National College of Ireland

Perfecting Intrusion Detection System using Machine Learning Algorithm

MSc Research Project
Data Analytics

Yogeshwar Bodicherla Rambob
Student ID: x22176322

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Yogeshwar Bodicherla Rambob
Student ID:	x22176322
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	20/12/2023
Project Title:	Perfecting Intrusion Detection System using Machine Learning Algorithm
Word Count:	5942
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Yogeshwar Bodicherla Rambob
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	✓
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	✓
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	✓

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Perfecting Intrusion Detection System using Machine Learning Algorithm

Yogeshwar Bodicherla Rambob
22176322

Abstract

The Intrusion Detection System is used in many IT industries now a days. As it has been becoming a stronger and complete tool for detecting intruders, whoever tries to enter the system. In this study, had been still trying to make the Intrusion system stronger and more efficient by finding out the default issues in it and trying to mitigate it by using five algorithms against eight types of attacks like User Datagram Protocol, Domain Name System, Lightweight Directory Access Protocol, Network Basic Input/Output System, Simple Network Management Protocol, Network Time Protocol, and Gaussian Naive Bayes. Perform the evaluation for all the samples as mentioned. In this thesis, predicted three best algorithms from that and filtered the best algorithm from the three. On top of chosen one of the best, the Random Forest algorithm as it was occurring in multiple results and its performing is faster and better results compared to the others.

1 Introduction

The computers play a very strong role in everyone's life. The dependency on the internet is also increasing daily day by day. And the cybercrime is parallelly increasing day by day. To overcomes this kind of situations. Everyone has starting using the antivirus in their systems. But still every day, we can find new kind of malwares in the system and its coming through the internet. Need of new technology was required to avoid these kinds of cybercrimes. Intrusion Detection System paved a new way to the decrease the crimes. As combinational friend, we also got the machine learning technology. Creating a combination of both helped us to learning about the new kinds of malicious virus and the activities which are about to happen from that those trojans. It also started to recognize the new kind of trojans, where it created a new technology to reduce the system attacks. The CICIDS was also one of the greatest research centre started to provide the captured samples of the malware for the real time prediction and to develop more advanced system for the future generations to be safe.

1.1 Motivation and Background

The critical need to increase cyber security in response to the rising and growing cyber threats motivates the choice of this work. Most cyber-attacks, including a wide range of incidents, such as data breaches and attacks supported by governments, have changed people, organizations, and governmental entities. Traditional rule-based Intrusion Detection Systems (IDS) have inherent limitations in their ability to adapt to emerging attack

vectors effectively and often generate false positive alerts. This study aims to fill the existing research difference by exploring the potential of Machine Learning Algorithms in enhancing the accuracy and effectiveness of Intrusion Detection Systems (IDS). This research supplies benefits to both academic and corporate sectors by contributing to the advancement of knowledge and supplying practical solutions for organizations to enhance their network security.

The research has significant importance because of its ability to enhance network security and efficiently address cyber-attacks. This research looks to improve the capabilities of Intrusion Detection Systems (IDS) by perfecting features, data processing methods, and Machine Learning algorithms. The aim is to enable IDS not only to find but also to neutralize a wide range of cyber-attacks.

1.2 Research Question and Objectives

This research on the integration of Machine Learning (ML) algorithms into Intrusion Detection Systems (IDS) holds significant implications for both academics and industry.In the industry, the significance lies in the strengthened security postures and improved operational efficiency. ML-based IDS not only enhances the precision of threat detection but also reduces false positives, allowing security teams to focus on critical tasks. The automation brought about by ML algorithms enables timely responses to security incidents, minimizing potential damages.

RQ:*How can the integration of Machine Learning Algorithms enhance the precision and effectiveness of Intrusion Detection Systems to strengthen network security, and what value does this hold for academics and industry in countering the evolving landscape of cyber threats?*

Implementing Machine Learning within Intrusion Detection Systems substantially enhances the accuracy of network security. This integration facilitates ongoing scholarly investigations into cyber threats and resilient security solutions. It improves the cyber resilience of industries by detecting and responding to threats autonomously and in real-time. Proactive defense measures are of utmost importance in mitigating the ever-changing landscape of cyber threats, providing advantages for scholarly investigation and industry-wide application.

Contribution - The dataset known as CICIDS 2019 supplies a significant addition to cyber security because it supplies a comprehensive and varied assortment of network traffic data that closely emulates real-world situations. The dataset presented in this study includes a diverse array of cyber-attack situations. It offers ground truth labels that enable correct evaluation. It includes normal and malicious network traffic, aiding the creation and testing of intrusion detection systems (IDS) and machine learning algorithms. The CICIDS 2019 dataset is significant in research, as it supplies a comprehensive collection of data across several classes. This dataset is a vital benchmark for researchers, allowing them to assess and enhance the effectiveness of different intrusion detection methods. This resource's practical significance and possible real-world uses contribute significantly to the advancement of intrusion detection and cyber security research.

ID	Objective
Obj1	Literature Review: Assess CICIDS 2019 dataset for IDS model training and testing, and review various research methodologies.
Obj2	Methodology Design: Tailor machine learning algorithms and data processing for IDS applications.
Obj3	Algorithm Assessment: Evaluate the efficacy of machine learning in IDS for accuracy and false positive reduction.
Obj4	Practical Recommendations: Offer actionable insights for enhancing network security in business and academia.
Obj5	Dataset Evaluation for Protocols: Assess dataset suitability for various network protocols like SSDP, NTP, SNMP, etc.
Obj6	Model Evaluation: Analyze model performance on precision, recall, F1 score, and accuracy, including a results summary table.
Obj7	Model Comparison: Contrast the developed model with those from literature reviews.
Obj8	Model Testing: Evaluate models on training and testing sets, including SMOTE for class balancing.

The succeeding sections of the report are structured as follows. The "Related Work" section offers a full analysis of earlier investigations performed in intrusion detection and machine learning. The "Research Methodology" presents a complete explanation of the chosen research strategy and the experimental setting used in the study. The "Design Specification" document defines the essential system requirements, exactly as the "Implementation" section explains the actual coding elements. The process of "evaluation" involves the presentation of results and performance measures. The "Conclusion and Future Work" merges the main results and supplies new avenues for further investigation in intrusion detection and network security. Using an organized plan ensures a complete and systematic presentation of the research investigation.

2 Literature Review on Intrusion Detection Systems with Machine Learning: Techniques and Applications

Machine learning (Machine Learning) techniques have been widely applied in various domains, including cybersecurity, where they have shown significant potential in enhancing intrusion detection systems (IDS). Intrusion detection plays a crucial role in identifying and mitigating malicious activities in network and IoT environments. This literature review synthesizes findings from recent studies focusing on the application of Machine Learning and Deep Learning in intrusion detection, highlighting the various techniques, datasets, and challenges encountered.

2.1 Machine Learning, Internet of Things Security and Network Intrusion Detection

The Author Saini et al. proposed a Machine Learning-based approach to mitigate security threats in IoT environments, highlighting the criticality of securing IoT devices (Saini et al. 2023). Similarly, Fatani et al. developed an IoT intrusion detection system using deep learning and an enhanced transient search optimization algorithm, emphasizing the effectiveness of Deep Learning in hanDeep Learninging complex and high-dimensional data prevalent in IoT networks (Fatani et al. 2021). Yadav et al. introduced an unsupervised federated learning-based IoT intrusion detection method, highlighting the potential of federated learning in distributed and privacy-sensitive IoT scenarios (Yadav et al. 2021). The Author Singhal et al. presented a hybrid Machine Learning and data mining approach for network intrusion detection, highlighting the constructive collaboration between different computational intelligence methods in enhancing detection performance (Singhal et al. 2021). Bharati and Tamane (Bharati & Tamane 2020) utilized deep and Machine Learning frameworks on the CICIDS2018 dataset with cloud computing, demonstrating the scalability and efficiency of cloud-based IDS solutions. Yin et al. (Yin et al. 2023), Panwar et al. (Panwar et al. 2019), Jairu and Mailewa (Jairu & Mailewa 2022), and Kurniabudi et al. (Kurniabudi et al. 2020) conducted extensive evaluations and feature analyses on the CICIDS-2017 dataset, providing valuable insights into the effectiveness of various Machine Learning algorithms in network intrusion detection.

2.2 Specialized Applications, Techniques, Evaluation and Performance Improvement

The Author Natarajan (Natarajan 2022) focused on detecting man-in-the-midDeep Learninge attacks using Machine Learning, addressing a specific type of network intrusion. Balyan et al. (Balyan et al. 2022) applied Machine Learning-based IDS to healthcare data, emphasizing the importance of securing sensitive health information. Wankhede and Kshirsagar (Wankhede & Kshirsagar 2018) and Rashid et al. (Rashid et al. 2020) addressed the detection of DoS attacks and phishing attempts, respectively, highlighting the versatility of Machine Learning in tackling different cyber threats. Several studies aimed at evaluating and improving the performance of Machine Learning-based IDS. Panwar et al. (Panwar et al. 2022) and (Ahanger et al. 2021) conducted extensive evaluations using the CICIDS-2017 dataset, exploring feature selection and various Machine Learning algorithms. AguilonGost et al. (AguilonGost et al. 2022) developed an IDS for both known and unknown anomalies, highlighting the challenge of zero-day attacks. Fosic et al. (Fosic et al. 2022) and Yedukondalu et al. (Yedukondalu et al. 2021) focused on network traffic verification and the development of a comprehensive IDS framework, respectively.

2.3 Hybrid Approaches, Emerging Trends and Specific Attacks Detection

The integration of various Machine Learning techniques has been a focal point of recent research. Paul et al. (PAUL et al. 2023) presented a hybrid IDS for detecting cross-layer DoS attacks in IoT, highlighting the trend towards integrated and multi-layered security solutions. Sharma et al. (Sharma et al. 2020), Abraham and Bindu (Abraham & Bindu 2021), and Samawi et al. (Samawi et al. 2022) delved into the application of hybrid

Machine Learning and Deep Learning approaches, highlighting the ongoing evolution and adaptation of Machine Learning in intrusion detection. Studies like Kristyanto et al. (Kristyanto et al. 2022) and Aljohani and Bushnag (Aljohani & Bushnag 2021) addressed specific attack vectors such as SSH brute force attacks and intrusion detection in local area networks, respectively. These works underscore the necessity of tailored machine-learning solutions for diverse attack scenarios.

2.4 Conclusion and contributes

This report builds upon and extends the existing body of research on machine learning-based intrusion detection systems by addressing several key limitations identified in the reviewed papers. Specifically, our work offers a novel solution that not only demonstrates strong performance in controlled evaluations but also places a significant emphasis on the practical implementation and deployment of the intrusion detection system in real-world settings. We address privacy concerns through advanced anonymization techniques and compliance with data protection regulations, ensuring the secure hanDeep Learninging of sensitive information. Moreover, our approach introduces innovative strategies to effectively detect and mitigate zero-day attacks, a persistent challenge in intrusion detection. By prioritizing scalability and efficiency, we strive to provide a comprehensive IDS framework that is not only robust and accurate but also adaptable to the demands of large-scale network and IoT environments, thus offering a practical and effective solution to the security challenges of today's interconnected digital world.

In conclusion, the reviewed papers collectively contribute to the field of machine learning-based intrusion detection systems by addressing various application domains and evaluation methodologies. They demonstrate the versatility of machine learning and deep learning techniques in enhancing security across the Internet of things, networks, and specialized contexts. However, challenges such as real-world deployment, privacy, and scalability remain areas of concern. While the studies provide valuable insights into performance, there is a need for future research to focus on practical implementation, privacy preservation, and innovative approaches to tackle zero-day attacks. Balancing the strengths of Machine Learning-based IDS with these considerations will be crucial for their continued evolution and effectiveness in safeguarding digital ecosystems.

3 Scientific Methodology Approach Used and Design Specification

A quantitative research design is employed, utilizing an empirical approach to systematically analyze network traffic data and assess the performance of different machine learning algorithms within the Intrusion detection system. The research follows an experimental design, facilitating a controlled environment to simulate network conditions, introduce cyber threats, and evaluate the intrusion detection system response.

3.1 Intrusion Detection Methodology Approach

An exhaustive description of the network security architecture outlines its essential components in Figure 1. The firewall is comprehensively described in the beginning, en-

compassing details such as its configuration, rule sets, and the criteria employed to differentiate between authorized and unauthorized packets. This component serves a pivotal function in impeding the interception of network transmissions, permitting passage exclusively to those authenticated. The integration described is critical for detecting potentially malicious communications that aim to evade the firewall. Furthermore, the significance of the Database/Management System is emphasized by its crucial position in the network architecture. This component ensures the creation of an exhaustive log of security events. It provides that the security configurations of the network are robust and in line with the intended security goals of the network. Finally, it is worth noting that the network security architecture comprises various network devices, such as email servers, web servers, and laptops. In the network framework, distinct roles and responsibilities are allocated to each of these devices, enhancing the system's overall security and functionality.

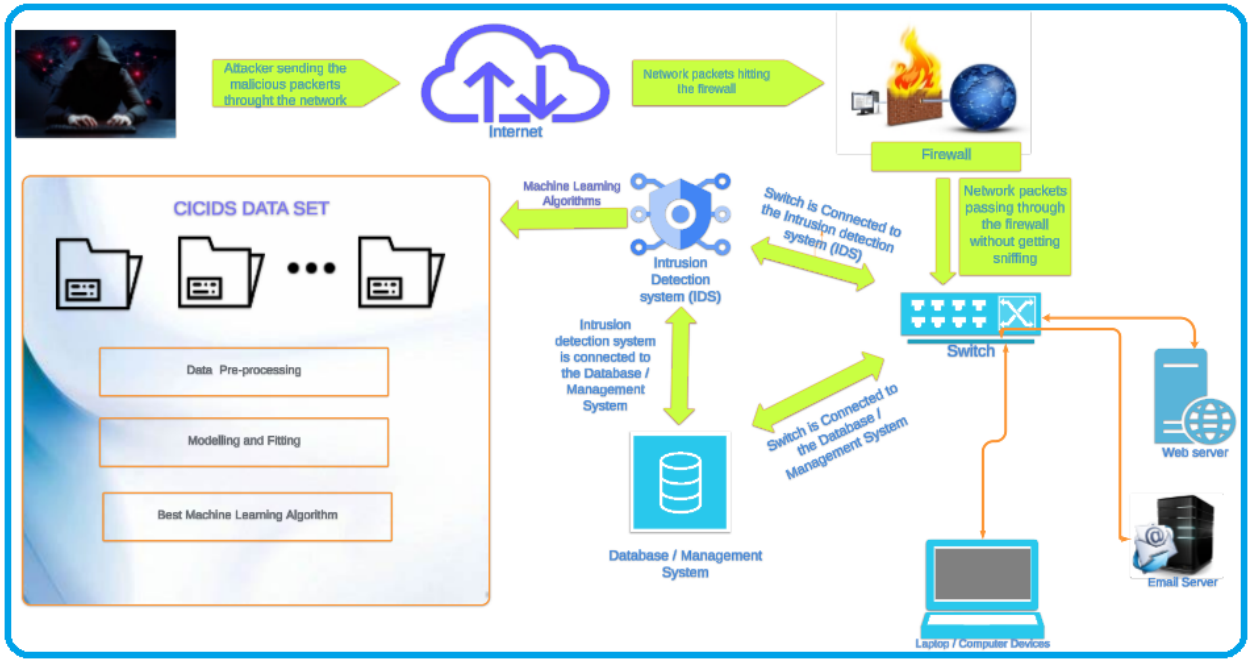

Figure 1: Intrusion Detection System Model

3.1.1 Role of Machine Learning in Strengthening Intrusion Detection Systems (IDS)

Machine learning plays a critical role in Intrusion Detection Systems (IDS) by automating a sequence of critical processes involving detecting and managing security threats. In Figure 2, the procedure commences with data collection, during which an intrusion detection system amasses a vast dataset about the activities of networks and systems. Before analysis, the raw data is preprocessed to remove extraneous noise and superfluous information. Feature engineering methods may be employed to extract noteworthy attributes. Following the preprocessing of the data, machine learning models are subsequently trained. These models may take diverse forms, including anomaly detection, signature-based, and behavioral models. Anomaly detection methods identify patterns that deviate from the norm, signature-based algorithms generate alarms for deviations from known attack patterns, and behavioral models recognize typical behavior patterns.

Upon detecting potential threats, the ML model identifies patterns or characteristics associated with recognized assaults or anomalies, generating alerts or responses. Constant improvement is required, and the model is consistently refreshed with new data to account for emerging hazards.

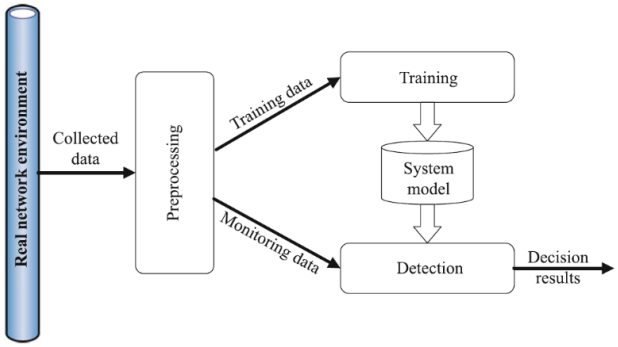

Figure 2: Machine Learning Implementation

3.1.2 Process Flow of Machine Learning in Intrusion Detection Systems

The preliminary phase consists of importing and preparing the dataset, addressing obstacles such as mixed categories and absent information in Figure 3.

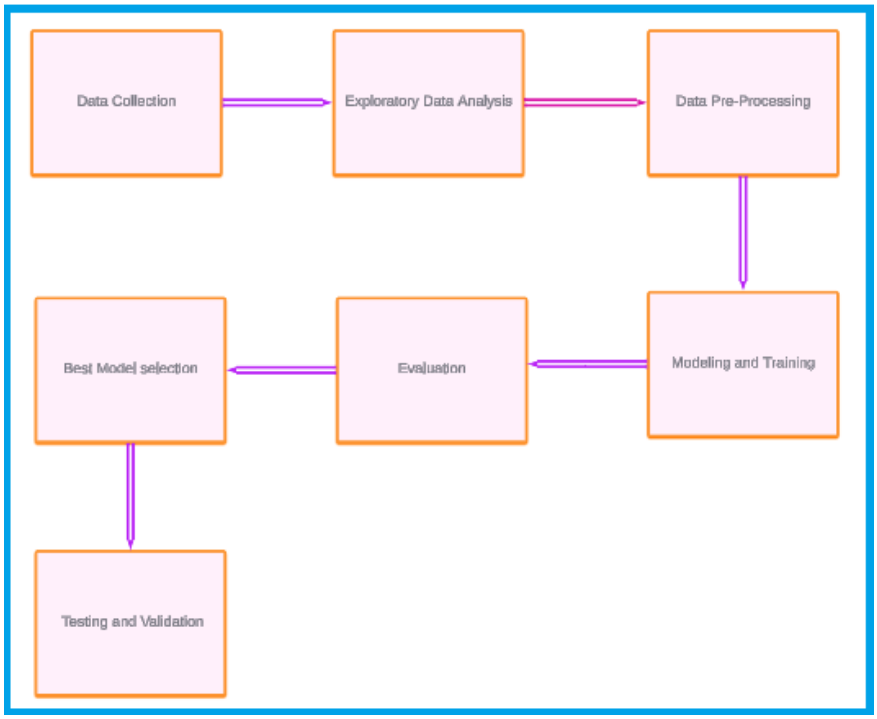

Figure 3: Methodology Flow

Data is subsequently partitioned into training and testing sets, and class imbalance is managed by utilizing the Random Under Sampler and Synthetic Minority Oversampling Technique (SMOTE). The fundamental basis of the architecture is composed of constructing and assessing machine learning models, which include the following: Gaussian

Naive Bayes, Random Forest, Logistic Regression, Decision Tree, and k-nearest Neighbours (KNN). The assessment centers on the recall metric, with particular attention paid to the precise detection of authentic affirmative instances in the context of intrusion detection. An additional element has been incorporated into the design to facilitate the comparison and evaluation of algorithms without the need for resampling methodologies. This feature offers greater flexibility in the approaches taken for model assessment. The ultimate component comprises critical metrics such as precision, recall, F1 score, and accuracy, providing a holistic assessment of the performance of each model. In Figure 4, the design process is specifically designed to manage data sets that contain an uneven distribution of classes effectively. It emphasizes the criticality of choosing suitable evaluation criteria when developing intrusion detection models.

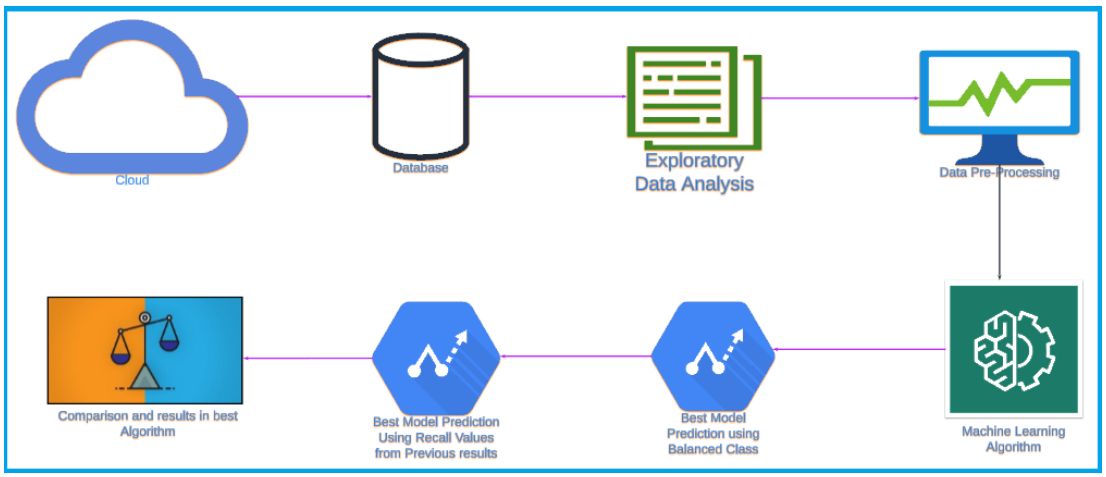

Figure 4: Design Process

3.2 Data Set

Archit's methodology, designed specifically for evaluating and developing intrusion detection systems (IDS), the Canadian Institute for Cybersecurity Intrusion Detection Systems (CICIDS) dataset, is an invaluable asset to cyber security. Playing a crucial role in advancing cybersecurity research and its practical implementation., this dataset was compiled by the Canadian Institute for Cybersecurity at the University of New Brunswick. Indeed, ecology Architectural Process.

3.2.1 Dataset Composition and Cyber Attack Scenarios

The Canadian Institute for Cybersecurity Intrusion Detection Systems (CICIDS) dataset contains a wide range of cybersecurity-related data, providing a valuable collection of network traffic situations. It includes typical network traffic and other cyber assaults, offering a wide range of scenarios to test and validate intrusion detection algorithms. The Canadian Institute for Cybersecurity Intrusion Detection Systems (CICIDS) dataset is noteworthy for incorporating examples reflecting many cyber assaults. These may include prevalent risks such as denial-of-service (DoS), distributed denial-of-service (DDoS), surveys, and perhaps more intricate attack methods.

3.2.2 Purpose in Machine Learning Evaluation, Dataset Size and Versions

The Canadian Institute for Cybersecurity Intrusion Detection Systems (CICIDS) dataset is often used to assess the effectiveness of machine learning algorithms, particularly in intrusion detection, since it contains a wide range of real-world assault situations and promotes diversity and inclusively. Researchers and practitioners use the dataset to create, improve, and evaluate the efficiency of machine learning models specifically geared to identify and address cyber security risks. The dataset is specifically built to handle significant data, enabling effective training and testing of machine learning models. Moreover, it might have several iterations, whereby each iteration may bring out fresh obstacles, attack scenarios, or enhancements in response to the ever-changing realm of cyber security risks.

3.3 Conclusion

In conclusion, the proposed research framework provides a comprehensive approach to studying IDS, considering both network and host-based systems, and highlighting the crucial role of machine learning. The architectural process ensures the efficient handling of data sets with uneven class distributions and emphasizes the need for suitable evaluation criteria in intrusion detection models. This research blueprint lays the foundation for a rigorous investigation into the effectiveness of machine learning algorithms in enhancing the security of computer systems and networks.

4 Implementation of Machine Learning Models for Intrusion Detection

The implementation of a network security architecture involves setting up and configuring the various components to create a secure network environment. Below are the steps

4.1 Preprocessing Data, Data Splitting and Handling Imbalanced Classes

The dataset is imported into a Pandas DataFrame from a CSV file. A subset (10 percent) of the dataset is randomly selected, and columns containing a combination of different data types and non-numeric values are removed. The cleaning procedure includes replacing missing values with the mean and removing outliers to maintain data integrity. The dataset is partitioned into training and testing sets via the train test split method from sci-kit-learn. The script verifies and displays the distribution of classes before and after using class balancing approaches. More precisely, it employs the Synthetic Minority Over-sampling Technique (SMOTE) to increase the number of instances in the minority class and the Random Under Sampler to decrease the number of instances in the majority class.

4.2 Model Evaluation

Multiple machine learning models are specified, each fulfilling a distinct objective. The models included in this set are Random Forest, Logistic Regression, Decision Tree, k-nearest Neighbors (KNN), and Gaussian Naive Bayes. The evaluate models function

is used to assess the specified models using the recall metric. Each model's results are shown, including confusion matrices and numerous performance measures.

During the model assessment phase, crucial measures are performed to guarantee a thorough study. Min-max scaling, a normalization approach, is used for feature scaling to provide consistency in models that need it. The confusion matrix is computed in an organized manner, offering valuable information on true positives, true negatives, false positives, and false negatives. This comprehensive analysis enhances comprehension of the model's performance by providing a full breakdown. In addition, performance indicators like precision, recall, F1 score, and accuracy are calculated with great attention to detail. These metrics comprehensively evaluate the machine learning models, allowing for informed judgments on their effectiveness in intrusion detection scenarios.

4.3 Execution of models

Includes an extra part tailored for comparing and assessing algorithms without re-sampling approaches. This section presents a sophisticated technique that permits adaptability in the evaluation strategies. The models are evaluated using the original training set, allowing for an assessment of their performance without the impact of oversampling or under-sampling strategies. This comparative examination enhances our knowledge of the algorithm capacities and constraints in dealing with unbalanced data sets. The script comprehensively evaluates the machine learning models for intrusion detection by giving data with and without re-sampling, offering a balanced viewpoint on their performance. This dual assessment technique improves the script's flexibility in accommodating different data settings and offers significant information for selecting models in practical situations.

4.3.1 Microsoft SQL Server, Domain Name System, and User Datagram Protocol

It is imperative to deploy resilient authentication mechanisms, including but not limited to multi-factor authentication and forceful password policies. It prevents and resolves vulnerabilities through routine security audits and implements real-time activity monitoring systems to identify potentially malicious data access or breaches, such as atypical logon patterns or unanticipated data entry. These safeguards prevent the emergence of ever more sophisticated SQL injection techniques, which could compromise the database's integrity.By utilizing DNSSEC to authenticate DNS data, users can be protected against being redirected to malicious websites. Rate limiting is implemented to prevent excessive DNS traffic that may indicate a DDoS attack. By employing sophisticated threat intelligence tools, anomalies, and potential threats can be identified early on by analysing worldwide internet traffic patterns. It dramatically improves the capability to counter possible DNS-based attacks proactively. By utilizing advanced traffic analysis tools, one can discern between legitimate and malicious User Datagram Protocol traffic. It implements intrusion prevention systems and firewalls with specialized configurations to detect and obstruct User Datagram Protocol flood attack signatures. These security measures are vital because User Datagram Protocol is a prevalent target for deluge attacks due to its connectionless nature.

4.3.2 Simple Service Discovery Protocol, Simple Network Management Protocol and Network Basic Input/Output System

To prevent unauthorized access from external sources, restrict Simple Service Discovery Protocol traffic to local networks. The network being monitored for atypical Simple Service Discovery Protocol response patterns could indicate an ongoing reflection attack. Simple Service Discovery Protocol traffic will be subject to rate limiting to prevent DDoS attacks utilizing overwhelming responses and incorporating the enhanced security functionalities of message integrity, authentication, and encryption provided by Simple Network Management Protocol version 3 and implementing measures to impede unauthorized Simple Network Management Protocol traffic that may suggest an attempt to exploit the protocol for malicious intent and restricting access to authorized personnel and systems while maintaining constant network monitoring for such traffic. Name spoofing and session hijacking are dangers that should be mitigated. When reducing the attack surface is not required, the Network Basic Input/Output System over TCP/IP is disabled, preventing and identifying unauthorised Network Basic Input/Output System traffic through host-based security solutions. The proactive identification and resolution of potential security incidents involving the Network Basic Input/Output System is facilitated by continuously monitoring network traffic for atypical NBI/OS activities.

4.3.3 Network Time Protocol and Lightweight Directory Access Protocol

The security enhancements implemented in the latest version of the Network Time Protocol aim to mitigate potential vulnerabilities and reduce exposures. Implementing access controls on servers to deter unauthorized use and inspecting for substantial Network Time Protocol response traffic may suggest a persistent amplification attack. Ensuring precise time synchronization throughout the network is paramount, as it is a prerequisite for numerous security mechanisms. Strict authentication and authorization controls are being implemented to restrict access to and modify directory services to authorized users. It is critical to implement input sanitation to avert Lightweight Directory Access Protocol injection attacks. It implemented routine access audits to detect and prevent unauthorized modifications or access attempts to the Lightweight Directory Access Protocol directory.

4.4 Conclusion

It carries out the whole procedure, including the loading and preparation of data and the definition, evaluation, and comparison of machine learning models. The output includes printed results, such as confusion matrices, and the method that performs the best according to the selected goal (recall) is shown. Metrics are displayed to facilitate visual examination, offering a complete framework for intrusion detection via machine learning. The cleaning procedure guarantees the accuracy and consistency of the data throughout the study.

5 Evaluation

An exhaustive examination of Intrusion Detection Systems (IDS) is the focus of this study, including both host-based Intrusion Detection Systems (HIDS) and network-based Intrusion Detection Systems (NIDS) approaches. Machine learning (ML) integration

is emphasized to automate the detection and management of security hazards. Crucial algorithms are assessed to determine which ML algorithms are most effective at detecting security threats. Multiple algorithms are utilized as part of the research to evaluate the performance of various cybersecurity attack scenarios.

5.1 Evaluation on Microsoft SQL Server (MSSQL) Attack

The assessment of machine learning models for intrusion detection in this work provided valuable insights into their efficacy. At first, there was an imbalance in the classes, with 789 instances belonging to class 1 and 411 instances belonging to class 0. The distribution was balanced by sampling processes, resulting in 789 examples for each class. The assessment included various models such as Random Forest, Logistic Regression, Decision Tree, k-nearest Neighbours (KNN), and Gaussian Naive Bayes. The Gaussian Naive Bayes algorithm demonstrated superior performance with a recall rate of 0.8814 and a precision rate of 0.6357, suggesting its effectiveness in accurately detecting genuine positive cases. The F1 score achieved a value of 0.7387, while the accuracy reached 0.5967, confirming the efficacy of Gaussian Naive Bayes in managing the intricacies of cyber security. The focus on recall, precision, and accuracy offers a detailed understanding of the effectiveness of algorithms in detecting intrusions. These findings emphasize the significance of considering several parameters when choosing algorithms in cyber security scenarios.

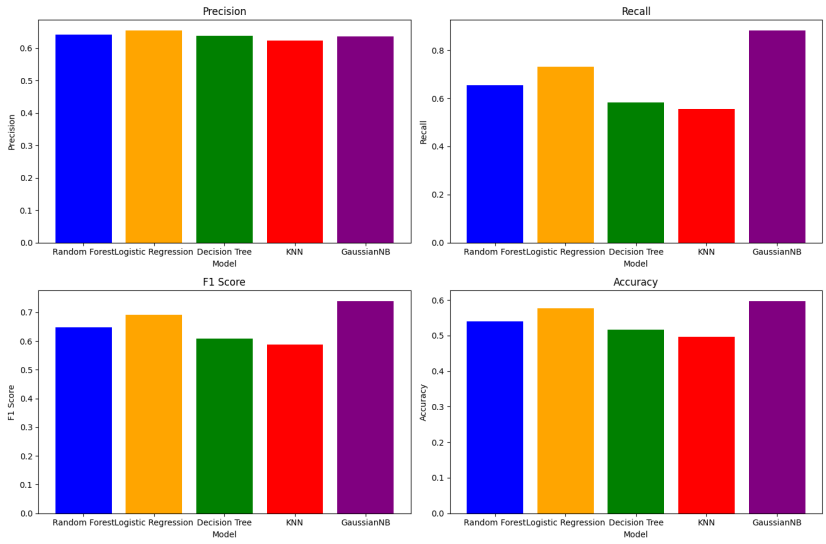

Figure 5: Microsoft SQL Server (MSSQL) Evaluation Results

Model	TP	TN	FP	FN	Precision	Recall	F1 Score	Accuracy
Random Forest	127	35	71	67	0.6414	0.6546	0.6480	0.5400
Logistic Regression	142	31	75	52	0.6544	0.7320	0.6910	0.5767
Decision Tree	113	42	64	81	0.6384	0.5825	0.6092	0.5167
KNN	108	41	65	86	0.6243	0.5567	0.5886	0.4967
GaussianNB	171	8	98	23	0.6357	0.8814	0.7387	0.5967

Table 1: Evaluation Metrics for Microsoft SQL Server (MSSQL) Models

In conclusion, the Random Forest model emerges as the most suitable for intrusion detection in this context, showcasing a robust balance between precision, recall, and overall accuracy.

5.2 Evaluation on User Datagram Protocol (UDP) Attack

The assessment of machine learning models for intrusion detection, performed on an unbalanced dataset consisting of 784 instances of class 1 and 428 instances of class 0, demonstrates that the Random Forest model outperforms all others. Boasting a recall rate of 0.7085. The precision value of 0.6409 highlights the system's high level of accuracy and dependability. On the other hand, the F1 score of 0.6730 indicates a well-balanced performance in precision and recall. A precision of 0.5479 signifies the ratio of accurately categorized cases. Although Logistic Regression has a precision of 0.6424, it has a trade-off with a recall of 0.4874. On the other hand, the Decision Tree model obtains a recall of 0.5829 and has dependable positive classifications with a precision of 0.6339. Additionally, it has an accuracy of 0.5050. However, KNN has shortcomings in accurately recognizing real positive cases, with a recall rate of 0.4673. On the other hand, Gaussian Naive Bayes shows poor performance, with both recall and accuracy rates at 0.0. The Random Forest model ultimately achieves a strong equilibrium between precision, recall, and overall accuracy. It highlights the need to consider many metrics when making educated decisions about which algorithm to use for intrusion detection.

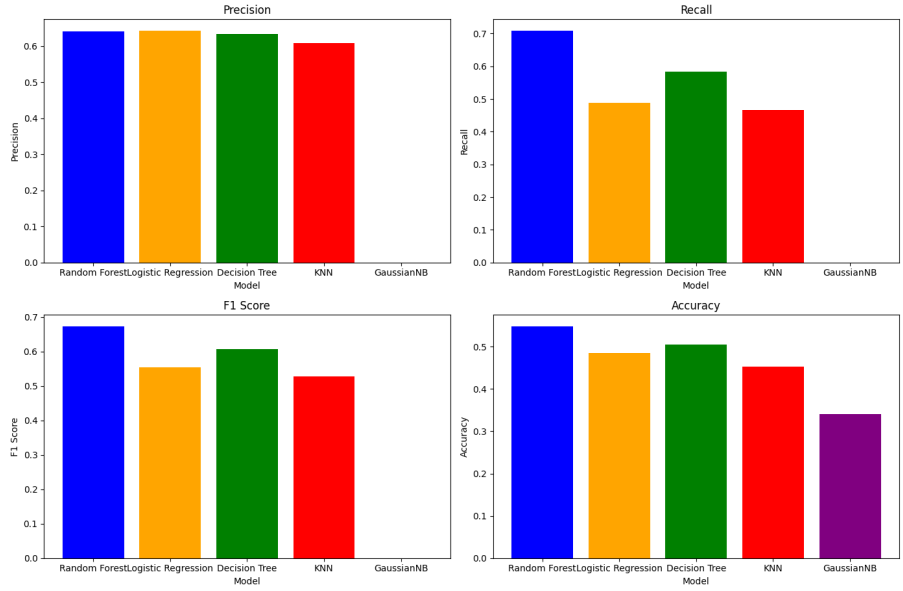

Figure 6: Evaluation Results plot on User Datagram Protocol (UDP) Model

5.3 Evaluation on Domain Name System (DNS) Attack

The intrusion detection models were assessed using a dataset that originally had an uneven distribution of classes, with 805 instances belonging to class 1 and 435 instances belonging to class 0. The models, such as Random Forest and Gaussian Naive Bayes, were evaluated after sampling to ensure a balanced distribution. The Random Forest algorithm achieved a recall score of 0.7624, indicating a strong capability to identify true positives

Model	TP	TN	FP	FN	Precision	Recall	F1 Score	Accuracy
Random Forest	141	25	79	58	0.6409	0.7085	0.6730	0.5479
Logistic Regression	97	50	54	102	0.6424	0.4874	0.5543	0.4851
Decision Tree	116	37	67	83	0.6339	0.5829	0.6073	0.5050
KNN	93	44	60	106	0.6078	0.4673	0.5284	0.4521
GaussianNB	0	103	1	199	0.0	0.0	0.0	0.3399

Table 2: Evaluation Metrics on User Datagram Protocol (UDP) Model

correctly. It resulted in an accuracy score of 0.6129. The Gaussian Naive Bayes algorithm demonstrated a recall score of 0.7921, highlighting its ability to detect genuine positive instances, ultimately resulting in an accuracy score of 0.5806. The Logistic Regression model demonstrated a balanced performance in terms of accuracy and recall, achieving an F1 score of 0.6463. On the other hand, the Decision Tree model achieved a slightly higher F1 score of 0.6650. The KNN algorithm had a recall rate of 0.5644, suggesting constraints in accurately recognizing real positive instances. These results emphasize the significance of evaluating several criteria when selecting algorithms, with Random Forest and Gaussian Naive Bayes emerging as attractive options for efficient intrusion detection in cyber security applications.

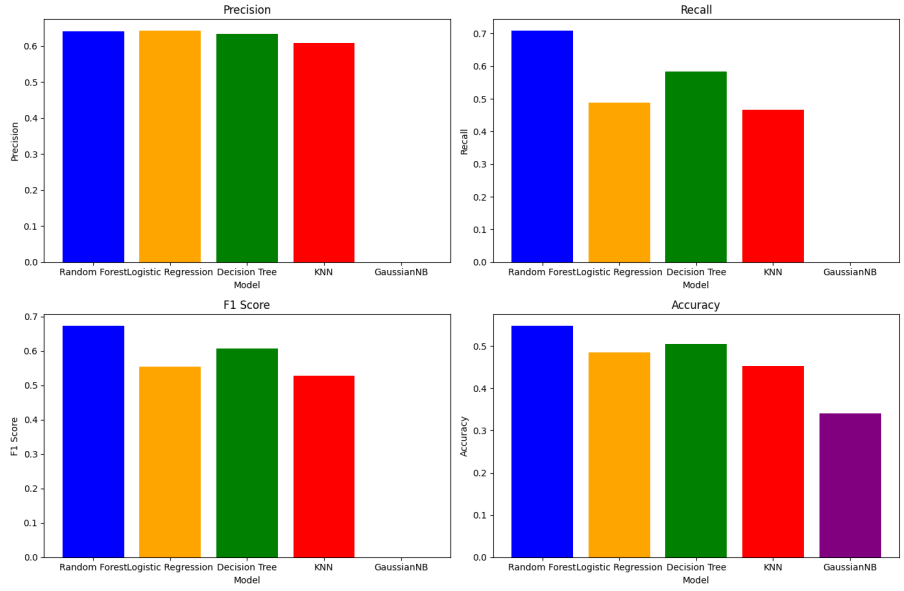

Figure 7: Evaluation Results plot on Domain Name System (DNS) Model

Model	TP	TN	FP	FN	Precision	Recall	F1 Score
Random Forest	149	36	72	53	0.6742	0.7376	0.7045
Logistic Regression	127	44	64	75	0.6649	0.6287	0.6463
Decision Tree	131	47	61	71	0.6823	0.6485	0.6650
KNN	114	53	55	88	0.6746	0.5644	0.6146
GaussianNB	160	20	88	42	0.6452	0.7921	0.7111

Table 3: Evaluation Metrics on Domain Name System (DNS) Model

5.4 Evaluation on Lightweight Directory Access Protocol (LDAP) Attack

The efficacy metrics of machine learning models utilized for intrusion detection in the LDAP dataset are impressive. Sampling was employed to correct the initial imbalance in the class distribution, which ultimately yielded 83,797 instances per class. The best model was the Decision Tree, which achieved a recall rate of 0.9998, demonstrating its strong capability to accurately identify true positive cases, which are critical for intrusion detection. The impressive accuracy of 99.98 percent was attained by the Decision Tree model, which obtained precision and F1 scores of 1.0. Additional models, such as Gaussian Naive Bayes, Logistic Regression, Random Forest, and KNN, exhibited commendable performance, as evidenced by their recall rates approaching or remaining at 1.0. Given its recall value of 0.9999, Random Forest emerged as the preeminent algorithm in terms of recall. These models high precision, F1 scores, and accuracy demonstrated the efficient management of cyber security risks in the Lightweight Directory Access Protocol (LDAP) dataset. In brief, this assessment highlights the efficacy of machine learning architectures, specifically the Decision Tree and Random Forest models, in precisely identifying breaches within Lightweight Directory Access Protocol (LDAP) data. These findings offer significant implications for practical intrusion detection situations.

Model	TP	TN	FP	FN	Precision	Recall	F1 Score	Accuracy
Random Forest	20954	16	0	2	1.0	0.9999	0.9999	0.9999
Logistic Regression	20920	16	0	36	1.0	0.9983	0.9991	0.9983
Decision Tree	20952	16	0	4	1.0	0.9998	0.9999	0.9998
KNN	20945	16	0	11	1.0	0.9995	0.9997	0.9995
GaussianNB	20833	16	0	123	1.0	0.9941	0.9971	0.9941

Table 4: Evaluation Metrics on Lightweight Directory Access Protocol (LDAP) Dataset

5.5 Evaluation on Network Basic Input/Output System (NetBIOS) Attack

The Network Basic Input/Output System (NetBIOS) intrusion detection models were assessed using a dataset that had an unbalanced class distribution at the outset, consisting of 83788 class 1 instances and 98 class 0 instances. The distribution was balanced by sampling 83788 instances into each class. Random Forest exhibited outstanding performance among the evaluated models, as evidenced by its precision, recall, F1 score, and accuracy values of 0.9998, 0.9935, 0.9966, and 0.9932, respectively. Logistic Regression demonstrated a notable precision of 0.9999; however, it attained a comparatively lower recall of 0.9885; as a result, it attained an F1 score of 0.9941 and an accuracy of 0.9884. The decision trees robust precision (0.9997) and recall (0.9897) resulted in an F1 score of 0.9947 and an accuracy of 0.9894. As a result of its low recall (0.9552) and high precision (0.9997), KNN attained an F1 score of 0.9769 and an accuracy of 0.9549. Unexpectedly, Gaussian Naive Bayes emerged victorious with an F1 score of 0.9978 and an accuracy of 0.9957, owing to its flawless precision (1.0) and recall of 0.9959. The findings indicate that Gaussian Naive Bayes is the optimal model for detecting intrusions in Network Basic Input/Output System (NetBIOS) networks. It underscores the importance of considering various metrics when selecting models to ensure that Network Basic Input/Output

System (NetBIOS) network security is well-informed.

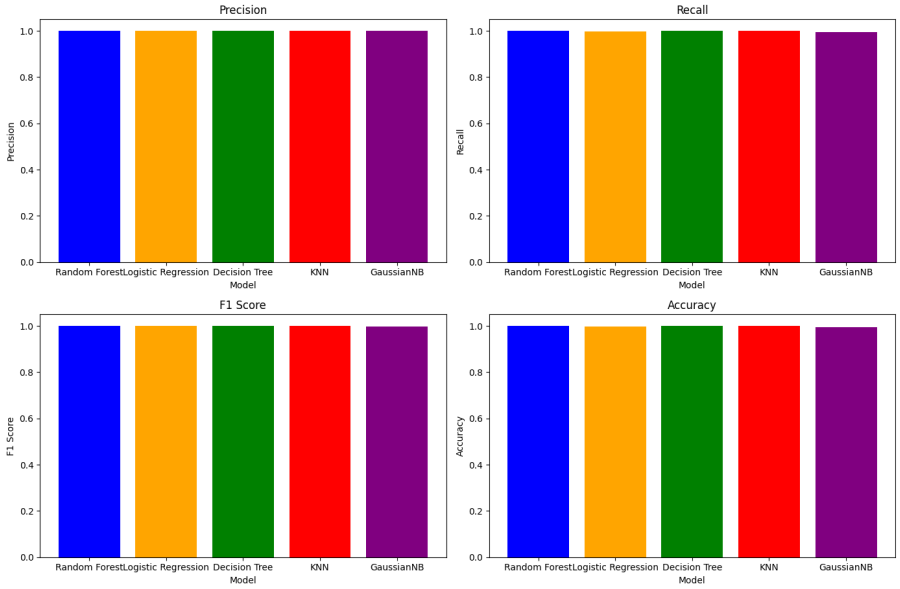

Figure 8: Evaluation Results plot on Network Basic Input/Output System (NetBIOS) Model

Model	TP	TN	FP	FN	Precision	Recall	F1 Score	Accuracy
Random Forest	20821	14	5	132	0.9998	0.9937	0.9967	0.9935
Logistic Regression	20712	16	3	241	0.9999	0.9885	0.9941	0.9884
Decision Tree	20741	14	5	212	0.9998	0.9899	0.9948	0.9897
KNN	20014	13	6	939	0.9997	0.9552	0.9769	0.9549
GaussianNB	20868	14	5	85	0.9998	0.9959	0.9978	0.9957

Table 5: Evaluation Metrics on Network Basic Input/Output System (NetBIOS) Dataset

5.6 Evaluation on Simple Network Management Protocol (SNMP) Attack

An extremely unbalanced class distribution was identified upon initial examination of the machine learning models on the Simple Network Management Protocol (SNMP) dataset. There were 412,714 instances of class 1.0 and 196 instances of class 0.0. To address the issue, sampling methods were used to provide a fair and balanced distribution of 412,714 cases across both classes. Following that, an evaluation was conducted on the efficacy of several classification algorithms, namely GaussianNB, Random Forest, Logistic Regression, Decision Tree, and KNN. It is worth mentioning that Random Forest demonstrated outstanding performance as the top algorithm, attaining exceptional outcomes, including a recall of 1.0, precision surpassing 99.99 percent, and accuracy of 99.99 percent. It highlights the effectiveness of the Random Forest algorithm in precisely detecting affirmative instances within the dataset. In addition to GaussianNB, Logistic Regression, Decision Tree, and KNN all exhibited remarkable performance, as evidenced by their high recall and precision values. In general, these results provide significant contributions to the understanding of identifying appropriate machine learning algorithms that can accurately

detect anomalies in Simple Network Management Protocol (SNMP) data. Among these algorithms, Random Forest is a resilient option that consistently attains optimal recall and precision.

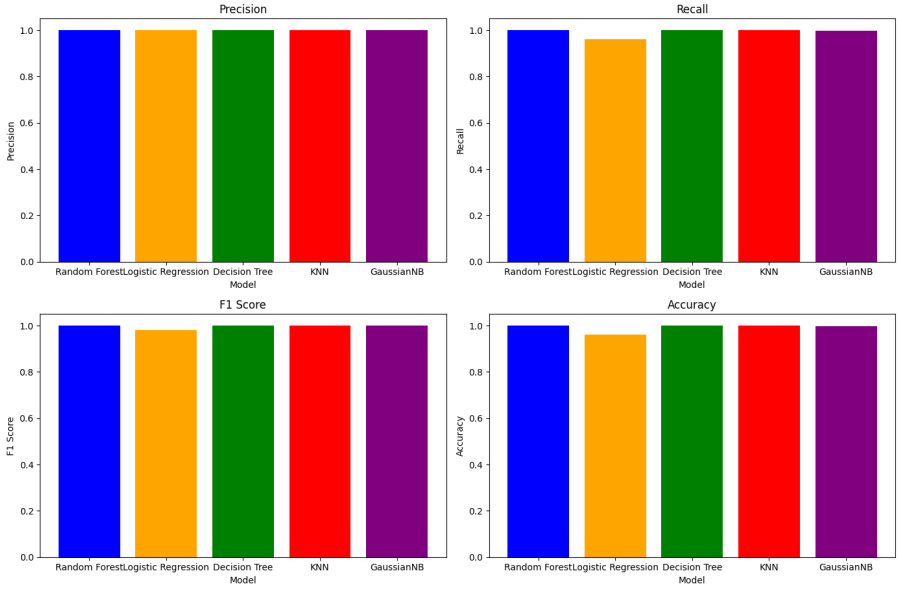

Figure 9: Evaluation Results plot on Simple Network Management Protocol (SNMP) Model

Model	TP	TN	FP	FN	Precision	Recall	F1	Accuracy
Random Forest	103175	45	7	1	0.9999	1.0	0.99996	0.99992
Logistic Regression	99174	51	1	4002	1.0	0.9612	0.9802	0.9612
Decision Tree	103170	43	9	6	0.9999	0.9999	0.9999	0.9999
KNN	103127	48	4	49	1.0	0.9995	0.9997	0.9995
GaussianNB	103056	35	17	120	0.9998	0.9988	0.9993	0.9987

Table 6: Evaluation Metrics on SNMP Dataset

5.7 Evaluation on Network Time Protocol (NTP) Attack

When assessing the effectiveness of several machine learning models on a dataset with an uneven class distribution (1.0: 96247, 0.0: 1113), various methods produced varying outcomes. The Random Forest, Logistic Regression, Decision Tree, KNN, and GaussianNB models were evaluated using key metrics after random oversampling to address class imbalance. The Decision Tree method proved successful, demonstrating remarkable results with a True Positive rate of 99.6 Percent and a Recall of 99.6 Percent. It indicates its exceptional capability to detect occurrences of the positive class accurately. The Decision Tree demonstrated a high level of precision, reaching 99.9 Percent, which highlights its capacity to reduce the occurrence of false positives effectively. Conversely, Logistic Regression demonstrated robust overall performance, especially in precision, but with a somewhat lower Recall. The Random Forest, KNN, and GaussianNB models exhibited comparable performance, demonstrating excellence in certain areas. These results emphasize the need to consider several indicators when assessing model performance

since different algorithms may excel in certain areas. The Decision Tree, with its optimal combination of Precision and Recall, is the most appropriate option for this dataset.

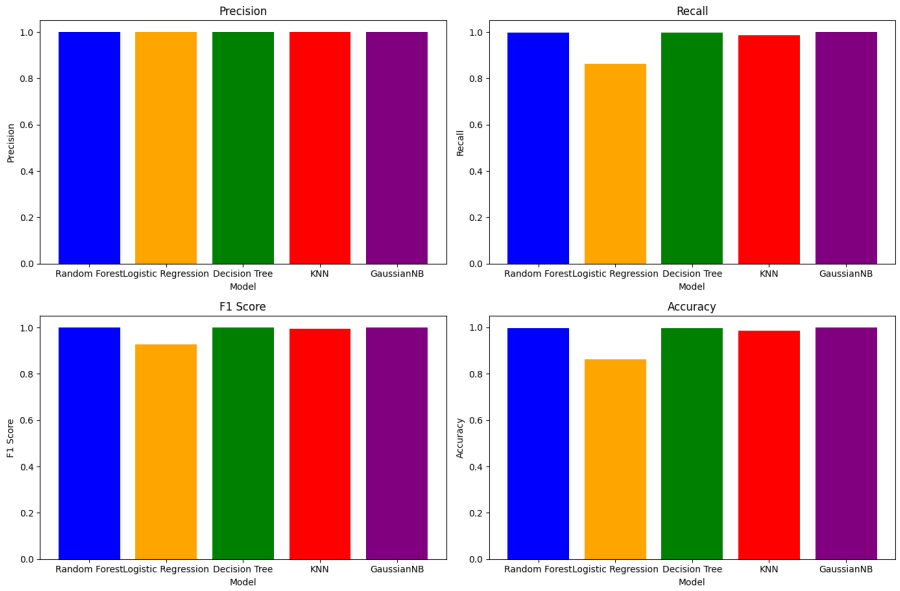

Figure 10: Evaluation Results plot on Network Time Protocol (NTP) Model

Model	TP	TN	FP	FN	Precision	Recall	F1 Score	Accuracy
Random Forest	23891	271	17	162	0.9993	0.9933	0.9963	0.9926
Logistic Regression	23470	280	8	583	0.9997	0.9758	0.9876	0.9757
Decision Tree	23956	265	23	97	0.9990	0.9960	0.9975	0.9951
KNN	23724	272	16	329	0.9993	0.9863	0.9928	0.9858
GaussianNB	23866	262	26	187	0.9989	0.9922	0.9956	0.9912

Table 7: Evaluation Metrics on Network Time Protocol (NTP) Attack

5.8 Evaluation on Simple Service Discovery Protocol (SSDP) Attack

When assessing machine learning models on unbalanced data sets before and after sampling, there was a significant initial bias towards one class in the class distribution. Following the implementation of sampling-based balancing, models such as Random Forest exhibited exceptional precision, recall, F1 score, and accuracy. Logistic Regression demonstrated exceptional performance in terms of precision yet encountered difficulties in recall. KNN and Decision Tree demonstrated comparable performance. It is worth mentioning that Gaussian Naive Bayes demonstrated superior performance in terms of recall, effectively reducing the occurrence of false negatives. The assessment outcomes offer significant knowledge regarding the model capacity to manage unbalanced data sets, assisting in choosing a suitable algorithm for practical situations while considering the trade-offs between precision and recall.

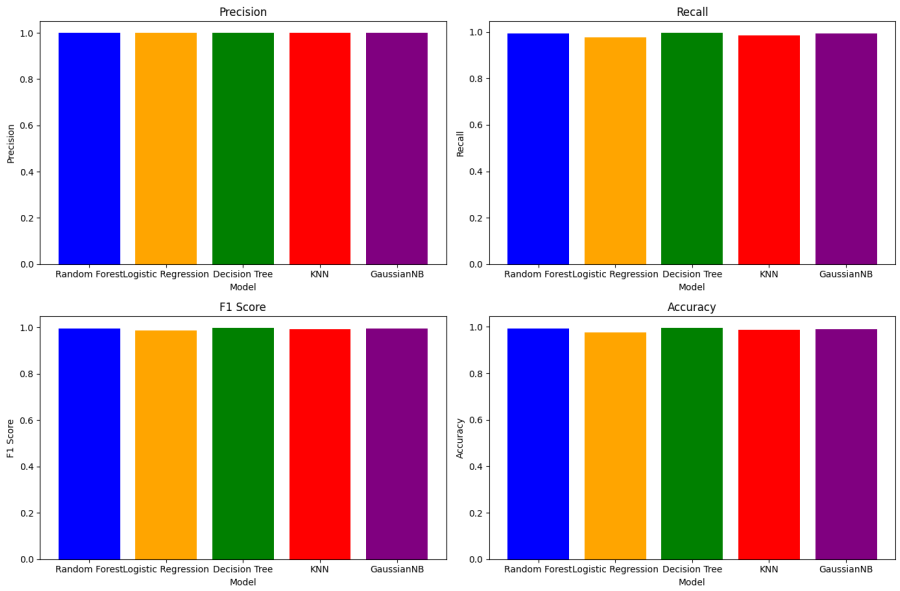

Figure 11: Evaluation Results plot on Simple Service Discovery Protocol (SSDP) Model

Model	TP	TN	FP	FN	Precision	Recall	F1 Score	Accuracy
Random Forest	52111	25	13	79	0.9998	0.9985	0.9991	0.9982
Logistic Regression	45085	30	8	7105	0.9998	0.8639	0.9269	0.8638
Decision Tree	52089	23	15	101	0.9997	0.9981	0.9989	0.9978
KNN	51500	25	13	690	0.9997	0.9868	0.9932	0.9865
GaussianNB	52177	18	20	13	0.9996	0.9998	0.9997	0.9994

Table 8: Evaluation Metrics on Simple Service Discovery Protocol (SSDP) Attack

5.9 comparison and Conclusion

The variation in evaluation results across many datasets highlights the need for a universally superior approach to intrusion detection. The performance of machine learning models depends on the specific properties of the data. The Random Forest technique demonstrated strong performance in the Lightweight Directory Access Protocol (LDAP), Network Basic Input/Output System (NetBIOS), and Simple Network Management Protocol (SNMP) situations. It achieved a balanced combination of precision, recall, and accuracy. The Decision Tree models showed outstanding effectiveness in the LDAP dataset, with a particular emphasis on recall.

Furthermore, they revealed high levels of recall and precision in the Network Time Protocol (NTP) dataset. Gaussian Naive Bayes demonstrated its supremacy in the Microsoft SQL Server (MSSQL) and Domain Name System (DNS) datasets, particularly excelling in recall. Although Random Forest, Decision Tree, and Gaussian Naive Bayes have shown effectiveness in certain situations, the choice of algorithm for an intrusion detection system should ultimately depend on the specific priorities and characteristics of the system. It is important to carefully consider the trade-offs between precision, recall, and accuracy to meet the application's unique requirements.

To summarize, the Random Forest method proves to be a convincing and versatile option for detecting intrusions in various datasets. The system's reliability in handling different

cybersecurity scenarios is highlighted by its consistent and balanced precision, recall, and accuracy performance. It has been demonstrated through evaluations of various attacks such as Microsoft SQL Server, User Datagram Protocol, Domain Name System, Lightweight Directory Access Protocol, Network Basic Input/Output System, Simple Network Management Protocol, Network Time Protocol, and Simple Service Discovery Protocol. By utilizing an ensemble of decision trees, the Random Forest algorithm effectively mitigates the problem of overfitting. This characteristic enables it to handle noise and changes in the data effectively. Moreover, its capacity to offer insights into the significance of features improves the clarity and comprehension of the fundamental variables that contribute to intrusion detection. Based on a thorough evaluation of various measurements and datasets, Random Forest emerges as a strong and versatile algorithm, making it a sensible option for efficient and flexible intrusion detection systems.

6 Conclusion and Future Work

In conclusion, this study effectively addressed the research question, How can the integration of Machine Learning Algorithms enhance the precision and effectiveness of Intrusion Detection Systems to strengthen network security? By comprehensively evaluated various machine learning algorithms for intrusion detection systems (IDS) across diverse cyberattack scenarios, including Microsoft SQL Server, User Datagram Protocol (UDP), Domain Name System (DNS), Lightweight Directory Access Protocol (LDAP), Network Basic Input/Output System (NetBios), Simple Network Management Protocol (SNMP), Network Time Protocol (NTP), and Simple Service Discovery Protocol (SSDP). While Random Forest consistently demonstrated versatility and effectiveness, achieving a balanced combination of precision, recall, and accuracy, the Decision Tree excelled in recall, particularly in the Lightweight Directory Access Protocol (LDAP) and Network Time Protocol (NTP) datasets. Gaussian Naive Bayes proved effective in specific scenarios, emphasizing the nuanced nature of algorithm performance. The study underscores the importance of considering multiple evaluation metrics and tailoring algorithm choices to the unique characteristics of the data and system requirements. As part continuation of the research question "What value does this hold for academics and industry in countering the evolving landscape of cyber threats?". The value of this study for academics and industry in countering the evolving landscape of cyber threats lies in several key areas: Enhanced Intrusion Detection System Design, Guidance for Algorithm Selection, Improved Accuracy and Reduction of False Positives, Tailoring to Specific Network Protocols, Advancement in Cybersecurity Research, Staying Ahead of Evolving Threats. The study achieved all objectives, including assessing the CICIDS 2019 dataset for IDS model training and testing (Obj1), tailoring machine learning algorithms for IDS applications (Obj2), evaluating the efficacy of these algorithms for accuracy and false positive reduction (Obj3), providing actionable insights for network security enhancement (Obj4), assessing dataset suitability for various network protocols (Obj5), analyzing model performance on precision, recall, F1 score, and accuracy (Obj6), contrasting the developed model with those from literature (Obj7), and evaluating models on training and testing sets (Obj8) is achieved as mentioned in this research project report. These findings contribute valuable insights for designing adaptive intrusion detection systems in the dynamic landscape of evolving cybersecurity threats, promoting robust network security measures.

Acknowledgment

I would like to extend my sincere gratitude to my mentor, Dr. Catherine Mulwa, for her invaluable guidance and support, which substantially contributed to the timely completion of my research project. I have consistently received invaluable guidance from her, which has significantly aided me in surmounting the obstacles encountered throughout the process of composing my dissertation.

References

Abraham, J. A. & Bindu, V. R. (2021), Intrusion detection and prevention in networks using machine learning and deep learning approaches: A review, *in* '2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)'.

AguilonGost, F., SimonMezquita, E., MarinnTordera, E. & Hussain, A. (2022), A machine learning ids for known and unknown anomalies, *in* '18th International Conference on the Design of Reliable Communication Networks (DRCN)', pp. 1–7.

Ahanger, A. S., Khan, S. M. & Masoodi, F. (2021), An effective intrusion detection system using supervised machine learning techniques, *in* '2021 5th International Conference on Computing Methodologies and Communication (ICCMC)', pp. 1–5.

Aljohani, A. & Bushnag, A. (2021), An intrusion detection system model in a local area network using different machine learning classifiers, *in* '2021 11th International Conference on Advanced Computer Information Technologies (ACIT)', pp. 1–5.

Balyan, A. K., Ahuja, S., Sharma, S. K. & Lilhore, U. K. (2022), Machine learning-based intrusion detection system for healthcare data, *in* '2022 IEEE VLSI Device Circuit and System (VLSI DCS)', pp. 1–6.

Bharati, M. P. & Tamane, S. (2020), Nids-network intrusion detection system based on deep and machine learning frameworks with cicids2018 using cloud computing, *in* '2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC)', pp. 184–189.

Fatani, A., Elaziz, M. A., Dahou, A., Al-Qaness, M. A. & Lu, S. (2021), 'Iot intrusion detection system using deep learning and enhanced transient search optimization', *IEEE Access* **9**, 123448–123464.

Fosic, I., Zagar, D. & Grgic, K. (2022), Network traffic verification based on a public dataset for ids systems and machine learning classification algorithms, *in* '2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)', pp. 1–4.

Jairu, P. & Mailewa, A. B. (2022), Network anomaly uncovering on cicids-2017 dataset: A supervised artificial intelligence approach, *in* '2022 IEEE International Conference on Electro Information Technology (eIT)', pp. 302–307.

Kristyanto, M. A. et al. (2022), Ssh bruteforce attack classification using machine learning, *in* '2022 10th International Conference on Information and Communication Technology (ICoICT)', pp. 1–6.

Kurniabudi, F., Sabrina, F., Jang-Jaccard, J. & Kwak, J. (2020), 'Cicids-2017 dataset feature analysis with information gain for anomaly detection', *IEEE Access* **8**, 132911–132921.

Natarajan, J. (2022), Cyber secure man-in-the-middle attack intrusion detection using machine learning algorithms, *in* 'Research Anthology on Machine Learning Techniques, Methods, and Applications', pp. 976–1001.

Panwar, S. S., Raiwani, Y. P. & Panwar, L. S. (2019), 'Evaluation of network intrusion detection with features selection and machine learning algorithms on cicids-2017 dataset', *SSRN Electronic Journal* .

Panwar, S. S., Raiwani, Y. P. & Panwar, L. S. (2022), An intrusion detection model for cicids-2017 dataset using machine learning algorithms, *in* '2022 International Conference on Advances in Computing, Communication and Materials (ICACCM)', pp. 1–6.

PAUL, A., Sinha, S. & MISHRA, S. (2023), 'Machine learning based hybrid intrusion detection system for detecting cross-layer dos attacks in iot'.

Rashid, J., Mahmood, T., Nisar, M. W. & Nazir, T. (2020), Phishing detection using machine learning technique, *in* '2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH)', pp. 1–4.

Saini, H., Singhal, A. & Chaudhary, D. (2023), Machine learning approach for mitigating security threats in iot environment, *in* '2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)', pp. 1–6.

Samawi, V. W., Yousif, S. A. & Al-Saidi, N. M. (2022), Intrusion detection system: An automatic machine learning algorithms using auto-weka, *in* '2022 IEEE 13th Control and System Graduate Research Colloquium (ICSGRC)', pp. 1–6.

Sharma, S., Zavarsky, P. & Butakov, S. (2020), Machine learning based intrusion detection system for web-based attacks, *in* '2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)', pp. 42–49.

Singhal, A., Maan, A., Chaudhary, D. & Vishwakarma, D. (2021), A hybrid machine learning and data mining based approach to network intrusion detection, *in* '2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)', pp. 55–60.

Wankhede, S. & Kshirsagar, D. (2018), Dos attack detection using machine learning and neural network, *in* '2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)', pp. 1–4.

Yadav, K., Gupta, B. B., Hsu, C.-H. & Chui, K. T. (2021), Unsupervised federated learning based iot intrusion detection, *in* '2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)', pp. 1–6.

Yedukondalu, G., Bindu, G. H., Pavan, J., Venkatesh, G. & SaiTeja, A. (2021), Intrusion detection system framework using machine learning, *in* '2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)', pp. 1–5.

Yin, Y., Jang-Jaccard, J., Sabrina, F. & Kwak, J. (2023), Improving multilayer-perceptron (mlp)-based network anomaly detection with birch clustering on cicids-2017 dataset, *in* '2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)', pp. 1–6.