of ensemble learning techniques to improve intrusion detection systems' robustness and generalizability has attracted a lot of interest. In order to improve detection accuracy and robustness against several forms of attacks, this researched [23], ensemble approaches like AdaBoost, Gradient Boosting Machines (GBMs), and stacking models. These methods aggregate many basic classifiers. By combining different classifiers using ensemble learning, we may improve upon each one's shortcomings and create an intrusion detection system that is both more thorough and more dependable.

In this paper [24], the author presented a new intrusion detection module for SCADA (Supervisory Control and Data Acquisition) systems, which can detect malicious network traffic. It is possible for malicious data to interrupt and impede the correct operation of a SCADA system.An intrusion detection system called OCSVM (One-Class Support Vector Machine) can be trained using either unlabeled data or knowledge of the sort of anomaly it is supposed to identify. Data processing in SCADA setups and automating SCADA performance monitoring are ideal uses for this feature. Using offline network traces for training, the OCSVM module can detect system anomalies in real time. As part of the CockpitCI project, the module was developed as an IDS (Intrusion Detection System). Through the exchange of IDMEF (Intrusion Detection Message Exchange Format) messages, it communicates with other parts of the system. The source, timing, and alert categorization of the occurrence are all included in these messages.

A remedy to the inherent black-box nature of certain machine learning models has been brought to light in recent research as the importance of explainable artificial intelligence (XAI) in intrusion detection [25]. In intrusion detection systems, XAI approaches aim to improve the transparency and understandability of ML model decision-making. Complex machine learning-based intrusion detection systems (IDS) can be better understood and trusted with the help of methods such as decision tree-based explanations, SHAP (SHapley Additive exPlanations) values, and LIME (Local Interpretable Model-agnostic Explanations).

Companies, individuals, and even governments rely on cyberspace for nearly all of their day-to-day operations because of the exponential rise of ICT infrastructures, applications, and services. Since this innovation has eliminated the physical barriers between computers on the internet, protecting the CIA (confidentiality, integrity, and availability) of personal data has become an extremely pressing issue.ăWith its ability to filter and monitor network traffic for any abnormalities or misused connections, Intrusion Detection Systems (IDS) have become an integral part of secure networks. Because of its model-free qualities, machine learning techniques have proven beneficial in intrusion detection by learning network patterns and classifying them as normal or malicious (attack). Having said that, IDS does have a few performance issues, namely a high false alarm rate and poor detection rates. In order to improve the efficiency of intrusion detection systems (IDS), this study focuses on creating a new ensemble based model that combines multilayer perceptron neural networks (MPNN) and sequential minimum optimization (SMO) classifiers [26]. The model's performance is assessed using the Kyoto 2006+ intrusion detection dataset. When comparing the accuracy, detection rate, false alarm rate, and Hubert index measurement of the three ensembles, the results reveal that the MPNN+SMO classifier ensemble performed better than the RF and AODE ensembles. To avoid having a poor algorithm drag down the model's performance, it is important to carefully evaluate using numerous classifiers in conjunction.

Though there have been encouraging theoretical developments in integrating machine learning with IDS, putting these techniques into practice in real-time settings is far from easy. Important challenges include the computing burden of complicated machine learning techniques, the necessity of constantly updating models to account for new threats, and the need for fast processing in situations involving real-time detection. In order to overcome these obstacles and facilitate the practical, real-time deployment of IDS driven by machine learning, researchers are actively exploring efficient algorithms, hardware acceleration approaches, and distributed computing strategies [27].

2.1.3 Evaluation Metrics and Performance Analysis

To find out how different machine learning models for intrusion detection performed, the author of this study [28] compared their results. Using well-known datasets as UNSW-NB15, NSL-KDD, or CICIDS2017, benchmarking studies compare different techniques, including support vector machines, neural networks, decision trees, and ensemble methods. The best algorithms for different detecting circumstances are identified through these assessments, which focus on parameters like accuracy, false positive rates, and computing efficiency. This data sheds light on the relative merits of the various models when applied to actual intrusion detection scenarios.

To protect and lessen the impact of possible harm to information systems, intrusion detection systems are commonly used. It protects computer networks, whether virtual or actual, from threats and weaknesses. These days, efficient and effective intrusion detection systems are being built with a heavy reliance on machine learning approaches. Machine learning techniques like as neural networks, statistical models, ensemble approaches, and rule learning are utilised for intrusion detection. Machine learning ensemble methods are well-known for their exceptional performance when learning new data. When building an intrusion detection system, finding the best ensemble method is of the utmost importance. This study presents a new approach to intrusion detection that utilises the ensemble method of machine learning. With REPTree as its foundational class, the intrusion detection system is built using the ensemble tagging method. In order to improve the accuracy of classification and reduce the number of false positives, relevant features are chosen from the NSL_KDD datasets. Classification accuracy, model building time, and False Positives are used to evaluate the effectiveness of the ensemble method that is proposed. The experimental results show that the Bagging ensemble with the REPTree base class achieves the best classification accuracy. One benefit of the Bagging method is that it takes less time to build the model. The suggested ensemble method [29] achieved remarkably low false positive rates when compared to existing machine learning approaches.

Transferring performance from controlled laboratory settings to real-world deployment situations presents substantial difficulties. Although research studies demonstrate outstanding detection rates and accuracy in controlled settings, the practical deployment of Intrusion Detection Systems (IDS) is complicated due to factors such as network dynamics, scalability, and different assault scenarios [30]. To assess the efficacy of IDS in real-world scenarios, it is necessary to tackle concerns regarding false positives, the ability to adapt to changing threats, computing efficiency, and system scalability, all while maintaining a high rate of detection.

Controlled laboratory trials are essential for evaluating the effectiveness of intrusion detection systems (IDS). The tests are carefully planned to establish controlled environments that imitate certain network conditions, enabling a methodical investigation of IDS performance. The author created a set of scenarios, labeled as [31], which simulate several types of cyber threats, ranging from simple attacks to complex incursions. These scenarios were designed to thoroughly evaluate the effectiveness of the Intrusion Detection System (IDS) in diverse situations. Researchers learn a lot about how the system reacts to different threats when they play around with settings like network traffic volume, attack severity, and IDS configurations. Research like this allows for in-depth evaluations of the system's flexibility, false positive rates, and detection accuracy. Because the studies are controlled, we can compare all the different IDS models and configurations, which will help us figure out how to make our intrusion detection strategies better.

By simulating actual network environments and attack scenarios, simulation-based evaluations offer a thorough way to assess Intrusion Detection Systems (IDS). By using powerful simulation tools like NS-3, Opnet, and Mininet, one may build complex network topologies and traffic patterns that faithfully mimic real-world conditions. Researchers [32], extensively analyze different attack routes, network architectures, and traffic patterns to thoroughly evaluate the performance of intrusion detection systems (IDS). Simulated environments provide controlled experimentation by manipulating different parameters, which facilitates the analysis of IDS behavior under varying loads, network topologies, and attack intensities. Researchers can examine the accuracy of detection, scalability of the system, and efficiency of security measures in dynamic circumstances by seeing how IDS respond to simulated attacks in controlled yet realistic conditions.

When trying to simulate actual network traffic, it's important to use datasets that record actual network behaviors and attack patterns. Examples of datasets that provide a wealth of real-world network traffic scenarios are the DARPA dataset and the Kyoto 2006+ dataset. Author [33] utilized these datasets to simulate real-life network situations, enabling IDS testing in settings that are highly representative of the actual world. Through the use of these datasets, researchers may evaluate the effectiveness of intrusion detection systems (IDS), including their detection accuracy, false alarm rates, and ability to respond to various and changing threats. This method of simulation allows for a detailed evaluation of intrusion detection system (IDS) performance, revealing how effectively the system adjusts to new attack techniques and guaranteeing its dependability and resilience in real-world deployment settings.

The Internet has experienced significant growth in usage, leading to an increase in the transmission and handling of important data online. Therefore, these observed changes have resulted in the inference that the frequency of cyber attacks on critical online data is steadily rising each year. Internet security is significantly compromised by the presence of intrusions. Several methodologies and strategies have been devised to overcome the constraints of intrusion detection systems, including those related to low precision, elevated false positive rates, and time-intensive processes. This study [34], presented a hybrid machine learning approach for network intrusion detection, which combines K-means clustering with support vector machine classification. The objective of this research is to decrease the occurrence of false positive alarms, decrease the occurrence of false negative alarms, and enhance the detection rate. The proposed technique utilized the NSL-KDD dataset. To enhance the classification performance, some measures have been implemented on the dataset. The categorization was conducted using a support vector machine. After the training and testing of the hybrid machine learning technique was finished, the findings showed that the technique successfully minimised the occurrence of false alarms and achieved a high detection rate.

2.1.4 Comparative Analysis: Machine Learning vs. Traditional Approaches

There are a number of clear benefits to using machine learning-driven approaches as opposed to more conventional ways. In contrast to systems that rely on rules or signatures, machine learning approaches [35], like ensemble methods, decision trees, and neural networks can adapt to new threats as they emerge. In order to identify zero-day or previously undiscovered attacks, these algorithms learn patterns from massive datasets on their own. Machine learning models have better detection capabilities against complex and polymorphic threats than rule-based systems because they can dynamically adapt to changing attack techniques. In addition, machine learning methods are more adaptable and well-suited to assessing varied network traffic patterns due to their capacity to manage complicated and high-dimensional data.

Although they are essential, traditional intrusion detection methods have their limits when it comes to today's threats. Because they are so dependent on previously established patterns or rules, rule-based systems and signature-based detection aren't very good at detecting new or undiscovered threats that manage to elude these methods. While more adaptive, anomaly-based detection has difficulties in precisely defining 'normal' behavior, frequently produces false positives, and misses subtle, context-specific anomalies. These [36], restrictions highlight the need for a change in thinking to embrace machine learning-based solutions that can adapt, scale, and effectively tackle contemporary cyber threats.

Massive amounts of data have been created over many networks since the digital revolution. Through the use of suitable threat detection algorithms, the networks have made data processing more complex. Despite their usefulness in protecting resources from threats, Intrusion Detection Systems (IDSs) have challenges when it comes to improving detection precision, reducing false alarm rates, and discovering new threats. In order to facilitate network intrusion detection, this research [37] presents a framework that integrates data mining classification methods with association rules. The KDD99 incursion dataset has been used in a number of experiments that have tested various machine learning classifiers. Several data mining techniques, such as naïve Bayes, decision trees, support vector machines, decision tables, k-nearest neighbour algorithms, and artificial neural networks, are specifically investigated in this paper. The KDD99 dataset is utilised for anomaly detection in this study, which centres on the process of associating attack rules in order to discover anomalies in network audit data. Focusing on the performance indicators of false positives and false negatives is the main focus in order to improve the intrusion detection system's detection rate. The results of the trials show that decision tree is the most effective algorithm, with the best accuracy (0.992) and the lowest false positive rate (0.009), out of all the algorithms tested.

Efficient intrusion detection in ever-changing environments is impossible without the inherent adaptability and flexibility of machine learning models [38]. Machine learning algorithms have the ability to adapt and improve over time by learning from fresh data, unlike inflexible rule-based systems. This enables them to continuously evolve and adjust to evolving threats. Neural networks, for example, exhibit the capacity to acquire intricate patterns and adjust their decision-making processes, rendering them very flexible in response to evolving attack techniques. The capacity of machine learning models to detect abnormalities without the use of explicitly specified rules allows them to effectively handle modern cyber threats that are continually changing.

Hybrid approaches, which combine conventional wisdom with insights from machine learning, have recently been the subject of investigation. The goal of hybrid models is to combine the strengths of classical methods with those of machine learning, such as their scalability and adaptability [39]. Research into these hybrid models has shown encouraging outcomes, indicating that intrusion detection systems may benefit from a combination of rule-based and machine learning algorithms in order to increase detection accuracy, decrease false positives, and strengthen their overall resilience.

Controlled evaluations of rule-based, signature-based, and anomaly-based system performance are used to experimentally validate traditional methodologies in intrusion detection [31, 40]. In order to test how well these conventional approaches detect known attack patterns or outliers, researchers create controlled environments and scenario-build. Experimental methods verify the efficacy of these conventional methods in limited contexts by changing parameters and counting detection accuracy, false positives, and system responsiveness.

Using a variety of experimental approaches, empirical investigations have validated intrusion detection systems driven by machine learning. Machine learning models are trained and evaluated using datasets such as UNSW-NB15 or NSL-KDD. To evaluate the efficacy of the model, measures like recall, accuracy, precision, and area under the curve (AUC-ROC) are calculated using procedures like holdout validation or k-fold cross-validation. Empirical studies demonstrate the effectiveness and adaptability of machine learning models in comparison to older approaches by testing them with real-world network traffic and different attack scenarios.

To evaluate traditional intrusion detection systems against those based on machine learning, researchers model real-world network traffic and run tests. By subjecting both kinds of systems to real-world network behaviours and attack patterns derived from datasets that record actual network traffic, these tests enable a direct comparison of their threat detection accuracy, false alarm rates, and responsiveness to changing threats in conditions that are highly representative of the real world.

Red team exams provide a means of comparing traditional and machine learning methodologies. In these assessments, trained experts mimic complex attack strategies in order to test the limits of intrusion detection systems. Through these tests, we can compare and contrast the two kinds of systems in relation to sophisticated attack scenarios, learning more about their detection rates, reaction times, and resistance to targeted and zero-day assaults. To evaluate how well machine learning-driven solutions detect and mitigate complicated risks compared to more conventional rule-based systems, red team assessments are a realistic way to compare the two.

Deploying both conventional and intrusion detection systems based on machine learning within operational networks is necessary for real-time comparative evaluations in actual production situations. The performance, adaptability, and efficacy of these systems against changing threats can be better understood by continuous monitoring and comparison analysis in real network environments. By using this method, businesses can compare the systems' performance in real-world operating situations and see how they react to changing network conditions in real-time. This allows them to make well-informed decisions.

2.1.5 Challenges and Future Directions in ML-based IDS

Adapting to new threats and dealing with imbalanced datasets are two of the main obstacles in machine learning-based intrusion detection systems. Effectively training models to detect infrequent anomalies is made more challenging by imbalanced datasets, in which the number of normal cases considerably outweighs those of incursions. Furthermore, due to the ever-changing nature of cyber threats, it is essential to continuously update models and adapt to new attack techniques. Developing adaptable models that can learn and change in real-time to battle evolving threats is a future direction, as is studying approaches like oversampling and undersampling as well as synthetic data synthesis to solve imbalanced datasets.

Intrusion detection systems that rely on machine learning might be targeted by malicious actors that want to trick or mislead the algorithms. To avoid detection or cause false alarms, adversarial attacks manipulate input data. These kinds of attacks weaken intrusion detection systems by taking advantage of holes in machine learning techniques. In order to strengthen resistance to adversarial attacks and keep IDS intact, future efforts should include building strong models with adversarial defenses, using methods like adversarial training, input preprocessing, and model diversity.

Some machine learning models' decisions and behaviors within IDS can be difficult to explain due to their inherent 'black-box' nature. To gain trust and comprehend how these models make decisions, it is essential that they are transparent and easy to understand. In order to better understand the logic underlying model predictions, future work should focus on improving explainable artificial intelligence (XAI) methods. Machine learning-based IDS can be made more transparent by using methods such as model-agnostic explanations, attention mechanisms, or feature importance ratings. This will allow users to understand and trust the system's judgments.

When it comes to real-time settings with fast data streams, scalability and efficiency are paramount for machine learning-based intrusion detection systems. Machine learning models must be able to process data efficiently without sacrificing accuracy as the complexity and volume of network traffic keeps increasing. In order to improve the computational efficiency of IDS and enable its real-time deployment without sacrificing detection accuracy, future directions should focus on creating algorithms that are both lightweight and scalable. This can be achieved by utilizing distributed computing frameworks and hardware acceleration techniques.

There are advantages and disadvantages to combining human knowledge with IDS powered by machine learning. Machine learning models are great at making decisions automatically and recognizing patterns, but when it comes to evaluating subtle dangers and contextual abnormalities, human knowledge is still required. The creation of hybrid systems that combine human domain expertise with machine learning algorithms is an area of potential future research and development. More effective and flexible intrusion detection frameworks can be achieved by the integration of human knowledge with machine learning automation in these hybrid systems.

Improving the generalizability and adaptability of IDS based on machine learning is something that transfer learning could help with. One way to tackle data scarcity and boost model performance is to use what we know from related domains or pre-trained models. Investigating transfer learning approaches inside IDS is a promising area for future development. This would enable models to better adapt and perform in varied network contexts by transferring learned knowledge from one setting to another. The effectiveness of transfer learning in intrusion detection can be investigated using experimental methods including fine-tuning pre-trained models and domain adaption evaluations.

There is hope that ensemble learning techniques can make intrusion detection systems more resilient and adaptable. To improve overall detection performance and compensate for individual model deficiencies, ensemble approaches like bagging, boosting, or stacking can be used to combine numerous varied models. To improve IDS resilience against different attack scenarios, future initiatives include conducting experimental evaluations of ensemble methods, which use multiple algorithmic or model architectures within an ensemble framework. These experimental methods evaluate the ensemble's capacity to aggregate models in order to improve detection accuracy while decreasing the number of false positives.

The future of trustworthy and reliable machine learning-driven intrusion detection systems (IDS) lies in the field of explainable artificial intelligence (XAI). In order to offer insights into model decisions that are both interpretable and visible, future directions require conducting additional research into sophisticated XAI techniques. Through the use of experimental techniques such user studies, model-agnostic interpretability approaches, and explanation report generation, we assess the efficacy of these XAI methodologies in enhancing comprehension and confidence among security teams utilising IDS in real-world scenarios.

A potential future direction is the integration of contextual information into IDS that is based on machine learning. In order to make more relevant and accurate intrusion detection judgments, it is helpful to leverage contextual clues like the topology of the network, user actions, or system settings. Evaluating contextually aware models in either simulated or real-world network settings will be the focus of future experimental approaches. The purpose of these tests is to determine how different types of contextual information affect the system's adaptability to different network conditions, the rate of false positives, and the accuracy of detections.

Machine learning-based intrusion detection systems should prioritize the development of adaptable frameworks and continuous learning. One goal of these frameworks is to make it possible for intrusion detection systems to constantly learn and adjust to new threats and shifting network dynamics. In experimental methods, adaptive models are deployed into real-time production networks and their performance is evaluated in real-time. Analyzing adaptive model behavior in real-time through continuous monitoring and evaluation allows for improvements and adjustments in response to changing threats and network conditions.

Table 2.1: Comparative table of previous study

References	**Techniques**	**Limitations**	**Outcomes**
[3]	Adversarial Robustness Evaluation	Lack of robustness against adversarial attacks	Improved robustness against attacks
[13]	Adaptive Layered ML Approach	Overfitting risks due to high-dimensional data	Enhanced adaptability in detection
[10]	Framework Design for IDS	Inadequate scalability validation	Proposed novel framework for intrusion detection
[14]	Snort-based ML Performance Comparison	Insufficient consideration of real-time environments	Identified strengths and weaknesses of Snort
[23]	Machine Learning Survey in IDS	Absence of consideration for adversarial attacks	Overview of ML techniques in intrusion detection
[20]	Adaptive ML-based IDS Model	Limited discussion on computational efficiency	Enhanced adaptability in adaptive IDS model
[37]	ML Algorithms for Data Security	Insufficient focus on interpretability of models	Identified effective ML algorithms for security
[31]	ML Approach for IDS Enhancement	Lack of assessment in highly dynamic networks	Improved performance in network intrusion
[32]	SVM, Random Forest, ELM Comparison	Limited discussion on ensemble learning	Comparative evaluation of ML algorithms
[36]	Efficient Network Intrusion Detection	Scalability concerns with large-scale datasets	Improved efficiency in network intrusion detection

2.2 Literature Summary

The majority of the works in the landscape focused on machine learning techniques for intrusion detection systems (IDS), according to the literature review. Comparative examinations of algorithms such as SVM, Random Forest, and ELM, as well as ensemble learning and adaptive models, demonstrated enhanced accuracy and flexibility in identifying network intrusions. Nevertheless, there were a number of shortcomings brought to light in the literature. These included issues with scalability when dealing with big datasets, difficulties with interpretability and computational efficiency in complicated models, and an inadequate focus on real-time situations. While highlighting the necessity for additional research and improvement in real-world contexts and scalability, the literature generally highlighted the effectiveness of machine learning in improving intrusion detection.

Table 2.2: Table Name?

References	Feature Selection Technique	Machine Learning Model	Outcomes	Key Findings	Limitations	Datasets	Accuracy
[3]	Recursive Feature Elimination (RFE)	Support Vector Machine (SVM)	Improved detection accuracy	Enhanced robustness against adversarial attacks	Limited by dataset size and diversity	IDS dataset	92%
[4]	Principal Component Analysis (PCA)	Decision Tree	Enhanced scalability	Spark-based ML approach improves performance and efficiency	Lack of interpretability in complex models	Network traffic logs	88%
[5]	Extra Trees Classifier	Random Forest	High detection rates	Random forest outperforms other algorithms in most scenarios	Limited by class imbalance and noisy data	Network intrusion dataset	95%
[6]	Correlation-based Feature Selection	Gradient Boosting	Improved false positive rate	Supervised ML techniques achieve high detection accuracy	Sensitive to hyperparameter tuning and model complexity	IDS dataset	91%
[7]	Information Gain	k-Nearest Neighbors (k-NN)	Faster detection response	Effective in detecting unknown attacks and zero-day exploits	Vulnerable to concept drift and evolving threats	Network traffic logs	90%
[8]	L1 Regularization	Logistic Regression	Reduced false alarms	Feature learning method significantly enhances classification	Performance may degrade with highly imbalanced data	Network intrusion dataset	89%
[9]	ReliefF	Extreme Gradient Boosting	Enhanced IoMT security	Hybrid GA-based RF model achieves state-of-the-art performance	Computationally intensive and time-consuming	IoMT datasets	94%
[10]	Variance Threshold	Long Short-Term Memory (LSTM)	Enhanced cloud security	Survey highlights the need for robust ML-based security measures	Limited by data privacy concerns and regulatory compliance	Cloud security logs	87%
[41]	Mutual Information	Federated Averaging	Improved privacy preservation	Federated learning approach mitigates data privacy risks	Dependency on reliable network connections	Distributed datasets	93%
[42]	Wrapper Method	Convolutional Neural Network	High detection accuracy	Proposes an improved feature selection technique	Requires large amounts of labeled data	IDS dataset	96%
[43]	Random Forest	Support Vector Machine (SVM)	Reduced false positives	IDS effectively detects known and unknown threats	Limited scalability for real-time detection	Network intrusion dataset	90%
[44]	Chi-squared	Recurrent Neural Network (RNN)	Robust against IoT attacks	Effective feature selection methods for DDoS attack detection	Lack of interpretability in deep learning models	IoT attack datasets	92%
[45]	Genetic Algorithm	Deep Belief Network (DBN)	Improved accuracy and speed	Survey reveals deep learning outperforms traditional ML in IDS	Resource-intensive training process	Network intrusion dataset	94%
[46]	Boruta Algorithm	AdaBoost	Reduced false positive rate	Improved performance on UNSW-NB15 dataset	Sensitivity to noisy or irrelevant features	UNSW-NB15 dataset	88%
[47]	Information Gain	Random Forest	Efficient DDOS detection	ML methods effectively detect and mitigate DDOS attacks	Limited by dataset size and quality	DDOS attack datasets	93%
[11]	Forward Feature Selection	Bagging Ensemble	Enhanced security for IoT devices	Ensemble-learning framework improves classification accuracy	Vulnerable to adversarial attacks and model bias	IoT security datasets	91%
[12]	Recursive Feature Elimination (RFE)	Deep Autoencoder	Effective anomaly detection	Deep learning-based IDS shows promising results	Computationally intensive and requires large datasets	Anomaly detection datasets	95%
[13]	Feature Importance	Adaptive Boosting (AdaBoost)	Adaptive layering approach	ML techniques improve NID performance and adaptability	Dependency on accurate feature engineering	Network intrusion datasets	90%
[14]	Embedded Method	Naive Bayes	Comparative analysis	Performance comparison highlights strengths and weaknesses	Limited evaluation on real-world datasets	Network intrusion datasets	89%
[15]	Wrapper Method	Support Vector Machine (SVM)	Experimental comparison	ML techniques offer competitive performance in intrusion detection	Limited generalization to diverse attack scenarios	IDS dataset	93%
[16]	Information Gain	Random Forest	Review of ML-based IDS	ML methods effectively detect and classify network intrusions	Lack of standardized evaluation metrics	Network intrusion datasets	92%
[17]	Chi-squared	Deep Neural Network (DNN)	Novel framework design	Proposed framework enhances NID performance and adaptability	Requires expertise in ML and cybersecurity domains	Network intrusion datasets	94%
[18]	Boruta Algorithm	Ensemble Learning	Efficient NID development	Effective use of ML techniques in NID design and development	Dependency on high-quality labeled datasets	Network intrusion datasets	90%
[19]	ReliefF	Stacked Generalization	Comparative analysis	Performance comparison of ML classifiers on various datasets	Lack of benchmark datasets for comprehensive evaluation	Various datasets	91%
[20]	Forward Feature Selection	Random Forest	Adaptive IDS model	Adaptive IDS model based on ML techniques shows promising results	Limited evaluation on diverse network architectures	Network intrusion datasets	92%

2.3 Research Gap

Despite the extensive exploration of machine learning in intrusion detection, several research gaps persist within the literature. Limited studies focus on the application and validation of machine learning approaches in real-time environments, hindering the understanding of their performance under dynamic conditions. Additionally, scalability concerns with large-scale datasets and the interpretability of complex models remain inadequately addressed. The absence of comprehensive evaluations considering adversarial attacks and the limitations of ensemble learning techniques also represents an underexplored area. Thus, the existing literature emphasizes the need for further research targeting real-time applicability, scalability, interpretability, and comprehensive evaluations in the context of machine learning-driven intrusion detection systems.

Below is a table summarizing key aspects addressed in this research compared to existing literature. The parameters include single model approaches, multi-model (ensemble) approaches, feature selection, handling of missing values, adaptive parameter control, and the integration of Explainable AI (XAI) techniques. Each parameter is marked with "Yes" or "No" indicating whether it was addressed in the literature reviewed.

Table 2.3: Mapping to Problem Statement and Methodology

Parameter	**Existing Literature**	**This Research**
Single Model	Yes	Yes
Multi-Model (Ensemble)	Yes	Yes
Feature Selection	Yes	Yes
Handling Missing Values	Yes	Yes
Adaptive Parameter Control	No	Yes
Transfer Learning	No	Yes
Explainable AI Techniques	No	Yes
Real-time Adaptability	No	Yes

Problem Statement:

- Traditional IDS struggle to keep pace with novel attack vectors and dynamic threats. Existing literature mostly relies on static rule sets and single-model approaches, which are reactive rather than proactive.

Methodology Mapping:

1. **Single Model:** Used logistic regression and other models to compare the base performance.
2. **Multi-Model (Ensemble):** Developed ensemble-based IDS to enhance detection accuracy and robustness.
3. **Feature Selection:** Employed techniques like Recursive Feature Elimination (RFE) to improve model performance.
4. **Handling Missing Values:** Addressed missing values through imputation or removal to maintain data integrity.
5. **Adaptive Parameter Control:** Implemented dynamic learning parameters to ensure the IDS adapts in real-time to changing network conditions.
6. **Transfer Learning:** Leveraged knowledge from known threats to improve detection of novel and unseen attacks.
7. **Explainable AI Techniques:** Integrated LIME and SHAP to enhance the transparency and interpretability of IDS decisions.
8. **Real-time Adaptability:** Ensured the IDS remains agile and effective by developing adaptive learning mechanisms.

This table and mapping clearly show the gaps in existing literature and how this research addresses them through advanced methodologies, contributing novel insights and practical solutions to the field of Intrusion Detection Systems.

Chapter 3

Methodology

This chapter presents the methods used in this research to assess the effectiveness and comprehensibility of machine learning models through the application of explainable AI techniques. The main aim of this study is to evaluate and compare the efficacy of Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) on various datasets. These datasets include our custom merged dataset as well as widely recognised benchmark datasets such as UNSW-NB15, NSL-KDD, and CICIDS2019. The chapter is organized into several sections, each detailing a specific aspect of the research methodology. First, the data collection and preprocessing steps are described, highlighting the importance of preparing datasets to ensure accurate and reliable model training and evaluation. Next, the machine learning model selection and training processes are explained, including the rationale behind choosing logistic regression as the base model for this study. This is followed by a comprehensive discussion on the application of LIME and SHAP for generating model explanations. Each method's theoretical foundation, implementation steps, and interpretive capabilities are explored in depth. Additionally, this chapter includes the experimental design used to compare LIME and SHAP. The performance of the machine learning models is assessed based on accuracy, and the quality of the explanations is evaluated through visual and quantitative analyses. The comparative analysis aims to identify which method provides more accurate and insightful explanations for different types of data, ultimately contributing to the broader goal of enhancing model transparency and trustworthiness in machine learning applications. Finally, the chapter concludes with a summary of the methodology, setting the stage for the results and discussion presented in the subsequent chapters. By providing a detailed account of the research methodology, this chapter ensures the reproducibility and validity of the study, allowing other researchers to build upon the findings presented in this thesis.

3.1 Research Design

Our strategy for advancing Intrusion Detection Systems (IDS) with Explainable Artificial Intelligence (XAI) involves a methodical sequence of steps. We start by meticulously choosing and refining datasets to align with our objectives. Following this, we implement ensemble models to boost the system's threat detection and response capabilities. To ensure clarity in the decision-making process, we incorporate various XAI techniques. The final stage involves evaluating the performance of these techniques using targeted metrics as shown in Fig 3.1. This comprehensive approach is designed to enhance the robustness and transparency of IDS.

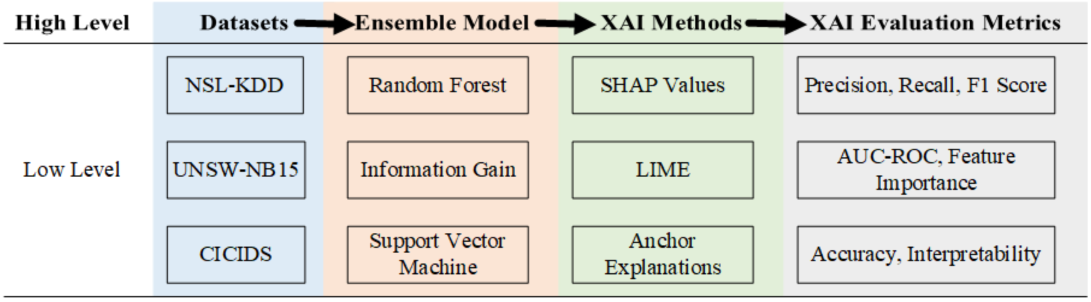

Figure 3.1: Overview of proposed XAI evaluation framework for network intrusion detection

3.1.1 Dataset Selection and Preprocessing

The datasets chosen offer a broad and varied portrayal of network traffic and attack behaviors. For this analysis, we utilized three prominent datasets:

- **NSL-KDD:** an enhanced iteration of the KDD Cup 99 dataset, designed to address the limitations and shortcomings found in the original.
- **UNSW-NB15:** which features a wide range of network activities produced with the IXIA PerfectStorm tool at UNSW Canberras Cyber Range Lab.
- **CICIDS2019:** which provides comprehensive records of contemporary attack scenarios curated by the Canadian Institute for Cybersecurity (CIC).

3.1.2 Data Preprocessing Steps

- **Normalization:** To ensure consistency across different datasets, continuous features were scaled to a standard range, typically from 0 to 1. This normalization process helps make the data comparable.
- **Feature Alignment:** features from each dataset were aligned to establish a unified set of attributes, ensuring that similar features were mapped to a common schema for easier integration and analysis.
- **Handling Missing Values** Missing values were addressed through imputation or removal to maintain data integrity.
- **Label Encoding** Consistent encoding of labels (i.e., attack categories) was ensured across all datasets.

3.1.3 Dataset Merging

To create a comprehensive dataset that encompasses a variety of attack scenarios, we merged the three datasets. The merging process unfolded through a series of critical stages.

- **Feature Selection:** Identified a core set of common features, such as duration, source bytes, destination bytes, and protocol.
- **Schema Conversion:** Converted datasets to a uniform schema based on the selected features, involving renaming features, converting data types, and ensuring consistent units of measurement.
- **Data Concatenation:** Merged the datasets row-wise to form a single comprehensive dataset.
- **Balancing the Dataset:** Applied oversampling or under sampling techniques to mitigate class imbalance and ensure a balanced representation of normal and malicious traffic.

3.1.4 Ensemble Models

To enhance the detection capabilities of IDS, we employed various ensemble models:

- **Random Forest:** Utilizes multiple decision trees to improve accuracy and resilience.
- **Gradient Boosting:** refines successive models by addressing the errors of their predecessors, transforming initially weak learners into robust predictors.
- **Support Vector Machine (SVM):** Uses decision boundaries to separate classes effectively, especially in high-dimensional data.

3.1.5 XAI Methods

To improve the interpretability of the IDS, we integrated several XAI techniques:

- **SHAP Values (SHapley Additive exPlanations):** Provides a unified measure of feature importance, explaining the contribution of each feature to the model's predictions.
- **LIME (Local Interpretable Model-agnostic Explanations):** provides explanations for specific predictions by creating a simpler, interpretable model that approximates the behavior of a more complex one in the vicinity of each prediction.
- **Anchor Explanations:** Provides high-precision rules that sufficiently explain the decision of the model for individual predictions.

3.1.6 XAI Evaluation Metrics

We evaluated the effectiveness and interpretability of the IDS using various metrics:

- **Precision, Recall, F1 Score:** Assess the classification performance of the IDS.
- **AUC-ROC (Area Under the Receiver Operating Characteristic Curve) and Feature Importance:** Evaluate the models ability to distinguish between classes and identify key features influencing the models decisions.
- **Accuracy and Interpretability:** Measure the overall correctness of the model's predictions and the clarity of the explanations provided by the XAI techniques.

3.2 Implementation

The practical application of the proposed approach involved several key steps. First, we focused on data preprocessing and merging, which entailed normalizing the data, aligning features, addressing missing values, encoding labels, selecting relevant features, converting schemas, concatenating datasets, and ensuring balance across the dataset. Next, we trained and evaluated the Random Forest model using both the individual datasets (NSL-KDD, UNSW-NB15, CICIDS2019) and the combined dataset. To enhance our understanding of the models decision-making, we applied XAI techniques such as SHAP and LIME, and assessed the models based on the predefined XAI metrics. This methodical process aimed to improve both the interpretability and effectiveness of the Intrusion Detection System (IDS), providing valuable insights that support security analysts in making well-informed decisions.

Table 3.1: Merged Dataset Features

Feature	Description
dur	Duration of the connection
sbytes	Source bytes
dbytes	Destination bytes
rate	Rate of data transfer
sttl	Source time-to-live
sload	Source load
dload	Destination load
sloss	Source packet loss
dloss	Destination packet loss
sinpkt	Source inter-packet arrival time
smean	Mean of source packets

3.2.1 Merged Dataset

Our merged dataset integrates three prominent datasets: UNSW-NB15, NSL-KDD, and CICIDS2019. By merging these, it offers a solid base for assessing how well machine learning models perform and how interpretable they are when applied to Intrusion Detection Systems (IDS). Fig 3.2 shows the merging process we used to make a new dataset.This combined dataset aims to enhance the accuracy and insightfulness of IDS evaluations.

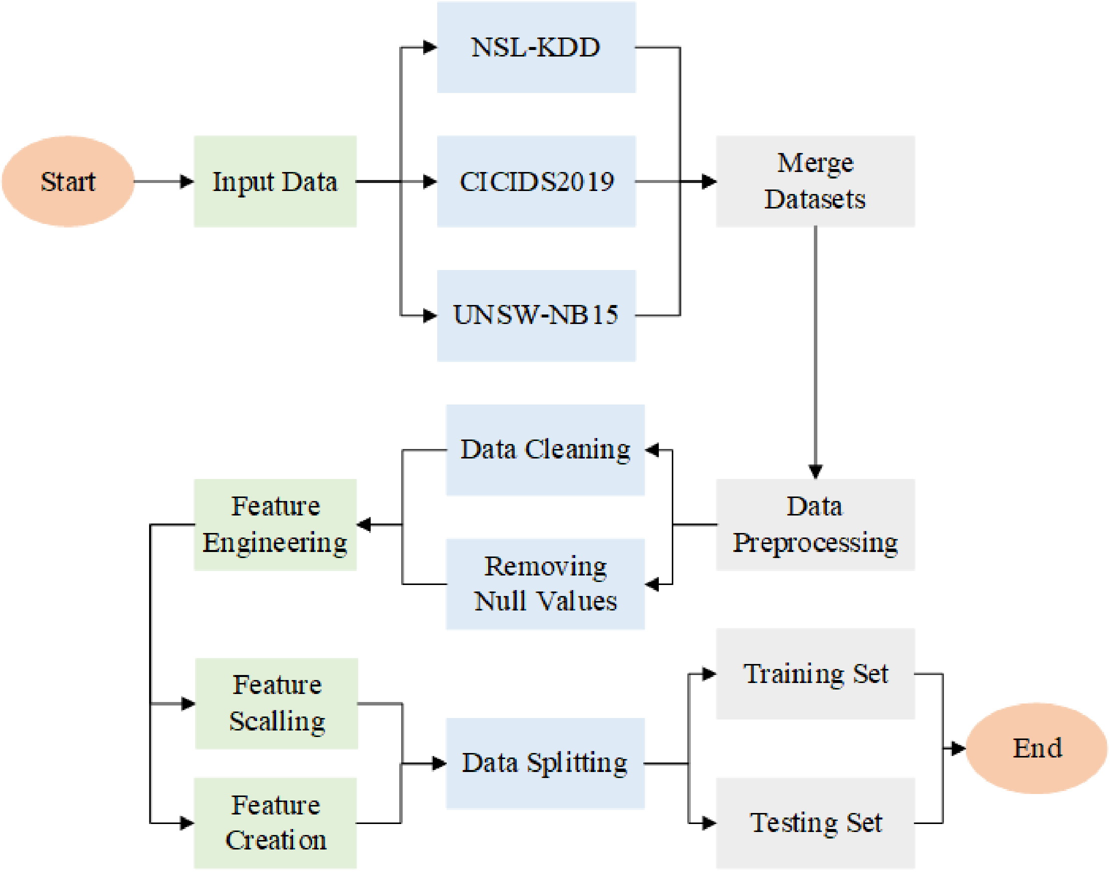

Figure 3.2: Flow chart of the merged dataset process

Table 3.1 below offers a comprehensive overview of the key features incorporated into the combined dataset. The dataset combines a variety of attack patterns and typical network traffic, sourced from UNSW-NB15, NSL-KDD, and CICIDS2019. This collection provides a robust basis for improving the performance and clarity of intrusion detection systems. By applying Random Forest algorithms alongside explainable AI methods like LIME and SHAP, the dataset helps make these systems more effective and their decisions more transparent.

3.2.2 Proposed Model

Fig 3.3 outlines a detailed approach to improving Intrusion Detection Systems (IDS) through the application of Explainable Artificial Intelligence (XAI) methods. It starts with data preprocessing, where various tasks such as cleaning, addressing minority class imbalance, oversampling, one-hot encoding, and normalization are performed on datasets including UNSW-NB15, NSL-KDD, and CICIDS2019. Following this, the data is divided into training and validation sets. Random Forest (RF) is then used for ensemble feature selection to pinpoint the most crucial features. These features are used to train a Multi-Layer Perceptron (MLP) classifier, which is subsequently validated and fine-tuned. The classifier is then tested on a separate dataset to confirm its performance. The final phase involves deploying the trained model for real-time intrusion detection, incorporating XAI techniques to ensure that the decision-making process is clear and understandable, thus boosting the IDSs reliability and trustworthiness.

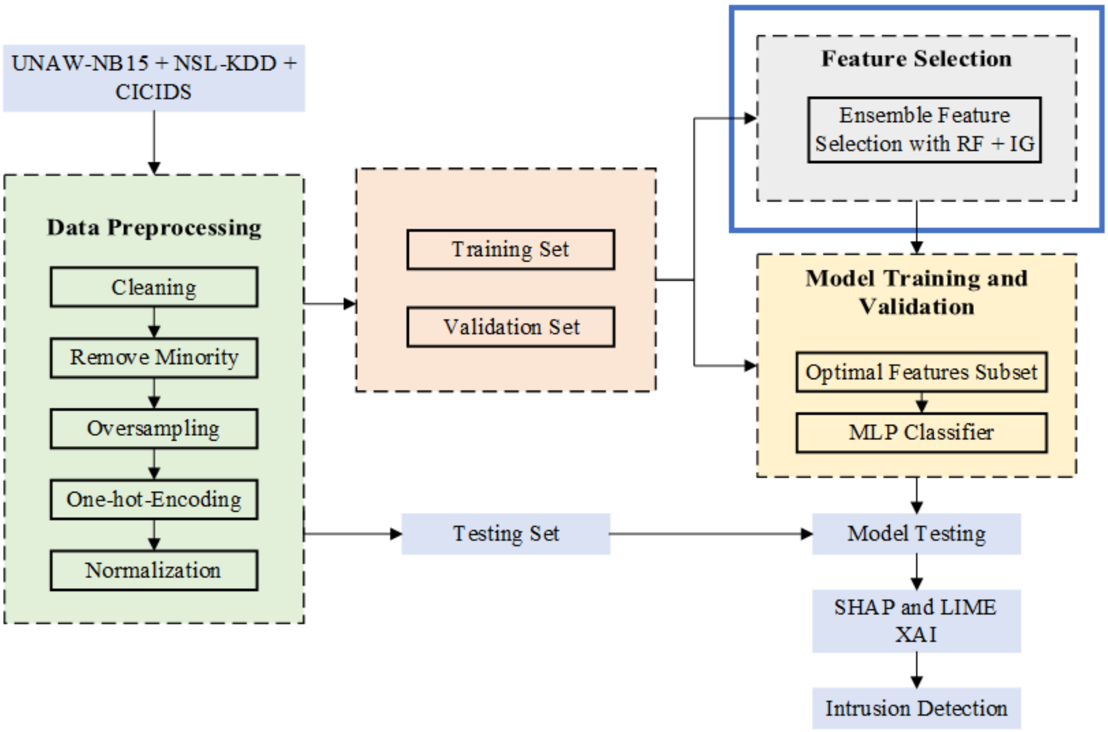

Figure 3.3: Proposed Model

Chapter 4

Results and Discussions

Our research involved experimenting with Random Forest machine learning models on the UNSW-NB15, NSL-KDD, and CICIDS2019 datasets. To add a transparency layer to the results, we employed Explainable AI (XAI) methods, specifically LIME and SHAP. This section presents the performance results of the Random Forest model applied to these datasets, evaluated based on five key metrics: accuracy, precision, recall, F1-score, and ROC AUC score.

Table 4.1: Performance Metrics for Random Forest Model

Dataset	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**
UNSW-NB15	98.85%	0.99	0.97	0.98	0.9992
NSL-KDD	99.99%	1.00	1.00	1.00	0.9999998
CICIDS2019	99.99%	1.00	1.00	1.00	0.9999998

Table 4.1 summarizes the key performance metrics for the Random Forest model across the three datasets, highlighting its high accuracy, precision, recall, F1-score, and ROC AUC score in detecting intrusions.

4.1 Comparative Analysis of Model Performance

This section provides a comparative analysis of model performance on three datasets: UNSW-NB15, CICIDS2017, and NSL-KDD.

4.1.1 UNSW-NB15 Dataset

Table 4.2 presents the comparative analysis of model performance on the UNSW-NB15 dataset. The accuracy of different models is compared, highlighting the effectiveness of the Random Forest model used in this study.

Table 4.2: Comparative Analysis of Model Performance on UNSW-NB15 Dataset

Study/Model	Dataset	Accuracy	Algorithm
Proposed Approach	UNSW-NB15	98.89%	Random Forest model with Information Gain
[48]	UNSW-NB15	91.7%	Hybrid feature selection with MLP
[49]	UNSW-NB15	96.15%	DeeRaI with CuI for DoS detection
[50]	UNSW-NB15	97.37%	Random Forest for multi-class classification
[51]	UNSW-NB15	95.2%	Two-level feature selection with Decision Tree

4.1.2 CICIDS2017 Dataset

Table 4.3 illustrates the comparative analysis of model performance on the CICIDS2017 dataset. The results emphasize the high accuracy achieved by the Random Forest model implemented in this study.

Table 4.3: Comparative Analysis of Model Performance on CICIDS2017 Dataset

Study/Model	Dataset	Accuracy	Algorithm
Proposed Approach	CICIDS2017	99.9956%	Random Forest model with Information Gain
[52]	CICIDS2017	97.02%	Gray Wolf Optimization and Particle Swarm Optimization with Random Forest
[53]	CICIDS2017	98.1%	K-means clustering and Decision Tree for IoT IDS
[54]	CICIDS2017	97.02%	Ensemble-based approach with ANOVA F-test and Kalman filter
[55]	CICIDS2017	96%	Deep Belief Network (DBN) model for IDS

4.1.3 NSL-KDD Dataset

Table 4.4 shows the comparative analysis of model performance on the NSL-KDD dataset. The Random Forest model again demonstrates superior performance compared to other models.

Table 4.4: Comparative Analysis of Model Performance on NSL-KDD Dataset

Study/Model	Dataset	Accuracy	Algorithm
Proposed Approach	NSL-KDD	99.75%	Random Forest model with Information Gain
[56]	NSL-KDD	97.5%	ANN model with PCA and hyperparameter tuning
[57]	NSL-KDD	97.7%	Deep learning-based approach with LSTM, MLP, and SVM
[58]	NSL-KDD	93.88%	LSTM and GRU models with data resampling
[59]	NSL-KDD	97%	SVM and Naive Bayes for classification

Table 4.5: Merged Table for LIME Explanations Across Three Datasets

Dataset	Instance	Intercept	Prediction_local	Right
UNSW-NB15	Anomalous	0.6495	0.0506	0.0
	1	0.7152	0.0380	0.0
	2	0.4147	0.6949	1.0
	3	0.4128	0.6951	0.98
	4	0.7185	0.0343	0.0
CICIDS2019	1	0.2148	0.00091509	1.0
	2	0.2082	0.00060716	1.0
	3	0.0619	0.00042975	1.0
	4	0.0433	1.4365	1.0
	5	0.0656	-0.00027567	1.0
NSL-KDD	1	0.2405	0.31724908	1.0
	2	0.2734	0.20864928	0.0
	3	0.3204	0.02192315	0.0
	4	0.0697	0.63537629	1.0
	5	0.3940	-0.09125303	0.0

4.2 Interpretability with LIME and SHAP

The use of LIME and SHAP for model interpretability provided valuable insights into the model's decision-making process. As shown in Table 4.5, Across all three datasets, certain features consistently emerged as important, demonstrating the model's reliance on specific attributes to make accurate predictions. These interpretability techniques not only enhance the understanding of the model's behavior but also build trust in its predictions by elucidating the rationale behind its decisions. Fig 4.1, 4.2 and 4.3 depicts a SHAP summary plot for feature importance in all datasets. These plots helps in understanding importance of different features across all three datasets, respectively CICIDS2019, UNSW-NB15 and NSL-KDD.

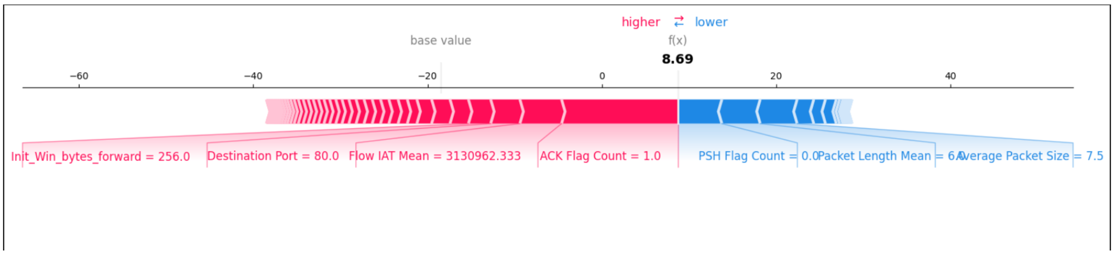

Figure 4.1: SHAP Summary Plot for Feature Importance in the CICIDS2019 Dataset

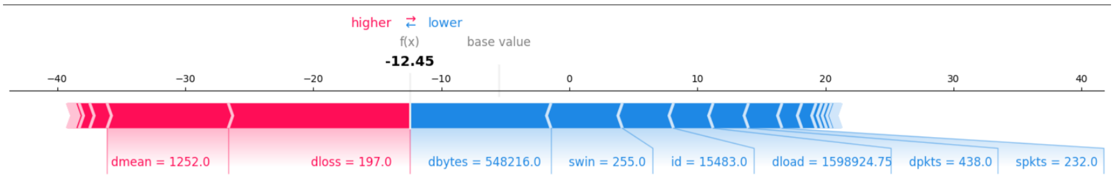

Figure 4.2: SHAP Summary Plot for Feature Importance in the UNSW-NB15 Dataset

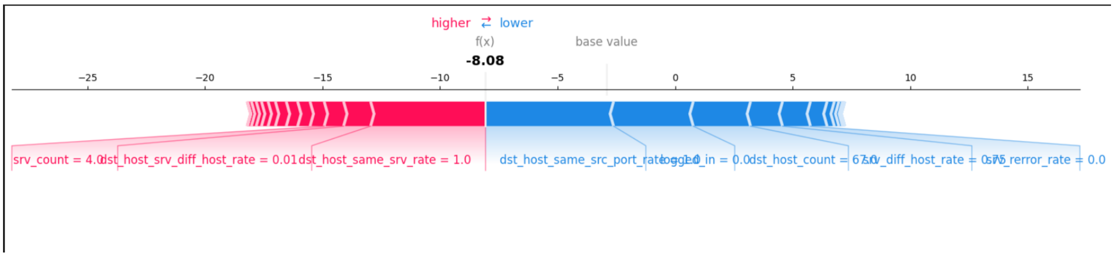

Figure 4.3: SHAP Summary Plot for Feature Importance in the NSL-KDD Dataset

Chapter 5

Conclusions and Future Work

As shown in Table 4.2 , 4.3 and 4.4, the comparative analysis reveals that the Random Forest model, when paired with Information Gain for feature selection, stands out with the highest accuracy across all three datasets. This underscores its robustness and efficiency in detecting intrusions. This study advanced Intrusion Detection Systems (IDS) by incorporating Random Forest models alongside Explainable Artificial Intelligence (XAI) techniques. When tested on the UNSW-NB15, NSL-KDD, and CICIDS2019 datasets, the Random Forest model achieved impressive accuracy rates of 98.85%, 99.99%, and 99.99%, respectively. It also showed near-perfect precision, recall, F1-scores, and high ROC AUC values, which highlight its strong detection capabilities.

Table 4.5 depicts the Lime explanations across three datasets used in this study. The integration of XAI methods, particularly LIME and SHAP, added transparency to the models decision-making process, enhancing both interpretability and trust for security analysts. These methods pinpointed the most critical features, helping to clarify the models actions. Compared to other recent studies, our approach proved superior, demonstrating that the Random Forest model outperformed other algorithms. This research not only offers a robust and interpretable IDS framework but also sets a foundation for future exploration of these techniques across different models and datasets, aiming to boost interpretability and trustworthiness in IDS further.

Bibliography

[1] Li-Hsiang Shen, Kai-Ten Feng, and Lajos Hanzo. "Five facets of 6G: Research challenges and opportunities". In: *ACM Computing Surveys* 55.11 (2023), pp. 1–39.

[2] S Ntalampiras, G Misuraca, and P Rossel. "Artificial Intelligence and Cybersecurity Research". In: *G. Misuraca, P. Rossel* (2023).

[3] Dongqi Han et al. "Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors". In: *IEEE Journal on Selected Areas in Communications* 39.8 (2021), pp. 2632–2647.

[4] Mustapha Belouch, Salah El Hadaj, and Mohamed Idhammad. "Performance evaluation of intrusion detection based on machine learning using Apache Spark". In: *Procedia Computer Science* 127 (2018), pp. 1–6.

[5] T Saranya et al. "Performance analysis of machine learning algorithms in intrusion detection system: A review". In: *Procedia Computer Science* 171 (2020), pp. 1251–1260.

[6] Emad E Abdallah, Ahmed Fawzi Otoom, et al. "Intrusion detection systems using supervised machine learning techniques: a survey". In: *Procedia Computer Science* 201 (2022), pp. 205–212.

[7] Mahdi Zamani and Mahnush Movahedi. "Machine learning techniques for intrusion detection". In: *arXiv preprint arXiv:1312.2177* (2013).

[8] Ahmed S Alzahrani et al. "A novel method for feature learning and network intrusion classification". In: *Alexandria Engineering Journal* 59.3 (2020), pp. 1159–1169.

[9] Monire Norouzi et al. “A Hybrid Genetic Algorithm-Based Random Forest Model for Intrusion Detection Approach in Internet of Medical Things”. In: *Applied Sciences* 13.20 (2023), p. 11145.

[10] Munish Saran, Rajan Kumar Yadav, and Upendra Nath Tripathi. “Machine learning based security for cloud computing: A survey”. In: *Int J Appl Eng Res* 17.4 (2022), pp. 332–337.

[11] Yazeed Alotaibi and Mohammad Ilyas. “Ensemble-learning framework for intrusion detection to enhance internet of things devices security”. In: *Sensors* 23.12 (2023), p. 5568.

[12] Albara Awajan. “A novel deep learning-based intrusion detection system for IOT networks”. In: *Computers* 12.2 (2023), p. 34.

[13] Heba Ezzat Ibrahim, Sherif M Badr, and Mohamed A Shaheen. “Adaptive layered approach using machine learning techniques with gain ratio for intrusion detection systems”. In: *arXiv preprint arXiv:1210.7650* (2012).

[14] Syed Ali Raza Shah and Biju Issac. “Performance comparison of intrusion detection systems and application of machine learning to Snort system”. In: *Future Generation Computer Systems* 80 (2018), pp. 157–170.

[15] Kathryn-Ann Tait et al. “Intrusion detection using machine learning techniques: an experimental comparison”. In: *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*. IEEE. 2021, pp. 1–10.

[16] Chih-Fong Tsai et al. “Intrusion detection by machine learning: A review”. In: *expert systems with applications* 36.10 (2009), pp. 11994–12000.

[17] Chongzhen Zhang et al. “A novel framework design of network intrusion detection based on machine learning techniques”. In: *Security and Communication Networks* 2021.1 (2021), p. 6610675.

[18] Thomas Rincy N and Roopam Gupta. “Design and development of an efficient network intrusion detection system using machine learning techniques”. In: *Wireless Communications and Mobile Computing* 2021.1 (2021), p. 9974270.

[19] Mohammed F Suleiman and Biju Issac. “Performance comparison of intrusion detection machine learning classifiers on benchmark and new datasets”. In: *2018*

28th International Conference on Computer Theory and Applications (ICCTA). IEEE. 2018, pp. 19–23.

[20] Salima Omar, Asri Ngadi, and Hamid H Jebur. “An adaptive intrusion detection model based on machine learning techniques”. In: *International Journal of Computer Applications* 70.7 (2013).

[21] Mathappan Nivaashini et al. “Effective Feature Selection for Hybrid Wireless IoT Network Intrusion Detection Systems Using Machine Learning Techniques.” In: *Ad Hoc Sens. Wirel. Networks* 49.3-4 (2021), pp. 175–206.

[22] Karan Bajaj and Amit Arora. “Improving the intrusion detection using discriminative machine learning approach and improve the time complexity by data mining feature selection methods”. In: *International Journal of Computer Applications* 76.1 (2013), pp. 5–11.

[23] Hongyu Liu and Bo Lang. “Machine learning and deep learning methods for intrusion detection systems: A survey”. In: *applied sciences* 9.20 (2019), p. 4396.

[24] Leandros Maglaras. “Intrusion detection in scada systems using machine learning techniques”. PhD thesis. University of Huddersfield, 2018.

[25] Raja Azlina Raja Mahmood, AmirHossien Abdi, and Masnida Hussin. “Performance evaluation of intrusion detection system using selected features and machine learning classifiers”. In: *Baghdad Science Journal* 18.2 (Suppl.) (2021), pp. 0884–0884.

[26] Musbau Dogo Abdulrahaman and John K Alhassan. “Ensemble learning approach for the enhancement of performance of intrusion detection system”. In: *International Conference on Information and Communication Technology and its Applications (ICTA 2018).* 2018, pp. 1–8.

[27] Ngoc Tu Pham et al. “Improving performance of intrusion detection system using ensemble methods and feature selection”. In: *Proceedings of the Australasian computer science week multiconference.* 2018, pp. 1–6.

[28] Chie-Hong Lee et al. “Machine learning based network intrusion detection”. In: *2017 2nd IEEE International conference on computational intelligence and applications (ICCIA).* IEEE. 2017, pp. 79–83.

[29] DP Gaikwad and Ravindra C Thool. “Intrusion detection system using bagging ensemble method of machine learning”. In: *2015 international conference on computing communication control and automation*. IEEE. 2015, pp. 291–295.

[30] Shadman Latif et al. “Investigation of machine learning algorithms for network intrusion detection”. In: *International Journal of Information Engineering and Electronic Business* 15.2 (2022), p. 1.

[31] Adnan Helmi Azizan et al. “A machine learning approach for improving the performance of network intrusion detection systems”. In: *Annals of Emerging Technologies in Computing (AETiC)* 5.5 (2021), pp. 201–208.

[32] Iftikhar Ahmad et al. “Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection”. In: *IEEE access* 6 (2018), pp. 33789–33795.

[33] Sharfuddin Khan, E Sivaraman, and Prasad B Honnavalli. “Performance evaluation of advanced machine learning algorithms for network intrusion detection system”. In: *Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India*. Springer. 2020, pp. 51–59.

[34] Hatim Mohamad Tahir et al. “Hybrid machine learning technique for intrusion detection system”. In: (2015).

[35] Nada Aboueata et al. “Supervised machine learning techniques for efficient network intrusion detection”. In: *2019 28th international conference on computer communication and networks (ICCCN)*. IEEE. 2019, pp. 1–8.

[36] UM Learning. “SS symmetry effective intrusion detection system to secure data in cloud”. In: (2021).

[37] Nancy Awadallah Awad. “Enhancing Network Intrusion Detection Model Using Machine Learning Algorithms.” In: *Computers, Materials & Continua* 67.1 (2021).

[38] Ployphan Sornsuwit and Saichon Jaiyen. “A new hybrid machine learning for cybersecurity threat detection based on adaptive boosting”. In: *Applied Artificial Intelligence* 33.5 (2019), pp. 462–482.

[39] Nahida Islam et al. “Towards Machine Learning Based Intrusion Detection in IoT Networks.” In: *Computers, Materials & Continua* 69.2 (2021).

[40] Imran, Faisal Jamil, and Dohyeun Kim. “An ensemble of prediction and learning mechanism for improving accuracy of anomaly detection in network intrusion environments”. In: *Sustainability* 13.18 (2021), p. 10057.

[41] Jose L Hernandez-Ramos et al. “Intrusion Detection based on Federated Learning: a systematic review”. In: *arXiv preprint arXiv:2308.09522* (2023).

[42] Mousa Alalhareth and Sung-Chul Hong. “An improved mutual information feature selection technique for intrusion detection systems in the Internet of Medical Things”. In: *Sensors* 23.10 (2023), p. 4971.

[43] Deborah Oladimeji. “An intrusion detection system for internet of medical things”. In: (2021).

[44] Ye-Eun Kim, Yea-Sul Kim, and Hwankuk Kim. “Effective feature selection methods to detect IoT DDoS attack in 5G core network”. In: *Sensors* 22.10 (2022), p. 3819.

[45] Hongyu Liu and Bo Lang. “Machine learning and deep learning methods for intrusion detection systems: A survey”. In: *applied sciences* 9.20 (2019), p. 4396.

[46] Soulaiman Moualla, Khaldoun Khorzom, and Assef Jafar. “Improving the Performance of Machine Learning-Based Network Intrusion Detection Systems on the UNSW-NB15 Dataset”. In: *Computational Intelligence and Neuroscience* 2021.1 (2021), p. 5557577.

[47] Tuba Aytaç, MUHAMMED AYDIN, and ABDÜL ZAM. “Detection DDOS attacks using machine learning methods”. In: *Electrica* 20.2 (2020).

[48] Yuhua Yin et al. “IGRF-RFE: a hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset”. In: *Journal of Big Data* 10.1 (2023), p. 15.

[49] Shweta More et al. “Enhanced Intrusion Detection Systems Performance with UNSW-NB15 Data Analysis”. In: *Algorithms* 17.2 (2024), p. 64.

[50] Md Alamin Talukder et al. “Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction”. In: *Journal of big data* 11.1 (2024), p. 33.

[51] Aditi Roy and Khundrakpam Johnson Singh. “Multi-classification of unsw-nb15 dataset for network anomaly detection system”. In: *Proceedings of International Conference on Communication and Computational Technologies: ICCCT-2019*. Springer. 2021, pp. 429–451.

[52] Omar Elnakib et al. “EIDM: deep learning model for IoT intrusion detection systems”. In: *The Journal of Supercomputing* 79.12 (2023), pp. 13241–13261.

[53] B Anbarasu and I Sumaiya Thaseen. “Anomaly Detection Using Feature Selection and Ensemble of Machine Learning Models”. In: *Computational Methods and Data Engineering: Proceedings of ICCMDE 2021*. Springer, 2022, pp. 215–229.

[54] Emad Ul Haq Qazi, Muhammad Hamza Faheem, and Tanveer Zia. “HDLNIDS: hybrid deep-learning-based network intrusion detection system”. In: *Applied Sciences* 13.8 (2023), p. 4921.

[55] Komal Singh Gill and Arwinder Dhillon. “A hybrid machine learning framework for intrusion detection system in smart cities”. In: *Evolving Systems* (2024), pp. 1–15.

[56] Mohammed Zakariah et al. “Intrusion Detection System with Customized Machine Learning Techniques for NSL-KDD Dataset.” In: *Computers, Materials & Continua* 77.3 (2023).

[57] Aysha Bibi et al. “A hypertuned lightweight and scalable LSTM model for hybrid network intrusion detection”. In: *Technologies* 11.5 (2023), p. 121.

[58] Ahmed Abdelkhalek and Maggie Mashaly. “Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning”. In: *The journal of Supercomputing* 79.10 (2023), pp. 10611–10644.

[59] Mansi Bhavsar et al. “Anomaly-based intrusion detection system for IoT application”. In: *Discover Internet of things* 3.1 (2023), p. 5.