CHAPTER ONE

INTRODUCTION

1.1 Background to the study

Online banking, often known as internet banking, e-banking, or virtual banking, is an electronic payment system that allows clients of a bank or other financial institution to execute a variety of financial transactions via the bank's website. The concept of online banking emerged in the late 20th century with the proliferation of the internet. In the mid-1990s, financial institutions started offering basic online services such as checking balances and viewing transaction history through web portals. In most cases, an internet banking system will link to or be a component of a bank's core banking system on a network.[1] To many people, electronic banking entails direct deposit of paychecks into checking or savings accounts or 24-hour access to cash through an automated teller machine (ATM). Online banking has revolutionized the way people conduct financial transactions. It allows customers to access their accounts, make payments, transfer funds, and perform various other banking activities through the internet. This convenience, however, comes with its own set of challenges, primarily in the form of security threats. Intrusion detection plays a critical role in safeguarding online banking systems against these threats. The evolution of online banking and the pivotal role of intrusion detection systems in ensuring its security. Over time, online banking evolved to include a wide range of services, including electronic funds transfers, bill payments, and investment management. This expansion was driven by advancements in internet technologies, encryption protocols, and the development of secure communication channels. However, there are numerous different sorts of transactions, rights, obligations, and occasionally fees, associated with using electronic banking. A variety of banking and other services or facilities that employ electronic technology are referred to as "banking services. “These consist of telephone banking, SMS banking, ATM and debit card services, electronic

alerts, mobile banking, money-transfer services, point-of-sale banking, e-statements, and other forms of e-commerce or services that create value[2]. Online banking has a lot of advantages. The two most crucial ones are convenience and speed. Online banking users get access to their accounts, statements, transactions, bill payment options, and more from the comfort of their homes or while on the road. These advantages are the main reasons why about 51% of EU adults utilize internet banking. Online banking does have its advantages, but there are also several unique problems and difficulties in the industry. These are extremely important for banks that provide online banking as well as for their clients, who depend on the banks' smooth operation[3]. While electronic banking benefits the financial system, it poses significant security risks to institutions and their clients. Before a user may access bank services, they must first enter an access code, which is usually in the form of a Personal Identification Number (PIN). This has not always protected banks from fraudsters' theatrics; fraudsters utilize a variety of methods to reveal or steal clients' secret access numbers[4]. Banks typically use manual inspection along with rules-based fraud detection technologies to find scams. In this method, banks establish rules that often describe the characteristics of questionable transactions or activities. They include atypical transaction quantities, transaction categories, or account numbers. These characteristics would mark a transaction or behavior as "fraudulent." If not, they are deemed to be "non-fraudulent." Therefore, rules-based fraud detection is like a gatekeeper in that it admits some people while rejecting others. This approach of fraud detection was very effective back then. However, recently it has demonstrated a constant pattern of failure, unreliable findings, intolerable false positives (as in when the system rejects genuine clients), and false negatives. A security system should be able to guard itself against external intrusions; otherwise, fraudsters may choose to attack the system by turning it off [5]. As a result, it's crucial that an electronic banking application has some level of security intelligence and can protect itself against

existential threats or intrusions. The Intrusion Detection System (IDS) is a powerful protection tool against both network and host-based threats. It gathers, analyzes, and audits security logs and network packets while monitoring important nodes of computer systems or networks. An IDS focuses on the proactive and timely detection of external attackers and unusual server behavior before they do such severe damage. Several cyberattacks have been in dangerous situations as of late, putting certain organizations' vital infrastructures at risk. A successful attack may have unfavorable effects, including but not limited to financial loss, the end of operations, and the revealing of secret information. Additionally, attackers have a greater possibility of success the larger the organization's network is. The network's complexity may also result in weaknesses and other specialized threats. As a result, security mitigation and protection techniques ought to be viewed as required [6]. A potential defense system, like intrusion detection, is essential since it uses preventive measures to get rid of any malicious activities within the computer network. An IDS looks at network and file access logs, audit trails, and other security-relevant data within the organization to detect and block threats without human interaction [7]. Applying algorithms that discover and recognize patterns in data is known as machine learning. Banks may considerably benefit from machine learning in terms of fraud detection and prevention because it allows them to automatically and precisely identify trends within massive quantities of transactions. In general, fraudulent transactions follow different patterns than legitimate ones—even if those differences are quite subtle. To distinguish between fraudsters and honest customers, machine learning algorithms are designed to recognize these suspicious behaviors.

Banks are now able to identify transactions that are most likely to be fraudulent while maintaining acceptable levels of false positives thanks to machine learning. A Creative Ensemble Method for Successful Intrusion Detection is utilized in systems in emerging fields including online banking, networks, healthcare, and e-governance to make choices on crucial

datasets. These techniques are useful and support improved dataset decisions across a variety of fields. The ensemble classifier, a collection of classification algorithms, is one of the most practical methods. Machine learning use these algorithms to carry out efficient classifications. Multiple learners are trained to tackle the same problem using the machine learning paradigm known as ensemble learning. Ensemble methods attempt to create a number of hypotheses and combine them to be used, as opposed to conventional machine learning approaches, which attempt to learn one hypothesis from training data. The ensemble technique combines the results of various classifiers to create a single response [8]. This strategy helps to achieve greater detection than the classification accuracy of a single classifier. Several trainable classifiers, such as base learners, form the foundation of an ensemble learner. Each base learner has been taught to predict for a specific class label, with the final prediction being formed using a specific blending method, such as a combiner. It is assumed that classifier ensembles now outperform individual classifiers for a variety of reasons, including statistical, computational, and representational ones [6]. The vast majority of studies on combining classifiers within the purview of IDS were initially started with a single justification.

Despite the fact that many actions have been indicated to mitigate intrusion detection including the use of ensemble classifier, there is a need for a more scientific approaches that can propel intrusion detection administration in developing countries were not underlined, which is a critical issue. It is of great significance to oversee intrusion and reduce its threat, which requires its detection and prediction at early stage. Despite the encouraging outcomes already conveyed, while implementing ensemble machine learning approaches, there exists uncertain concerns in the use of an enhanced intrusion detection model for online banking using ensemble machine learning approach such as the application of fuzzy logic for refining the features of each dataset and the usage of an add-on optimizers algorithm in in advancing the value of the learning algorithm.

Hence, this research work proposed an enhanced model using ensemble machine learning technique to detect and predict intrusion in online banking. Moreover, the study applied a fuzzy logic method to appraise the performance of the ensemble machine learning algorithm on historical intrusion banking dataset with the aim of advancing the value of datasets.

1.2 Problem Statement

The increase in the usage of smartphones application due to technological advancement and the ease of performing banking activities with little or no security has naturally increased the possibility of intrusion which has become a threat to human life and the world economy.

Several approaches, as reported in literature, have been in use to detect and predict intrusion in online banking (approaches without machine language). However, there is a knowledge gap in understanding the performance of the Homogeneous Boosting technique specifically for online banking network intrusion detection. The use of ensemble classifier has attracted various intertest in cybersecurity research and its application in an IDS domain is not an exception. Recent methods include ensemble machine learning approaches such as (voting, bagging, RF) among others can detect and predict intrusion. The problem is that most of these methods unveil weakness in the application of data decomposition methods for refining the feature of each dataset: instead of analysis a complete set of data, decompose it to a smaller and more manageable data that can be analyzed independently. This underscored the need for this study to employ fuzzy logic approach on various datasets alongside homogenous ensemble machine learning techniques.

Thus, the study was concerned with the use of a fuzzy logic approach for the performance evaluation of intrusion detection model that can increase the detection rate and improve the management of possible online banking risk using homogenous ensemble machine learning techniques.

1.3 Justification for the Study

The justification for this study is based on the critical need for effective online banking network intrusion detection systems. Online banking institutions are increasingly at risk from cybercriminals who use sophisticated techniques to infiltrate banking networks, steal sensitive information, and perpetrate fraudulent activities. To counteract this threat, online banking institutions need robust intrusion detection systems that can detect and respond to security breaches in real-time.

The proposed study aims to evaluate the effectiveness of the Homogeneous Boosting technique for online banking network intrusion detection. This technique combines multiple weak classifiers to create a strong classifier capable of accurately detecting intrusion attempts in real-time. The study's findings will help to determine whether the Homogeneous Boosting technique is a viable solution for improving the accuracy and efficiency of online banking network intrusion detection.

The justification for this study lies in its potential to contribute to the development of more effective online banking network security solutions. The study's findings can inform the development of intrusion detection systems that can detect and responding to security breaches in real-time, thus reducing the risk of financial losses, reputation damage, and customer attrition. The study can also serve as a basis for further research on the use of machine learning techniques for online banking network security.

The study's findings can also have practical implications for cybersecurity experts and online banking institutions. By evaluating the performance of the Homogeneous Boosting technique, the study can provide valuable insights into the strengths and weaknesses of the technique and help to identify ways to optimize its performance. This information can inform decisions

about the selection and implementation of intrusion detection systems for online banking networks.

The study on the "Performance Evaluation of Homogenous Boosting Technique for Intrusion Detection in Online Banking " is substantiated by a multitude of compelling justifications, underscoring its relevance and potential impact in the domains of cybersecurity, financial stability, and technological advancement. The contemporary landscape of online banking is fraught with increasingly sophisticated cyber threats. The proliferation of malware, phishing attacks, and other malicious activities poses a significant risk to the security of online banking networks. This study addresses the urgent need for robust intrusion detection systems in safeguarding sensitive financial information. Intrusion detection forms the cornerstone of cybersecurity for financial institutions, especially in the context of online banking. Detecting and thwarting unauthorized access, malware infiltrations, and other security breaches are paramount to maintaining the integrity and trustworthiness of online banking systems. The homogenous boosting technique, as an ensemble learning method, has demonstrated significant potential in improving the performance of machine learning models. Its ability to combine multiple weak classifiers into a robust, accurate predictor makes it an intriguing candidate for enhancing intrusion detection capabilities in online banking. While there exists a substantial body of research on intrusion detection systems, the evaluation of the homogenous boosting technique specifically within the context of online banking network security represents a relatively unexplored territory. This study seeks to bridge this gap by providing empirical insights and assessments that are tailored to the unique challenges of the online banking environment. The study aligns with the broader trend of harnessing advanced machine learning techniques for cybersecurity applications. By evaluating the homogenous boosting technique, this research contributes to the ongoing quest for innovative, effective solutions that can adapt to the evolving tactics of cyber adversaries. In an era where digital

banking is becoming the norm, ensuring the security of online transactions and financial data is of paramount importance. The study's focus on online banking networks directly addresses the needs of a digitally driven financial landscape, making its findings immediately applicable and crucial. Positive outcomes from this study may have far-reaching implications for the banking industry as a whole. The adoption of the homogenous boosting technique could signify a significant advancement in online banking security practices, potentially setting new industry standards and influencing regulatory guidelines. Beyond its practical implications, the study enriches both the academic and practical domains. It adds to the academic discourse by providing empirical evidence on the efficacy of the homogenous boosting technique. Simultaneously, it offers tangible, practical insights that financial institutions can leverage to fortify their online banking security measures. The study on the "Performance Evaluation of Homogenous Boosting Technique for Online Banking Network Intrusion Detection" is justified by the pressing need to bolster cybersecurity in the realm of online banking. Its potential to yield novel insights, influence industry practices, and fortify financial stability positions it as a significant contribution to the fields of cybersecurity and digital finance. The study's outcomes hold promise for a more secure, resilient online banking ecosystem in the face of evolving cyber threats.

1.4 Aim and Objectives

The aim of this study is to design a fuzzified classifier model based on homogenous boosting ensemble machine learning algorithm for improved detection of intrusion through performance evaluation. The following are the specific objectives to be achieved in this study.

1. apply a fuzzy logic feature selection technique on the selected dataset to determine the objectivity of the homogenous boosting ensemble machine learning algorithms for the performance evaluation of intrusion detection model.

2. evaluate the performance of the homogenous boosting machine learning algorithms using the selected metrics.
3. perform a comparative analysis of homogeneous boosting ensemble algorithm based on the evaluation criteria.

1.5 Significance of the Study

The significance of this study lies in the critical need for effective online banking network intrusion detection systems. With the increasing reliance on online banking, cybercriminals are continually developing new techniques to infiltrate banking networks, steal sensitive information, and perpetrate fraudulent activities. To counteract this threat, online banking institutions need robust intrusion detection systems that can detect and respond to security breaches in real-time.

The proposed study aims to evaluate the effectiveness of the Homogeneous Boosting technique for online banking network intrusion detection. This technique combines multiple weak classifiers to create a strong classifier capable of accurately detecting intrusion attempts in real-time. The study's findings will help to determine whether the Homogeneous Boosting technique is a viable solution for improving the accuracy and efficiency of online banking network intrusion detection. Given the escalating sophistication of cyber threats targeting online banking systems, evaluating the efficacy of the homogenous boosting technique in intrusion detection addresses a critical need. The study's findings can lead to enhanced security protocols, thereby fortifying the resilience of online banking networks against evolving threats.

By providing a robust evaluation of the homogenous boosting technique, the study equips financial institutions with a potent tool to augment their cybersecurity arsenal. This empowerment is pivotal in safeguarding sensitive financial information, instilling confidence

in users, and ensuring the stability of the financial sector. Through its specialized focus on the homogenous boosting technique, the study contributes to the cutting edge of intrusion detection methodologies. This pioneering spirit fosters innovation and propels the field forward, potentially inspiring novel approaches to cybersecurity in online banking. Recognizing the distinct challenges posed by online banking environments, the study tailors its investigation to this specific domain. This bespoke approach acknowledges the unique intricacies of securing financial transactions, crafting solutions finely tuned to the intricacies of the online banking landscape. In an era where online banking has become integral to financial transactions, the study's focus on bolstering security measures couldn't be timelier. Its findings are directly applicable and relevant, providing immediate insights and potential solutions to the banking industry grappling with the omnipresent threat of cyberattacks. Positive outcomes from the study could permeate the banking industry, potentially influencing security practices across financial institutions worldwide. The adoption of the homogenous boosting technique may represent a paradigm shift in the approach to online banking security, potentially setting new industry standards. Beyond its practical implications, the study enriches the academic landscape by contributing empirical evidence and insights into the performance of the homogenous boosting technique in intrusion detection. This body of knowledge serves as a cornerstone for future research endeavours, inspiring further exploration in the domains of cybersecurity and machine learning. In a world fraught with cyber risks, effective intrusion detection systems are pivotal in mitigating risks associated with online banking. The study's endeavours, aimed at enhancing the performance of the homogenous boosting technique, directly contribute to financial protection and risk management strategies, ultimately safeguarding the interests of online banking users. In summation, the study on the "Performance Evaluation of Homogenous Boosting Technique for Intrusion Detection in Online Banking" carries far-reaching implications. It transcends the

confines of a singular field, resonating in the realms of cybersecurity, financial stability, technological advancement, and academic enlightenment. Its contributions have the potential to reverberate across industries, shaping the future landscape of online banking security.

1.6 Scope of the Study

The scope of this study is to evaluate the performance of the Homogeneous Boosting technique for intrusion detection in online banking. The study will focus on the application of the technique to a simulated online banking network environment to determine its effectiveness in detecting intrusion attempts in real-time.

The study will utilize a variety of datasets and evaluation metrics to assess the performance of the Homogeneous Boosting technique. The datasets will include both synthetic and real-world data to ensure that the results are representative of the actual conditions encountered in online banking networks. The evaluation metrics will include accuracy, sensitivity, specificity, Kappa statistic and AUROC to provide a comprehensive assessment of the technique's performance.

The study will also consider the potential limitations of the Homogeneous Boosting technique, including the impact of data quality, the presence of noise and outliers, and the need for continuous updating of the model. The scope of the study is limited to the evaluation of the Homogeneous Boosting technique for online banking network intrusion detection. The study will not consider other intrusion detection techniques or alternative applications of the Homogeneous Boosting technique in cybersecurity. The study will also not address broader issues related to online banking security, such as authentication, authorization, and access control.

Overall, the scope of the study is to provide a comprehensive evaluation of the Homogeneous Boosting technique's performance for online banking network intrusion detection and to identify areas for further research and optimization.

1.7 Limitations of the Study

There are several limitations to consider for the proposed study on the Performance Evaluation of Homogeneous Boosting Technique for Intrusion Detection in Online banking. Some of these limitations include:

1. Limited Data Availability: The availability of large datasets containing real-world online banking intrusion attempts may be limited due to the sensitivity and privacy of the data. This may impact the study's ability to accurately assess the performance of the Homogeneous Boosting technique.
2. Limited Generalizability: The study's findings may not be generalizable to all online banking network environments. The Homogeneous Boosting technique's performance may vary depending on the specific network architecture, hardware and software configurations, and other factors unique to each online banking institution.
3. Finance was an impediment to gathering facts and datasets.
4. Computational Requirements: The Homogeneous Boosting technique is a computationally intensive machine learning algorithm that may require significant computing resources to implement. The study may not be able to evaluate the technique's performance under realistic computational constraints.
5. Limited Impact: The study's findings may have limited practical impact if the Homogeneous Boosting technique's performance does not significantly improve on existing intrusion detection systems.

6. Imbalanced Class Distribution: In intrusion detection datasets, the occurrence of normal traffic is often much higher than that of actual intrusions. This class imbalance can affect the performance evaluation, as the model may become biased towards the majority class.
7. Dynamic and Evolving Threat Landscape: The nature of cyber threats is constantly changing, with new attack vectors and techniques emerging over time. The study may not capture the full spectrum of potential threats, and the effectiveness of the chosen boosting technique may vary in different threat scenarios.
8. Overfitting and Model Complexity: Complex models like boosting algorithms have the potential to overfit to the training data, especially if hyperparameters are not appropriately tuned. Balancing model complexity with performance is a crucial consideration.
9. Evolving Network Infrastructure: The study may not account for changes in the network infrastructure over time, such as hardware upgrades, changes in protocols, or the adoption of new security measures, which can impact the effectiveness of the intrusion detection system.
10. Real-time Considerations: The study may not address real-time constraints, as online banking systems require timely detection and response to potential intrusions. The processing time of the proposed technique should be evaluated in this context.

Despite these limitations, the proposed study still has the potential to provide valuable insights into the performance of the Homogeneous Boosting technique for online banking network intrusion detection.

1.8 Outline of the Thesis

Chapter one – This gives an introduction of the research study and the theoretical

background.it comprises of the statement of the problem, aim and objectives, scope and limitations pertaining to the study and the organization of the research.

Chapter two – This chapter covers an extensive review of scholarly articles, papers and contributions on the research discussed.

Chapter three– The methodology and system analysis design are exemplified in this chapter. This includes the process of extraction, data collections, processing of data and for the algorithms to be used in this research.

Chapter four - This is the system implementation and execution of processes stated in chapter three and documentation of findings.

Chapter fiveis on the summary and, conclusion of the research thesis, recommendations and future works are also considered.

1.9 Operational Definition of Terms

i. **Algorithm:** a finite sequence of well-defined, computer-implementable instructions, typically to solve a class of specific problems or to perform a computation.

ii. **Dataset**: corresponding to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question.

iii. **Intrusion Detection System**: is a tool or software that works with your network to keep it secure and flag when somebody is trying to break into your system.

iv. **Machine Learning:** the study of computer algorithms that can improve automatically through experience and using data.

v. **Online Banking:** allows a user to conduct financial transactions via the internet. online banking is also known as internet banking or web banking.

vi. **Un-supervised learning**: the use of artificial intelligence (ai) algorithms to identify patterns in data sets containing data points that are neither classified nor labelled.

Endnotes

[1.] Nedumaran, Dr G., and M. Baladevi. "Impact on customer perceptions of green banking process with special reference in Rajapalayam Taluk." (2020).

[2.] Lesjak, Dušan. "Electronic Banking: Presence and Trends." In *MIC 2019: Managing Geostrategic Issues; Proceedings of the Joint International Conference, Opatija, Croatia, 29 May–1 June 2019*, pp. 111-120. University of Primorska Press, 2019.

[3.] Eurostat (2018), *Internet banking on the rise.* https://ec.europa.eu/eurostat/web/products-eurostatnews/-/DDN-20180115-1

[4.] Mawutor, John KM, et al. "Fraud and Performance of Deposit Money Banks." Accounting and Financial Research 8.2 (2019): 202-213.

[5.] Eneji, Samuel Eneeji, Maurice UdieAngib, Walter Eyong Ibe, and Kelechukwu ChimdikeEkwegh. "A study of electronic banking fraud, fraud detection and control." ***International Journal of Innovative Science and Research Technology*** **4**, no. 3 (2019): 708-711.

[6.] Tama, Bayu Adhi, and Sunghoon Lim. "Ensemble learning for intrusion detection systems: A systematic mapping study and cross-benchmark evaluation." *Computer Science Review* 39 (2021): 100357.

[7.] Gupta, Rajesh, Sudeep Tanwar, Sudhanshu Tyagi, and Neeraj Kumar. "Machine learning models for secure data analytics: A taxonomy and threat model." *Computer Communications* 153 (2020): 406-440.

[8.] Rajasekaran, M., and A. Ayyasamy. "A novel ensemble approach for effective intrusion detection system." In *2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM)*, pp. 244-250. IEEE, 2017.

CHAPTER TWO

LITERATURE REVIEW

This portion of the work contains a review of past relevant work done on curbing intrusion in online banking. It also notes that various works have been completed in this area; yet, just for this study, the review is limited to only the important and closely connected materials.

2.1 Conceptual Review

2.1.1 Online Banking

Online banking, also known as internet banking or e-banking, refers to the electronic platform that allows customers to perform various financial transactions over the internet. It has become an integral part of modern banking, providing convenience and accessibility to millions of users worldwide. This part explores the evolution of online banking, its benefits, challenges, and the factors influencing its adoption[1].The concept of online banking emerged in the late 20th century with the proliferation of the internet. In the mid-1990s, financial institutions began offering basic online services, such as checking balances and viewing transaction history through web portals. Advancements in internet technologies, encryption protocols, and the development of secure communication channels facilitated the expansion of online banking services. This evolution led to the integration of more sophisticated features, including electronic funds transfers, bill payments, and investment management.

The advent of smartphones further transformed online banking. Mobile banking applications provided customers with the ability to conduct financial transactions from their handheld devices, leading to a surge in mobile banking adoption. Online banking offers customers the flexibility to manage their finances anytime, anywhere. They can perform transactions, check account balances, and pay bills without the need to visit a physical branch. For both customers and financial institutions, online banking reduces the costs associated with

physical infrastructure, such as brick-and-mortar branches and paper-based transactions. This translates to potential cost savings for all parties involved. Online banking platforms provide a range of services beyond basic transactions, including loan applications, investment management, and real-time financial insights. These additional features empower customers to have greater control over their financial portfolios. One of the primary concerns associated with online banking is security. Customers are vulnerable to various cyber threats, including phishing attacks, malware, and identity theft. Financial institutions must employ robust security measures to protect customer data. Not all individuals have equal access to technology or the internet, leading to a digital divide. This can exclude certain demographics from benefiting fully from online banking services. Establishing trust between customers and financial institutions is crucial for the success of online banking. Incidents of security breaches or system failures can erode customer confidence. The perceived ease with which customers can navigate and utilize online banking platforms significantly influences their adoption. Intuitive interfaces and user-friendly features enhance the overall experience. Customers' perception of the security measures implemented by financial institutions is a critical factor in their decision to adopt online banking. Clear communication about security protocols is essential. Age, education level, and technological proficiency are key demographic factors influencing online banking adoption. Younger, tech-savvy individuals are more likely to embrace these services. The integration of artificial intelligence and machine learning technologies is poised to enhance the personalization and security of online banking services. Blockchain technology and cryptocurrencies have the potential to revolutionize the financial industry, potentially impacting the future landscape of online banking Blockchain technology and crypto technologies are closely related concepts that form the foundation of cryptocurrencies like Bitcoin. They have also found applications in a wide range of industries beyond finance. Blockchain is a distributed ledger technology that

records transactions across multiple computers in a way that ensures the security, transparency, and immutability of the data.

It operates on a decentralized network of nodes (computers), which means no single entity has control over the entire network. This decentralized nature makes it resistant to manipulation and fraud.

Online banking has evolved from a rudimentary service to a sophisticated platform offering a wide array of financial services. While it brings immense convenience and efficiency, challenges such as security risks and the digital divide persist. Factors like ease of use, security, and demographic considerations play pivotal roles in its adoption. As technology continues to advance, online banking is likely to undergo further transformation, shaping the future of financial services globally.

Today, the service business is concentrating on automating transactions via the use of cutting-edge technology solutions. Banks, in particular, play an important role in a country's economic development. Internet banking (also known as e-banking) is one of the technical applications that has triggered a revolution in the financial business. Physical branch banking has now been supplanted by e-banking, in which consumers physically visiting the branch[1]. This transition to e-service has given various benefits to both banks and clients, including tailored services, transaction security, transaction speed, and overall enhanced quality of service[2]. Online banking as a payment system that allows clients of a bank or other financial institution to execute a variety of financial transactions via the financial organization's website often known as internet banking, e-banking, or virtual banking[3]. The convenience of internet banking is a significant benefit. Paying bills and moving payments between accounts are simple banking procedures that may be completed 24 hours a day, seven days a week, wherever the customer desires. However, it was pointed out that for new online banking users, using online systems might offer obstacles that prohibit transactions from being performed,

which is why some customers prefer face-to-face transactions with a teller. Online banking is primarily created by both commercial and public sector banks to accomplish two goals, the primary goal is to promote customer convenience by meeting customer needs such as viewing of account details via the internet, statement information, bill payment via the internet, money transfers, applying for accounts and e-clearance such as rent, loan payment, and so on. The second goal is to lower operating costs. So many people still find it difficult to embrace the online banking paradigm[3]. A research model to explore the key factors affecting consumers' willingness to use online banking was developed. This process has two steps. To begin, the decision-making trial and evaluation laboratory (DEMATEL) and analytic network process (ANP) were utilized to investigate the essential variables of firms using internet banking Secondly, structural equation modeling (SEM) was used to investigate the most important aspects of consumers' actual use of internet banking. The findings revealed disparities in the criteria used by businesses and consumers. Based on the findings, businesses may change their business strategies and increase consumers' desire to use online banking[4]. Trust is the most important component for both businesses and customers. Research was carried out that analyzed Internet banking clients in the Coimbatore region in India to better understand various facets of Internet banking services and consumer concerns about security measures.

The findings of the Internet banking research project aided in the development of a preventative checklist for a variety of difficulties that may arise in the Internet banking age. It was however noticed that not enough sample size was used. More banks should be incorporated into the research as samples. In the present scenario majority of the customers are accepting online banking transactions because of its many favorable factors. It is a border less entity permitting anytime, anywhere and anyhow banking to its customers. On the other hand, it also aggravates the use of traditional banking risk [5]. There is a significant relationship

between e- banking and customer satisfaction. The customers are committed to using the service, as well as banks is able to retain the major interest of its users. Beside the enormous benefits e-banking is a difficult business and face a lot of challenges like usefulness, security, and privacy. Many customers think that it is not easy to use online banking system as people want their money to be safe and secure. To overcome with the issues, the banks should provide more facilities and convenience to the customers by taking all steps and measures to make online transactions safer and secure for the customers. But e-banking is a difficult business and banks face a lot of challenge[6].

2.1.2Types of Online Banking

In the fast-paced digital era, online banking has become an integral part of our financial lives. It offers a convenient and secure way to manage our finances from the comfort of our homes or on the go. Understanding the different types of online banking can empower individuals to make informed decisions about their financial management.

1. Internet-only Banks

Internet-only banks, also known as virtual banks or digital banks, operate solely online. They have no physical branches, which allows them to offer higher interest rates on savings accounts and lower fees for services. These banks often rely on partnerships with ATM networks to provide customers with free access to cash withdrawals. As the name suggests, internet-only banks do not have physical locations. All transactions and interactions with the bank are conducted through online platforms, including websites and mobile apps. Without the expenses associated with maintaining physical branches and staffing them, internet-only banks can offer higher interest rates on savings accounts and lower fees for services. These banks prioritize their online presence, offering user-friendly interfaces, secure login processes, and a wide range of services such as account management, bill payment, fund transfers, and more. These banks often leverage cutting-edge technology to provide a seamless and secure

online banking experience. Internet-only banks prioritize security, employing robust encryption, multi-factor authentication, and other measures to protect customer information. Examples of internet-only banks include Ally Bank, Chime, and Simple. They are known for their user-friendly interfaces, competitive interest rates, and robust online security measures.

2. Traditional Banks with Online Services

Most traditional brick-and-mortar banks now offer online services to their customers. This includes features such as online account management, bill payment, fund transfers, and mobile banking apps. These services aim to provide the convenience of online banking while still maintaining physical branches for in-person services. Examples of traditional banks with comprehensive online services include First Bank, Wema Bank, and Access Bank. Customers can enjoy the benefits of both online and in-person banking experiences. In addition to their physical presence, traditional banks provide secure and user-friendly online banking platforms. These platforms enable customers to perform various banking activities, including account management, bill payment, fund transfers, and more. These banks typically offer mobile apps that allow customers to access their accounts, make mobile check deposits, transfer funds, and perform other banking tasks via their smartphones or tablets. Customers of traditional banks with online services can use the bank's ATMs for cash withdrawals and deposits. Some banks may also have partnerships with ATM networks for added convenience. These banks often offer a wide range of financial products and services, including savings and checking accounts, loans, mortgages, credit cards, investment options, and more.

3. Credit Unions

Credit unions are similar to traditional banks, but they are not-for-profit institutions owned by their members. They often offer competitive rates and low fees. Like traditional banks, credit unions now provide robust online banking services. Members can access their accounts, pay bills, and perform various transactions through secure online platforms.

Notable credit unions with strong online banking services include Navy Federal Credit Union and Alliant Credit Union.

4. Mobile Banking Apps

Mobile banking apps are a subset of online banking, designed specifically for smartphones and tablets. They provide a streamlined and user-friendly way to manage finances on the go. These apps often include features such as mobile check deposit, fund transfers, bill pay, and account alerts. Leading mobile banking apps include those offered by major banks like First Bank and Access Bank, as well as standalone services like PiggyVest and Opay. Many apps enable users to deposit checks by simply taking a picture of the front and back of the check. This saves a trip to a physical branch or ATM. Users can pay bills directly through the app, set up recurring payments, and even receive alerts for upcoming due dates. Mobile banking apps facilitate easy and secure transfers of funds between different accounts, both within the same bank and to external accounts. Apps often include features to help users locate nearby ATMs and branches of their bank. Users can set up alerts for various account activities, such as low balance notifications, large withdrawals, and more. banking apps implement robust security measures including encryption, biometric authentication (like fingerprint or facial recognition), and multi-factor authentication.

5. Robo-Advisors

While not traditional banks, robo-advisors are a form of online financial management. They provide automated, algorithm-driven investment services with little to no human supervision. Users answer questions about their financial goals and risk tolerance, and the robo-advisor then constructs and manages a portfolio on their behalf. Robo-advisors are accessible to a wide range of investors, regardless of their wealth or investment experience. They usually charge lower fees compared to traditional human financial advisors, making them an economical option for investors. Robo-advisors employ sophisticated algorithms to build diversified portfolios that align with the investor's risk tolerance and goals. They eliminate emotional biases from investment decisions, which can lead to more disciplined and rational investment strategies. Investors can set up automatic contributions and withdrawals, allowing for a hands-off approach to managing investments. Robo-advisors often provide clear fee structures and transparent reporting on portfolio performance. While robo-advisors offer automated investment management, they lack the personalized advice and human touch provided by traditional advisors. Some robo-advisors may have limited options for highly customized investment strategies. Like all investments, robo-advised portfolios are still subject to market risks and fluctuations. Reliance on technology means that occasional outages or technical issues may occur. Examples of popular robo-advisors include Betterment, Wealthfront, and Robinhood.

6. Cryptocurrency Exchanges

For those interested in the world of cryptocurrencies, online platforms known as cryptocurrency exchanges facilitate the buying, selling, and trading of digital currencies. These platforms often come with their own digital wallets and various tools for managing cryptocurrency portfolios. Well-known cryptocurrency exchanges include Coinbase, Binance, and Kraken. Online banking has evolved into a diverse ecosystem with options to suit various financial needs and preferences. As technology continues to advance, the landscape of online

banking is likely to evolve, providing even more options for consumers in the future. Exchanges offer various trading pairs, which represent the combinations of cryptocurrencies that can be traded against one another. For example, Bitcoin (BTC) to Ethereum (ETH) or Bitcoin to USD. Users can execute market orders, which are executed immediately at the current market price, or set limit orders, which are executed only when the price reaches a specific level. Many exchanges provide digital wallets for users to store their cryptocurrencies securely. However, it's generally recommended to store significant amounts of cryptocurrencies in more secure hardware wallets. Established exchanges tend to have higher liquidity, meaning there are more buyers and sellers available, resulting in more competitive prices. Reputable exchanges implement robust security measures like two-factor authentication (2FA), cold storage for funds, and encryption to protect user accounts and assets.

2.1.3Types of Network Intrusion

Online banking networks can face various types of intrusions and cyber threats. These intrusions aim to compromise the security of the network, potentially leading to unauthorized access, data breaches, or financial losses. Here are some common types of online banking network intrusions:

1. **Phishing Attacks**: Phishing attacks involve sending deceptive emails or messages that appear to be from legitimate sources, such as a bank. These emails often contain links or attachments that, when clicked, lead to fake websites designed to steal login credentials or personal information. Phishing attackers often disguise themselves as legitimate entities through emails, websites, or messages. They may use convincing logos, language, and formatting to create a sense of authenticity. Phishing emails often employ urgent or threatening language to create a sense of urgency. They may claim that immediate action is required, such as verifying an account or preventing a

security breach. Phishing emails may contain links to fake websites that mimic legitimate ones. These websites are designed to steal login credentials or install malware on the victim's device. Phishing attacks often rely on psychological manipulation to exploit human behaviour. This can include tactics like exploiting trust, creating a sense of urgency, or appealing to emotions.

2. **Malware Infections**: Malware, short for malicious software, includes viruses, worms, trojans, and other types of malicious code. When a user unknowingly downloads or interacts with this software, it can compromise their system and potentially allow attackers to gain unauthorized access to online banking accounts. Malware is a broad term that encompasses various types of malicious software, including viruses, worms, trojans, ransomware, spyware, adware, and more. Malware may have various objectives, including: Stealing sensitive information like login credentials, financial data, or personal information, Ransomware encrypts files and demands a ransom for decryption, Causing system or network failures, leading to downtime, Gathering information for intelligence or corporate espionage, Creating networks of infected devices for coordinated attacks or spam distribution.
3. **Man-in-the-Middle (MitM) Attacks**: In a MitM attack, an attacker intercepts and potentially alters the communication between a user's device and a website. This can allow the attacker to capture sensitive information, including login credentials and financial details. In a MitM attack, the attacker positions themselves between the two parties in a communication exchange. This can occur in various scenarios, including Wi-Fi networks, public hotspots, or compromised routers. The parties involved in the communication are typically unaware of the presence of the attacker. They believe they are communicating directly with each other. The attacker simply listens in on the communication to gather sensitive information, such as passwords, financial details,

or personal conversations. The attacker may alter the content of the communication, inserting or changing information. This can lead to misinformation or fraudulent activity. The attacker may seize control of an established session (such as a web session) between the two parties. This allows them to impersonate one party and potentially gain unauthorized access. This attack downgrades a secure HTTPS connection to an unencrypted HTTP connection, allowing the attacker to intercept and manipulate the data. MitM attacks pose a significant threat to the confidentiality and integrity of communications.

4. **SQL Injection**: SQL injection attacks target vulnerabilities in web applications that interact with databases. Attackers inject malicious SQL code into input fields, exploiting vulnerabilities to gain unauthorized access to the database or execute unauthorized commands. SQL injections typically occur when web applications do not properly validate or sanitize user input before interacting with the database. This allows attackers to inject their own SQL code. Classic SQL Injection involves injecting malicious SQL code directly into input fields, potentially allowing unauthorized access to the database, blind SQL Injection involves attackers exploit vulnerabilities that don't display the results of the injected code, making it more challenging to execute, but still possible while in Time-Based Blind SQL Injection Delays in database responses are used to infer whether the injected code is executed or not.

5. **Cross-Site Scripting (XSS)**:XSS attacks involve injecting malicious scripts into web pages that are viewed by other users. These scripts can steal cookies, session tokens, or other sensitive information, potentially compromising online banking sessions. The malicious script is permanently stored on the target server, often in a database or file,

and served to users whenever they access a particular page or resource. The malicious script is embedded in a URL or form input and is only served to users who visit a specific URL with the injected payload. The attack occurs entirely on the client-side, manipulating the Document Object Model (DOM) of a web page after it has been loaded. Attackers can steal session cookies, allowing them to impersonate the victim and gain unauthorized access to their account.

6. **Distributed Denial-of-Service (DDoS)**:DDoS attacks flood a website or network with an overwhelming amount of traffic, rendering it inaccessible to legitimate users. While this type of attack doesn't directly result in unauthorized access, it can disrupt online banking services and create opportunities for other attacks. DDoS attacks operate by coordinating a large number of compromised devices, often spread across the internet, to simultaneously send a flood of data packets to a specific target. The primary goal is to render the targeted service or network unavailable to legitimate users. DDoS attacks are often executed using botnets, which are networks of compromised devices (computers, IoT devices, etc.) controlled by a single entity.
7. **Credential Stuffing**: In a credential stuffing attack, attackers use previously obtained usernames and passwords (often from breaches on other websites) to attempt unauthorized access to online banking accounts. They rely on the likelihood of users reusing passwords across multiple platforms. The main goal of credential stuffing attacks is to gain unauthorized access to online accounts, which can lead to various malicious activities such as data theft, financial fraud, or account takeover. Attackers obtain lists of usernames and passwords from previous data breaches, black markets, or other sources. They use automated tools or scripts to rapidly input these stolen credentials into login pages of various online services. Since many users reuse

passwords across multiple platforms, attackers are often successful in gaining unauthorized access.

8. **Social Engineering**: Social engineering attacks manipulate individuals into revealing sensitive information or performing actions that compromise security. This can include tactics like impersonation, pretexting, or baiting.
9. **Insider Threats**: Insider threats involve individuals within an organization who misuse their privileges to gain unauthorized access or compromise security. This could be employees, contractors, or partners with access to the online banking network.
10. **Zero-Day Exploits**: Zero-day exploits target vulnerabilities in software or hardware that are not yet known to the vendor. Attackers exploit these vulnerabilities before a patch or update is available, potentially gaining unauthorized access. Online banking networks employ a combination of security measures, including encryption, firewalls, multi-factor authentication, and intrusion detection systems, to mitigate the risks associated with these types of intrusions.

2.1.4 Internet Fraud

Internet fraud, also known as online fraud or cyber fraud, refers to deceptive practices carried out over the internet with the intent of financial gain or theft of personal information. It encompasses a wide range of fraudulent activities, including phishing, identity theft, online scams, and various forms of cybercrime. Phishing attacks involve sending deceptive emails or messages that masquerade as legitimate sources, aiming to trick recipients into revealing sensitive information like passwords or credit card details. These attacks often lead to financial losses and identity theft. Identity theft occurs when a perpetrator steals personal information, such as Social Security numbers or bank account details, and uses it for fraudulent activities, including unauthorized transactions and opening fraudulent accounts.

Online scams encompass a broad range of fraudulent schemes, such as lottery scams, advance-fee fraud, and fake online marketplaces. These scams prey on victims' trust or vulnerabilities to extract money or valuable information. Cyberextortion involves threatening a victim with the release of sensitive information or disruption of services unless a ransom is paid. Ransomware, a form of cyberextortion, involves encrypting a victim's files or systems and demanding payment for their release. Human vulnerabilities, such as gullibility, lack of awareness, or a desire for quick financial gains, are often exploited in internet fraud. Social engineering techniques play a significant role in deceiving victims. Flaws in authentication mechanisms, such as weak passwords or inadequate multi-factor authentication, provide opportunities for fraudsters to gain unauthorized access to accounts or systems.

Unpatched or outdated software, as well as insecure coding practices, can be exploited by cybercriminals to gain unauthorized access or deploy malicious software.

Victims of internet fraud often suffer significant financial losses, including unauthorized charges, stolen funds, or even complete financial ruin in severe cases. For businesses, falling victim to internet fraud can lead to reputational damage, loss of customer trust, and potentially legal consequences, affecting long-term viability. Internet fraud can have a profound emotional and psychological impact on victims, leading to stress, anxiety, and a sense of violation.

Raising awareness about common internet fraud techniques and providing education on best practices for online safety is crucial in preventing victimization. Implementing strong authentication methods, such as multi-factor authentication, and ensuring proper authorization protocols can significantly reduce the risk of unauthorized access. Regularly updating and using antivirus and anti-malware software helps in detecting and mitigating potential threats from malicious software. Collaboration between law enforcement agencies and regulatory bodies is essential in investigating and prosecuting internet fraud cases, as

well as implementing and enforcing regulations to protect consumers. As technology advances, new forms of internet fraud, such as deepfake scams and AI-driven attacks, are likely to emerge, requiring innovative countermeasures. While blockchain technology holds promise for enhancing security in financial transactions, it also presents new challenges in combating fraud, particularly in the realm of cryptocurrencies. Internet fraud represents a pervasive and evolving threat in the digital age, encompassing various deceptive practices with significant financial and emotional consequences for victims. Understanding the types, vulnerabilities, and impact of internet fraud is crucial in developing effective prevention and mitigation strategies. As technology continues to advance, ongoing research and collaboration between industry stakeholders, law enforcement, and regulatory bodies will be essential in staying ahead of emerging threats.

The FBI described fraud as the use of Internet services or software with an Internet connection to deceive or otherwise exploit people. Internet crime schemes steal millions of dollars from victims each year and continue to haunt the Internet through a variety of tactics[7]. Electronic banking frauds as scams linked with electronic banking that are committed via ATM, POS, internet, and mobile banking systems. They went further to point out that fraud could be pulled off by impersonation, phishing(which is a type of social engineering where an attacker sends a fraudulent message designed to trick a human victim into revealing sensitive information to the attacker or to deploy malicious software on the victim's infrastructure like ransomware.), hacking, trojan horse, etc. Fraud can be identified by detecting unusual patterns in financial transactions, monitoring suspicious transactions (i.e. activities when the user is unclear what to do), keeping track of new fraud schemes [8]. The World Wide Web spawned security fraud on the Internet. This sort of deception is defined as the purposeful manipulation of the "securities market for profit." Market manipulation, "fraudulent offers of securities," and unlawful touting are the three principal forms of such

scams, according to the experts[9]. The first type of fraud involves the manipulation of stock prices on the Internet market by an entity or individual who alters the natural flow of demand and supply. The second type of Internet securities fraud is focused on the development of particular websites[9]. These portals are "particularly designed to offer securities falsely". The designers of such websites sell their goods by promising unreasonably high profits to stakeholders. However, the product does not exist, and the funds are sent directly to the criminals' accounts. In some of these schemes, earlier investors earn a portion of the returns generated by the future contributors' expenses. Unlawful touting, on the other hand, is the dissemination of misleading information about an already existent firm over the internet. In such a scheme, the culprits promote a business's product in return for money from that business[10]. For financial organizations, the unfettered flow of digital information means that the backdoor is always theoretically vulnerable to loss, as in the case of Russian hacker Vladimir Levin, who stole $10 million from Citibank, and the funds were moved to bank accounts all through the globe[11]. Clearly, internet fraud is a serious concern for the financial services sector; nonetheless, financial institutions are progressively providing their consumers with online banking services[12].Banks, then, require a continuous improvement mentality that evaluates and retests the bank's e-fraud defenses, because hackers usually exploit the flaw quickly and silently, and are long gone before the bank or its customers discover the problem[13]. Phishing is an indirect fraud that also threatens customers and banks. It may easily duplicate the original daily deal email, substituting legitimate links with malicious ones, such as key loggers that record passwords for banking or other sensitive sites[14].

2.1.5 Intrusion Detection Systems

Intrusion detection systems consists of two main types, Network-based (NIDS) and Host based (HIDS) intrusion detection systems [15].

2.1.5.1 Network-based Intrusion Detection System (NIDS)

A Network-Based Intrusion Detection System (NIDS) is a critical component of network security, designed to monitor and analyse network traffic for suspicious or malicious activities. It plays a pivotal role in identifying and mitigating potential security breaches, providing an additional layer of defence against cyber threats. It uses a permissive interface to search network traffic for signs of attacks. Only packets that were sent over the network segment to which it is connected can be seen by it. If network communication is encrypted, NIDS cannot scan the protocols or content. Additionally, NIDS, also called as "Wireless Intrusion Prevention System" at several manufacturers, analyses network packets at all OSI (Open System Interconnection) layers[16].

2.1.5.2 Host-Based Intrusion Detection System (HIDS)

This is a piece of software that relies on the documentation, files, and activity trees of the operating to operate. It won't scan the entire network; it will simply scan the independent hosts or devices on it. A HIDS keeps track of the packets coming from the tool or device and notifies the administrator if any suspicious activity is found. A notification was sent to the administrator if any changes were made. Additionally, it can function in encrypted contexts and has a platform-specific, high overhead OS and greater management (deployment expenses). Remarkably different from one another, the two varieties of intrusion detection systems work effectively together. The top IDS tools incorporate both strategies into a single administration console. In this manner, the user receives thorough protection, ensuring that they are protected from as many dangers as possible[16]. Responses from intrusion detection systems: There are two groups of responses for intrusion detection systems. Both are reactive and passive systems.

i. **IDS passive system**: The IDS in this system recognizes a possible security breach, logs the data, and issues an alarm.

ii. **IDS Reactive System:** As a result of the suspicious behavior, the IDS on this system either logs off the user or reprograms the firewall to prevent network traffic from the allegedly hostile source[17].

2.1.6 Intrusion Detection Systems Techniques

There are two fundamental IDS methods that are used to identify, Misuse and anomaly detection systems are examples of such. Misuse detection system (pattern or signature detection Monitoring network traffic and analyzing it are examples of detection against a particular, predetermined attack. IDSs are nearly all signature-based, also referred to as knowledge-based. It is set up to do a certain packet or piece of data should be interpreted contained as an attack in those packets. most analyses of signatures Systems operate using straightforward pattern matching algorithms. The IDS typically just searches a stream for a substring network packet carrying a variety of data. There are downsides because they are unable to recognize new attacks, experiencing false warnings and require new programming to recognize each new pattern.

Anomaly detection system (Behavior detection): It serves as a noise characterization model for how the network is typically used. Any noise that stands out has been interpreted as an incursion action. Techniques for behavior-based intrusion detection make the assumption that an intrusion may be found by looking for a change in the users' or the system's usual or expected behavior. It can recognize new attacks, which is one of its advantages. They would produce a lot of false alerts as a downside, which would reduce the IDS's effectiveness. Additionally, this detection method has employed a number of techniques. For instance, statistical methods, pattern-predicting methods, and machine learning.

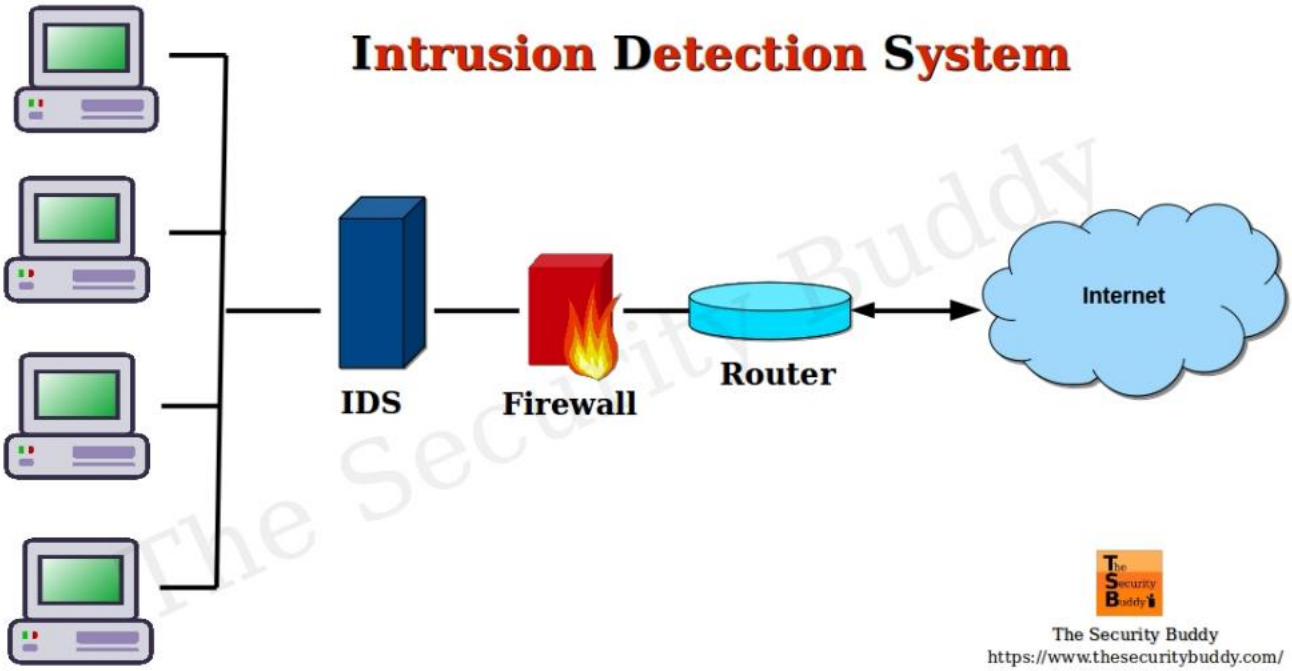


Figure 2.1: Intrusion Detection System

Source: https://www.thesecuritybuddy.com/data-breaches-prevention

2.1.7 Machine Learning

Machine learning (ML) is a field of artificial intelligence (AI) that focuses on the development of algorithms and models that allow computers to learn and make predictions or decisions without being explicitly programmed. It has witnessed significant advancements in recent years, revolutionizing various industries and applications. Machines are becoming increasingly intelligent, cars now drive themselves, Alexa now understands her owner, Siri interprets sounds, and Google interprets webpages. The essential idea behind these achievements is both quantitative and mathematical. Machine learning is a data analysis technique that automates the creation of analytical models. It is a sub-field of artificial intelligence that is predicated on the concept that systems can learn from data, spot patterns, and make decisions with little or no human interaction, the emergence of statistical learning theory, which provided a mathematical framework for understanding the capabilities and limitations of learning algorithms. This period also saw the development of algorithms like

Support Vector Machines (SVMs) and Decision Trees. Machine learning was initially defined as software that learns to complete a job or make a decision automatically from data rather than having the behavior explicitly coded. Machine learning presently is not the same as machine learning in the past due to advances in computer technology. It arose from pattern recognition and the idea that computers may learn without being taught to execute certain tasks; artificial intelligence researchers sought to investigate if computers could learn from data. a science that is not new but has garnered new traction. While many machine learning techniques have been known for a long time, the capacity to automatically apply complex mathematical computations to large amounts of data repeatedly and quicker is a relatively new phenomenon[18]. Machine learning is a computer science subfield that evolved from the study of pattern recognition and computational learning theory in artificial intelligence.In recent years, the availability of large datasets and advancements in computing power have enabled the rise of deep learning. Deep neural networks, with their ability to automatically learn hierarchical representations, have achieved remarkable success in tasks like image recognition and natural language processing.It is the branch of study that enables computers to learn without being explicitly programmed. Machine learning investigates the study and development of Algorithms that can learn from and forecast data. Machine learning techniques are being used for general-purpose approaches to learning functional relationships from data (without requiring them to be defined). Machine learning is concerned with the development and evaluation of algorithms that facilitate pattern recognition. Models derived from existing data are used for recognition, classification, and prediction[18].

A computer program is said to learn from experience E in relation to a set of tasks Tand measure P, if its performance at tasks in T, as measured by P, improves experience E. Example: playing checkers

• E = the experience of playing many games of checkers

• T = the task of playing checkers

• P = the probability that the program will win the next game

• grow out of work in AI

• new capability for computers

• Examples:

• database mining

Large datasets from growth of automation/web e.g. web click data, medical records, biology and engineering. Machine Learning aims to make decisions based on unknown facts without being explicitly told how to do so supervised learning, unsupervised learning, and semi-supervised learning are the three kinds of machine learning tasks. Supervised machine learning is a sort of machine learning in which a model is "trained" on a set of labeled "training instances" before being applied to new data to make predictions. Classification problems are another term for supervised learning problems. When a model is developed on unlabeled data by detecting patterns and relationships in the data, it is known as unsupervised machine learning. Clustering difficulties are another term for unsupervised learning problems. Semi-supervised machine learning is a type of machine learning that uses labeled examples to train the model and then uses unlabeled data to refine the class boundaries. Credit card fraud detection, character recognition, speech understanding, face characterization, product suggestion, the health sector, consumer segmentation, form identification, and sign language interpretation are just a few examples of real-world Machine Learning applications. Machine Learning operates by looking for patterns in data (associated or not with given classes). Data is first turned into a representation (a set of features) that a computer can understand in all of the following applications. In order to use Machine Learning methods, text must be translated into a quantitative (or discrete) representation in NLP (natural language processing). Similarly, Computer Vision works with images to turn them into a format that a computer can

understand. Data collection and preprocessing, feature engineering, model training, and system testing are all part of a typical machine learning application[19]. Machine learning has witnessed significant growth and innovation, transforming industries and applications across the board. From early developments in statistical learning to the deep learning revolution, the field continues to evolve rapidly. Addressing challenges like data quality and interpretability will be crucial in harnessing the full potential of machine learning. With ongoing research and technological advancements, machine learning is poised to continue revolutionizing various fields and shaping the future of artificial intelligence.

Machine learning, a subset of artificial intelligence, is a rapidly evolving field that focuses on enabling computers to learn and make predictions or decisions based on data. There are several distinct approaches within machine learning, each with its own strengths, weaknesses, and applications.

2.1.8 Types of Machine Learning

This provides an in-depth exploration of the various types of machine learning, including supervised learning, unsupervised learning, reinforcement learning, semi-supervised learning, and self-supervised learning.

I. Supervised Learning

A. Definition and Process

Supervised learning is a type of machine learning where the model is trained on a labeled dataset, which means each data point is associated with a known output or target. The model learns to map inputs to outputs, making it capable of making predictions on new, unseen data. A model is trained on a labelled dataset, meaning the dataset includes input data along with corresponding correct output. The goal of supervised learning is to learn a mapping from the input data to the output, so that when presented with new, unseen data, the model can make accurate predictions or decisions.

In supervised learning, the model iteratively makes predictions on the training data and is corrected by comparing its predictions to the actual labelled outputs. Through this process, the model adjusts its internal parameters to minimize the error or the difference between its predictions and the actual outputs. This is typically done using various optimization algorithms.

B. Applications

Supervised learning is widely used in various domains, including:

1. **Image and Object Recognition**: Recognizing objects in images or classifying images into different categories. Image recognition, also known as image classification, involves the task of categorizing or labelling an entire image into predefined classes or categories. It's a form of supervised learning where a model is trained on a dataset of labelled images. The goal is to enable the model to correctly identify and assign the appropriate label to new, unseen images.

   Applications of image recognition include: Identifying objects in photographs (e.g., distinguishing between cats and dogs), Medical image analysis for tasks like diagnosing diseases from medical images (e.g., X-rays, MRI scans), Autonomous vehicles to detect and understand the environment around them. Object recognition is a broader task that involves not only identifying objects in an image but also locating and outlining their boundaries (object localization) and sometimes even recognizing multiple objects in a single image (object detection). This often involves more complex models and techniques. Augmented reality, where digital information is overlaid onto the real world in a way that interacts with recognized objects. Robotics, where robots use object recognition to interact with and manipulate objects in their environment. Video surveillance for security and monitoring purposes. Both image and object recognition have seen significant advancements with the advent of deep

learning and convolutional neural networks (CNNs), which are particularly well-suited for processing visual data. These technologies have enabled breakthroughs in areas like facial recognition, automated medical diagnoses, and more.

2. **Natural Language Processing (NLP)**: Tasks like sentiment analysis, language translation, and named entity recognition. Natural Language Processing (NLP) is a subfield of artificial intelligence and computer science that focuses on enabling computers to understand, interpret, and generate human language in a way that is valuable and useful.
3. Here are some key components and tasks within NLP:

   i) **Text processing**

   -Tokenization: Breaking text into individual words or tokens.

   -Stemming and Lemmatization: Reducing words to their base or root form.

   -Stop word Removal: Removing common, uninformative words (e.g., "the", "and").

   ii) Syntax and Grammar:

   -Parsing: Analyzing the grammatical structure of a sentence.

   -Part-of-Speech Tagging: Identifying the grammatical parts of words (e.g., noun, verb).

   iii) Semantics:

   -Word Sense Disambiguation: Determining the correct meaning of a word in a particular context.

   -Named Entity Recognition (NER): Identifying and classifying named entities (e.g., names of people, places, organizations).

   iv) Discourse and Pragmatics:

   -Coreference Resolution: Identifying when different words refer to the same entity.

   -Anaphora Resolution: Understanding pronouns in relation to their antecedents.

v) Sentiment Analysis:

-Determining the sentiment or emotion expressed in a piece of text (e.g., positive, negative, neutral).

vi) Machine Translation:

-Automatically translating text from one language to another.

vii) Question Answering:

-Understanding questions and providing relevant answers based on a given text or database.

viii) Chatbots and Conversational Agents:

-Interacting with users in a natural, conversational manner.

ix) Text Summarization:

-Automatically generating concise summaries of longer texts.

x) Information Extraction:

-Identifying specific pieces of information (e.g., dates, names) from unstructured text.

NLP is widely used in various applications, including:

Chatbots and Virtual Assistants: Providing automated customer support and assistance.

Sentiment Analysis for Social Media: Analysing public opinion on platforms like Twitter or Facebook.

Machine Translation: Services like Google Translate use NLP techniques.

Information Retrieval and Search Engines: Understanding user queries and retrieving relevant results.

Recent advances in deep learning, particularly with models like transformers, have led to remarkable progress in NLP tasks and applications, allowing computers to process and generate human language with unprecedented accuracy and fluency.

4. **Regression Analysis**: Predicting a continuous variable, such as housing prices or stock market trends. Regression analysis is a statistical method used to examine the relationship between one dependent variable (often denoted as "Y") and one or more independent variables (often denoted as "X"). It is particularly useful for understanding how changes in the independent variables are associated with changes in the dependent variable.
5. **Classification Problems**: Identifying the category or class to which a data point belongs, like spam detection in emails. In machine learning, classification is a type of supervised learning task where the goal is to predict the categorical class or label of a new data point based on its features. The output variable in a classification problem is discrete, meaning it falls into a specific category or class.

II. Unsupervised Learning

A. Definition and Process

Unsupervised learning involves training models on unlabelled data. The goal is to discover hidden patterns or structures within the data without any predefined labels.

B. Applications

Unsupervised learning has diverse applications, including:

1. **Clustering**: Grouping similar data points together based on features or characteristics.
2. **Dimensionality Reduction**: Techniques like Principal Component Analysis (PCA) help reduce the complexity of high-dimensional data.
3. **Anomaly Detection**: Identifying unusual or outlier data points.
4. **Generative Models**: Creating new data samples similar to the existing dataset.

III. Reinforcement Learning

A. Definition and Process

Reinforcement learning involves agents learning to make decisions by interacting with an environment. They receive feedback in the form of rewards or penalties based on their actions, allowing them to learn the best strategies to achieve a specific goal.

B. Applications

Reinforcement learning is applied in scenarios like:

1. **Game Playing**: Agents learning to play games by trial and error, such as AlphaGo in the game of Go.
2. **Robotics**: Teaching robots to perform complex tasks by rewarding successful actions.
3. **Autonomous Systems**: Training self-driving cars to navigate and make decisions on the road.

IV. Semi-Supervised Learning

A. Definition and Process

Semi-supervised learning combines elements of both supervised and unsupervised learning. It involves training a model on a dataset that contains both labelled and unlabelled data. This approach can improve performance, especially when acquiring labelled data is costly or time-consuming.

B. Applications

Semi-supervised learning is valuable in scenarios like:

1. **Text and Document Classification**: When it's impractical to label large volumes of text data, semi-supervised learning can enhance classification accuracy.
2. **Speech Recognition**: Utilizing a combination of labelled and unlabelled audio data to improve speech recognition systems.

V. Self-Supervised Learning

A. Definition and Process

Self-supervised learning is a relatively new paradigm that leverages the inherent structure or content within data to generate labels automatically. It involves creating tasks where the data itself provides the supervision.

B. Applications

Self-supervised learning is showing promise in areas such as:

1. **NLP**: Tasks like masked language modelling, where the model learns to predict missing words in a sentence.
2. **Computer Vision**: Techniques like contrastive learning, where the model learns to differentiate between similar and dissimilar images without explicit labels.

The various types of machine learning offer a diverse set of tools and techniques for solving different types of problems. While supervised learning is well-suited for labelled data, unsupervised learning excels in tasks with no predefined labels. Reinforcement learning enables agents to learn from interactions, and semi-supervised learning leverages both labelled and unlabeled data for improved performance. Self-supervised learning, on the other hand, is a promising area that leverages the data itself for training.

Choosing the appropriate type of machine learning depends on the nature of the data, the problem at hand, and the available resources. As the field continues to advance, hybrid approaches and novel techniques are likely to emerge, further expanding the capabilities and applications of machine learning across various domains.

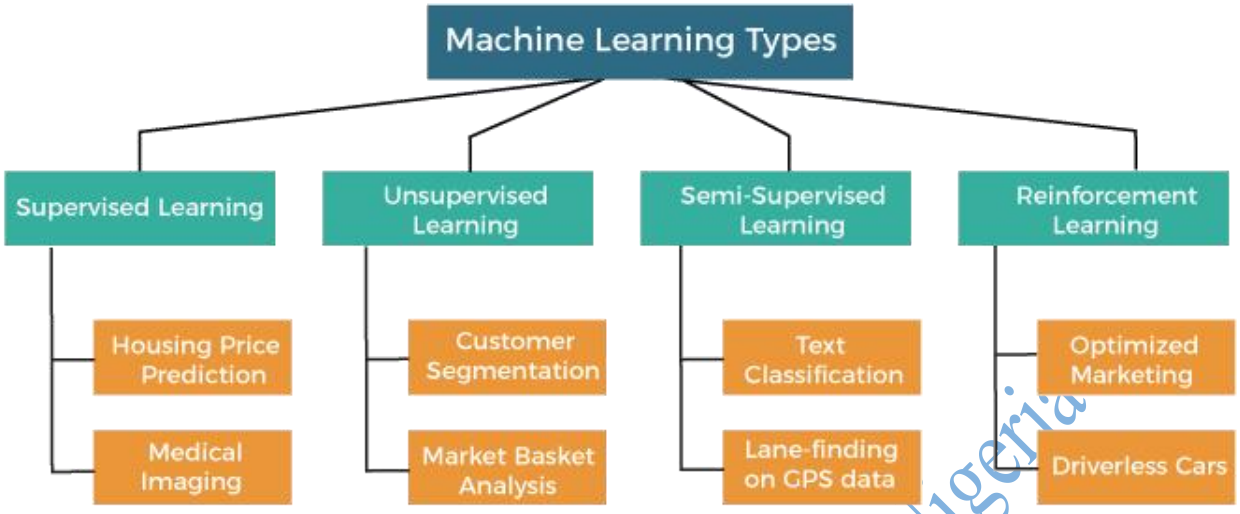


Figure 2.2: Types of Machine Learning

Source: https://www.javatpoint.com/types-of-machine-learning

2.1.9 Ensemble Learning

Ensemble learning is a powerful machine learning technique that leverages the collective intelligence of multiple models to improve predictive accuracy and generalization. By combining the strengths of different models, ensemble methods have demonstrated exceptional performance in a wide range of applications. Using intelligently mixed numerous analytics, ensemble learning is a type of hybrid learning system that aims to produce better (more accurate, more robust, etc.) results than a single insight can. Ensemble learning is based on the principle that combining the predictions of multiple models often yields better results than relying on a single model. This is analogous to the "wisdom of crowds" phenomenon, where a group of individuals collectively provides more accurate predictions than any single individual. Here, bagging, boosting, and stacking are three methods of ensemble learning that are discussed. In a bagging procedure, replacement random sets of data are generated N times, and from these subsets, nonpruned classification (decision) trees are built. Ensemble methods aim to strike a balance between reducing bias and variance. By aggregating multiple models, ensemble techniques can often achieve lower prediction errors

compared to individual models, which may suffer from high bias or high variance. Replacement is crucial since it guarantees that each potential decision tree branching will be represented in the ensemble with an equal likelihood. This is designed to offer the domain space the best possible coverage. After N iterations of this process, the categorization of each sample in the overall data set is chosen by majority vote based on the decision trees. A subpopulation of the many decision trees that were created in this way was maintained as a model for the classifier, and they can be graded for their overall accuracy (if needed). The central limit theorem, which goes along with the averaging of the several decision trees, prevents overfitting during bagging. It is possible that not enough samples are included in the decision tree(s) needed to assign a classification if the domain space is huge. Should this happen with any samples, closest neighbor or other decisioning techniques can be used to assign them. This straightforward but frequently efficient architecture can be made more resilient and path-covered by randomization, which involves introducing a modest random bias to decision tree node splits. Bagging typically produces excellent rank bias, pushing the correct classification higher up the ranking, but it does not always significantly enhance accuracy. By taking into account the outcomes from the samples in proportion to their (beneficial) impact on the accuracy of the system as a whole, boosting is an additional method of ensemble learning that enables the system to maximize its choice. In boosting, the samples are originally evenly weighted. The samples that were correctly assigned are weighted less after each algorithm iteration than the samples that were mistakenly assigned. This is like how a support vector is created, with the exception that samples in a support vector are zero-weighted unless they are close to a class boundary. Also known as layered generalization, stacking. Since this method applies to numerous models created by two or more learning algorithms for instance, a Bayesian and a decision tree approach, it is challenging to mathematically examine. Stacking moves toward the architectural strategies

connected to meta-algorithms, although stacking does not offer any particular design patterns to use. The application of output probabilities for each class and weighted voting based on adding these probabilities for each sample and each algorithm mark the conclusion of conventional stacking methods. Like bagging and boosting, stacking typically gives intelligent systems a good rank bias [20]. Ensemble methods benefit from using base learners that are diverse in terms of the algorithms they employ, their feature representations, or their initializations. This diversity allows the ensemble to capture a wider range of patterns in the data. The way predictions are combined plays a crucial role in the effectiveness of ensemble learning. Strategies include averaging, voting, weighted averaging, and more sophisticated techniques like stacking. Fine-tuning the hyperparameters of the individual models and the ensemble itself is critical for achieving optimal performance. Techniques like grid search and random search are commonly used. Ensemble methods can potentially be overfit to the training data, especially if the base learners are overly complex or highly correlated. Properly tuning the models and ensuring diversity are important mitigation strategies. Ensemble methods can be computationally intensive, particularly when training many base learners. This may limit their applicability in real-time or resource-constrained environments. Ensemble models, especially those with many base learners, can be challenging to interpret. Understanding the contributions of individual models to the final prediction can be non-trivial. Advances in online learning techniques are leading to the development of ensemble methods that can adapt and learn continuously from streaming data. Integrating transfer learning principles into ensemble methods allows models to leverage knowledge gained from one task and apply it to related tasks, potentially improving performance. Efforts are underway to make ensemble models more interpretable and explainable, which is crucial for gaining trust in critical applications like healthcare and finance. Ensemble learning stands as a powerful approach in machine learning, harnessing the collective intelligence of multiple models to

achieve superior predictive performance. From the principles of the "wisdom of crowds" to the diverse array of ensemble techniques, this field continues to evolve. Overcoming challenges related to overfitting, computational complexity, and interpretability will be crucial in realizing the full potential of ensemble learning. With ongoing research and technological advancements, ensemble methods are poised to play a pivotal role in the advancement of machine learning and artificial intelligence.

2.1.10 Simple Ensemble Techniques

In this section, a few uncomplicated yet effective methods were explored, specifically:

1. Max Voting
2. Averaging
3. Weighted Averaging

Max Voting

The max voting approach is commonly employed in classification tasks. This technique involves using multiple models to generate predictions for each data point. Each model's prediction is treated as a 'vote'. The final prediction is determined by the majority of the models.

For instance, imagine asking five colleagues to rate a movie (on a scale of 1 to 5); let's say three of them rated it as 4 and two of them gave it a 5. Since the majority favoured a rating of 4, the final rating would be recorded as 4. This can be likened to finding the mode of all the predictions.

The outcome of max voting might look like this:

Table 2.1: Max voting

Colleague 1	Colleague 2	Colleague 3	Colleague 4	Colleague 5	Final rating
5	4	5	4	4	4

Averaging

Like the max voting approach, the averaging technique involves generating multiple predictions for each data point. In this method, the predictions from all models are averaged to derive the final prediction. Averaging is applicable for generating predictions in regression scenarios or for computing probabilities in classification tasks.

For instance, in the given scenario, the averaging method would calculate the mean of all the values.

i.e. (5+4+5+4+4)/5 = 4.4

Table 2.2: Averaging

Colleague 1	Colleague 2	Colleague 3	Colleague 4	Colleague 5	Final rating
5	4	5	4	4	4.4

Weighted Average

This approach is an extension of the averaging technique. Each model is assigned a specific weight, which determines its influence on the prediction. For example, if two of your colleagues are seasoned critics, whereas the others lack expertise in this domain, the input from these two knowledgeable individuals carries more significance compared to the rest.

The result is computed as [(*50.23) + (4*0.23) + (*50.18) + (4*0.18) + (4*0.18)] = 4.41.

Table 2.3: Weighted Average

Colleague 1	Colleague 2	Colleague 3	Colleague 4	Colleague 5	Final rating
weight	0.23	0.23	0.18	0.18	
rating	5	4	5	4	4.41

2.1.11 Fuzzy Logic

Fuzzy logic is a mathematical framework that allows for the representation of uncertainty and imprecision in decision-making. It extends classical binary logic by introducing degrees of truth, which enables reasoning in situations where information is vague or ambiguous. Fuzzy logic introduces the concept of degrees of truth, allowing statements to be partially true or partially false. This contrasts with classical logic, which assumes binary true/false values. Membership functions are a key component of fuzzy logic. They assign degrees of membership to elements of a set, indicating the extent to which an element belongs to that set. Fuzzy rules define relationships between inputs and outputs in terms of linguistic variables and their associated membership functions. Fuzzy inference combines these rules to make decisions or predictions. The Mamdani-type fuzzy system employs linguistic variables, fuzzy rules, and fuzzy inference to make decisions or control systems. It is widely used in applications involving human-like reasoning. Sugeno-type fuzzy systems use linear or nonlinear mathematical functions to represent the relationship between inputs and outputs. This type is often used for modelling systems where precise mathematical relationships can be defined. Hybrid fuzzy systems combine fuzzy logic with other computational techniques, such as neural networks or genetic algorithms, to enhance performance in specific applications. Fuzzy clustering algorithms, such as Fuzzy C-Means (FCM), allow for the classification of data points into clusters with varying degrees of membership. Fuzzy control systems use fuzzy logic to control complex, nonlinear systems, where precise mathematical modelling may be impractical or too complex. Fuzzy pattern recognition techniques apply fuzzy logic to classify data patterns based on their similarity to known prototypes. Fuzzy logic is extensively used in control systems for various applications, including industrial processes, automotive systems, and consumer electronics. In healthcare, fuzzy logic aids in decision-making for diagnosis, treatment planning, and patient monitoring, particularly in

situations involving uncertainty. Fuzzy logic is employed in tasks like image segmentation, object recognition, and edge detection, where handling uncertainty and imprecision is crucial. Complex fuzzy systems can be challenging to interpret, which may hinder their acceptance in critical applications where transparency is crucial. Obtaining accurate and comprehensive linguistic rules and membership functions can be a labour-intensive task, requiring expert knowledge in the domain. Fuzzy systems may face challenges in scaling to handle large datasets or complex, high-dimensional spaces. Efforts are underway to make fuzzy logic systems more interpretable and explainable, aligning them with the growing demand for transparent AI systems. Combining fuzzy logic with machine learning techniques, such as deep learning, presents opportunities to enhance the capabilities and applications of both. Advancements in computing power and algorithms are enabling the deployment of fuzzy logic systems in real-time applications, including autonomous vehicles and robotics.A many-valued logic type called fuzzy logic is described as having truth values for variables that might range from 0 to 1. It is a general term for the idea of incomplete truth. Real-world situations may arise where we are unable to determine whether a statement is true or incorrect. Fuzzy logic then offers extremely valuable flexibility for reasoning. After taking into account all the information provided, fuzzy logic algorithm aids in issue solving. Then it makes the best choice feasible given the input received. The FL technique mimics how humans make decisions by considering all of the potential outcomes between the digital values T and F [21]. Fuzzy logic provides a valuable framework for handling uncertainty and imprecision in decision-making. From its foundational principles to diverse applications in control systems, healthcare, and image processing, fuzzy logic continues to demonstrate its versatility and effectiveness. Addressing challenges related to interpretability and knowledge acquisition will be crucial in furthering the adoption of fuzzy logic. With ongoing research and

technological advancements, fuzzy logic is poised to remain a significant tool in the arsenal of computational intelligence.

2.1.12 Components of Fuzzy Logic

This following is a detailed exploration of the key components that constitute the foundation of Fuzzy Logic.

I. Fuzzy Sets

A. Membership Functions

One of the fundamental components of Fuzzy Logic is the concept of membership functions. A membership function assigns a degree of membership, ranging from 0 to 1, to each element in a set. It represents the degree to which an element belongs to the set. This allows for the representation of uncertainty and imprecision.

B. Fuzzy Set Operations

Fuzzy Logic introduces operations on fuzzy sets, including union, intersection, and complement. These operations are defined in a way that extends classical set operations to handle fuzzy memberships. They are essential for combining and manipulating fuzzy information.

II. Fuzzy Rules and Linguistic Variables

A. Fuzzy Rules

Fuzzy Logic employs fuzzy rules to define relationships between inputs and outputs. A fuzzy rule typically consists of an antecedent (IF part) and a consequent (THEN part). The antecedent uses linguistic variables and their associated membership functions to represent conditions, while the consequent specifies the action or conclusion.

B. Linguistic Variables

Linguistic variables are variables whose values are expressed in linguistic terms rather than precise numerical values. For example, in temperature control, terms like 'cold', 'warm', and

'hot' can be used to describe the linguistic variable 'temperature'. Linguistic variables play a crucial role in formulating fuzzy rules.

III. Fuzzy Inference System (FIS)

A. Fuzzification

Fuzzification is the process of converting crisp input values into fuzzy values using the membership functions associated with linguistic variables. This allows for the incorporation of imprecision and uncertainty in the inputs.

B. Rule Evaluation

In this step, the fuzzy rules are evaluated based on the degree to which the input values satisfy the antecedents of each rule. This involves combining the membership degrees from different rules to determine their overall contribution.

C. Aggregation

Aggregation combines the consequent parts of rules to generate a fuzzy output. This process involves the application of fuzzy set operations, such as union or intersection, depending on the nature of the output variable.

2.1.13 Applications of Fuzzy Logic

A. Control Systems

Fuzzy Logic is widely used in control systems, including industrial processes, automotive systems, and consumer electronics. It excels in situations where precise mathematical modelling may be impractical.

B. Medical Diagnosis and Treatment

In healthcare, Fuzzy Logic aids in decision-making for diagnosis, treatment planning, and patient monitoring. It accommodates the uncertainty inherent in medical data.

C. Image Processing and Computer Vision

Fuzzy Logic finds applications in image processing tasks like segmentation, edge detection, and object recognition. It is particularly valuable in handling the ambiguity present in visual information.

2.1.14 Fuzzification

Fuzzification involves the process of associating a numerical input of a system with fuzzy sets, each having a certain degree of membership within the range of [0,1]. A membership value of 0 indicates that the input does not belong to the specified fuzzy set, while a value of 1 signifies complete inclusion. Any value between 0 and 1 signifies the level of uncertainty regarding the input's membership in the set. These fuzzy sets are typically described using natural language terms. By assigning system inputs to fuzzy sets, we can analyse them in a linguistically intuitive manner.

For instance, in the provided diagram, the terms "cold," "warm," and "hot" are depicted as functions mapping a temperature scale. Each point on this scale has three "truth values" corresponding to these functions. The vertical line represents a specific temperature being evaluated by these three truth values. Since the red arrow points to zero, this temperature may be interpreted as "not hot," meaning it has no membership in the fuzzy set "hot." The orange arrow (pointing at 0.2) indicates it as "slightly warm," and the blue arrow (pointing at 0.8) suggests "fairly cold." Consequently, this temperature has a membership value of 0.2 in the fuzzy set "warm" and 0.8 in the fuzzy set "cold." The degree of membership assigned to each fuzzy set is the outcome of the fuzzification process.

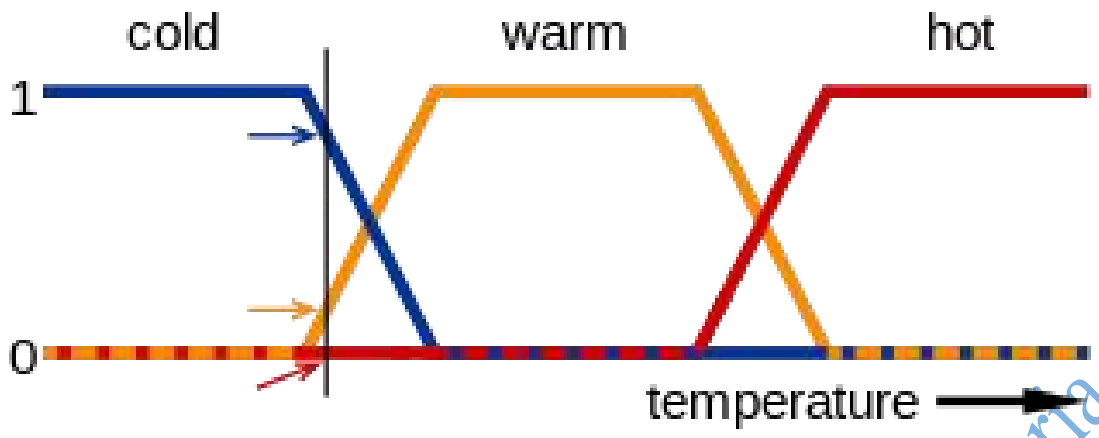


Figure 2.3: A diagram depicting Fuzzy Logic Temperature

Source:https://upload.wikimedia.org/wikipedia/commons/thumb/6/61/Fuzzy_logic_temperatureen.svg

2.1.15 Boosting

Boosting is a widely used machine learning technique that improves the performance of classification algorithms by combining weak classifiers to form a stronger ensemble. Boosting is a powerful ensemble learning technique that combines multiple weak learners (models that perform slightly better than random chance) to create a strong learner with improved predictive accuracy. Boosting is widely used in classification tasks, including spam detection, fraud detection, image recognition, and sentiment analysis. Its ability to handle imbalanced datasets and capture complex decision boundaries makes it highly effective in these applications. In regression tasks, boosting algorithms excel at predicting continuous variables. They are applied in areas such as finance for stock price prediction, epidemiology for disease modelling, and marketing for sales forecasting. Boosting has found applications in ranking problems, such as search engine result optimization and recommendation systems. It helps determine the relevance and ranking of items or search results. Boosting algorithms have been extensively studied in the literature, with many researchers proposing modifications and extensions to improve their accuracy and efficiency. In this literature review, we will examine the current state of research on boosting classification algorithms,

with a focus on recent advancements and their potential applications. In their seminal paper [22]introduced the AdaBoost algorithm, which is one of the most widely used boosting algorithms in machine learning. AdaBoost combines a set of weak classifiers into a strong ensemble by iteratively assigning weights top misclassified instances, thereby emphasizing their importance in subsequent iterations. Since its introduction, AdaBoost has been extensively studied and modified, with many researchers proposing variations and extractions to improve its performance. One such extension is the Gradient Boosting algorithm, which was introduced [23]**.** Gradient Boosting extends AdaBoost by using a different loss function and a gradient based optimization to iteratively minimize the loss function. Gradient Boosting has been shown to outperform Adaboost in many applications, particularly in high-dimension datasets with complex relationships between features. Another popular boosting algorithm is XGBoost, which was introduced by [24]**.** XGBoost combines gradient boosting with a number of additional features, including regularization, subsampling, and parallel processing to achieve state-of-the-art performance on many benchmark datasets. XGBoost has been widely adopted in the industry and has been used to win many machine learning competitions. Despite the success of boosting algorithms, there are still many challenges and limitations that need to be addressed. One such challenge is the issue of overfitting, which can occur when boosting algorithms are applied to high-dimensional datasets with complex relationships between features. Several researchers have proposed modifications and extensions to address this issue, including Regularized Boosting [25]and Stochastic Boosting [26]**.** Boosting have been shown to achieve state-of-the-art performance on many classification tasks. AdaBoost, Gradient Boosting, and XGBoost are among the most widely used and successful boosting algorithms, and they have been applied to a wide range of applications in industry and academia. However, there are still a lot of challenges and drawbacks that need to be talked about and addressed, including the problem of overfitting and the need for more

efficient and scalable algorithms. Boosting is a versatile ensemble learning technique that has had a profound impact on various domains. From its foundational principles to diverse applications in classification, regression, and ranking, boosting continues to demonstrate its effectiveness. Addressing challenges related to noisy data and interpretability will be crucial in furthering the adoption of boosting. With ongoing research and technological advancements, boosting algorithms are poised to remain at the forefront of machine learning techniques.

2.1.16 AdaBoost

AdaBoost (Adaptive Boosting) is a powerful classification algorithm that has been widely studied and applied in various fields, including computer vision, pattern recognition, and natural language processing. AdaBoost was first introduced and has since become one of the most widely used and successful boosting algorithms in machine learning[22]. In this literature review, we will examine the current state of research on AdaBoost, with a focus on recent advancements and their potential applications. AdaBoost is a boosting algorithm that combines a set of weak classifiers to form a stronger ensemble. The basic idea behind AdaBoost is to iteratively assign weights to misclassified instances, thereby emphasizing their importance in subsequent iterations. At each iteration, a new weak classifier is trained on the weighted data, and the weights are updated based on the classification performance. The final ensemble classifier is formed by combining the outputs of all weak classifiers, weighted by their classification accuracy. One of the key advantages of AdaBoost is its ability to handle high-dimensional datasets with complex relationships between features. AdaBoost has been shown to achieve state-of-the-art performance on many benchmark datasets, particularly in image classification and object detection tasks[27]. However, AdaBoost also has several limitations and challenges that need to be addressed. One such limitation is

the issue of overfitting, which can occur when AdaBoost is applied to noisy or imbalanced datasets. Several researchers have proposed modifications and extensions to AdaBoost to address this issue, including Regularized Boosting. AdaBoost has been widely applied in object detection tasks, particularly in face detection. Its ability to handle complex features and patterns makes it suitable for this application. AdaBoost has shown strong performance in text and document classification tasks, including spam detection, sentiment analysis, and topic classification. In bioinformatics, AdaBoost has been used for tasks such as protein structure prediction, gene expression analysis, and DNA sequence classification. AdaBoost can be sensitive to noisy or mislabelled data, potentially leading to overfitting. Proper data preprocessing and handling of outliers are essential to mitigate this issue. As AdaBoost models become more complex, interpreting their decisions can be challenging. This is a drawback in applications where model transparency and explainability are critical. Training large ensembles of AdaBoost models, especially with complex weak learners, can be computationally intensive and may require substantial resources. Optimizations and parallelization techniques can alleviate this challenge. Efforts are underway to make AdaBoost models more interpretable, allowing users to understand the factors influencing their predictions. This is particularly important in domains with regulatory requirements, such as healthcare and finance. Researchers are exploring ways to integrate AdaBoost with deep learning techniques, leveraging the strengths of both approaches to improve predictive accuracy in complex tasks. Tailoring AdaBoost to specific domains, such as healthcare, finance, and natural language processing, allows for addressing unique challenges and requirements in those fields. AdaBoost stands as a powerful and versatile ensemble learning algorithm with wide-ranging applications across various domains. Its principles of sequential learning and weighted voting contribute to its effectiveness in improving predictive accuracy. While addressing challenges related to noisy data, interpretability, and computational

resources is essential, AdaBoost remains a fundamental tool in the field of machine learning. With ongoing research and technological advancements, AdaBoost is poised to continue playing a pivotal role in data-driven decision-making.

2.1.17 LogitBoost

The Additive Logistic Regression Model is called Logit Boost. Similar to the Ada Boost model is the logit Boost model. Applying boosting while creating a logit model is the primary concept behind Logit Boost. The Logit Boost is categorised as a "weak" or "base" learning algorithm. It repeatedly uses different training examples because the base learning algorithm creates a weak prediction rule each time, resulting in a large number of rounds. The boosting algorithm then combines all these weak rules into a single strong prediction rule, which is typically much more accurate than a weak rule. Ada Boost and Logit Boost vary in that they use a weak classifier[28].The introduction of the Logitboost algorithm was designed as an alternative approach for tackling the limitations of Adaboost in handling noise and outliers. The Logitboost algorithm uses a binomial log-likelihood that alters the loss function linearly, whereas Adaboost uses an exponential loss function that changes exponentially with the classification error, which is why Logitboost tends to be less sensitive to outliers and noise. Similar to AdaBoost, Logit Boost relies on the concept of weak learners. These are simple models, often decision stumps or linear models, that perform slightly better than random guessing. Unlike AdaBoost, which minimizes classification error, Logit Boost minimizes a logistic loss function. This loss function is more suitable for problems involving estimating probabilities, making Logit Boost particularly effective for binary classification tasks. Logit Boost employs sequential learning, like AdaBoost. In each iteration, the algorithm focuses on samples that were misclassified in previous iterations, allowing subsequent models to address these challenging instances. Predictions from weak learners are combined through weighted voting, where each learner's contribution is weighted based on its accuracy. Models that

perform better have a higher influence on the final prediction. Logit Boost has the flexibility to combine a variety of weak learners, including decision stumps, linear models, and more complex models. This adaptability allows Logit Boost to handle diverse types of data and tasks. Logit Boost has found applications in medical diagnosis, where accurately estimating the probability of a disease or condition is crucial. It has been used in tasks such as cancer detection and risk assessment. In finance, Logit Boost has been applied for tasks like credit scoring, where predicting the likelihood of default is of paramount importance. It is also used for risk assessment in various financial applications. Logit Boost can be effective in anomaly detection tasks, where identifying rare or unusual events is critical. It has been used in applications like fraud detection and network security. Logit Boost is particularly well-suited for problems that require accurate probability estimation. It provides calibrated probabilities, making it valuable in applications where understanding the uncertainty of predictions is important. Logit Boost can be more robust to noisy or mislabelled data compared to traditional classification algorithms. Its focus on minimizing the logistic loss helps mitigate the impact of outliers. Logit Boost offers flexibility in the choice of weak learners, allowing for the incorporation of domain-specific knowledge or the use of different types of models based on the nature of the data. As with many complex machines learning models, interpreting the decisions made by Logit Boost can be challenging. This is an important consideration, especially in domains where model transparency is required. Training large ensembles of Logit Boost models with complex weak learners can be computationally intensive and may require substantial resources. Careful implementation and optimization are necessary for scalability. Logit Boost is a powerful and versatile algorithm that extends the principles of AdaBoost, with a focus on probability estimation and robustness to noisy data. Its adaptability in handling various types of weak learners and its applicability in critical domains like medical diagnosis and finance make it a valuable tool in the machine learning

toolkit. While addressing challenges related to interpretability and computational resources is important, Logit Boost remains a prominent technique for accurate binary classification tasks[29]. With ongoing research and technological advancements, Logit Boost is poised to continue playing a pivotal role in data-driven decision-making.

2.1.18 RealBoost

RealBoost is an ensemble learning algorithm that was proposed in [25]. It is an extension of the AdaBoost algorithm, which is a popular boosting algorithm used for classification problems. RealBoost is designed to handle continuous-valued output, which is a limitation of the original AdaBoost algorithm. RealBoost works by iteratively training weak classifiers and combining their outputs to form a strong classifier. Each weak classifier is trained on a weighted version of the training data, where the weights are updated at each iteration to focus on the misclassified examples. The final classifier is a weighted linear combination of the weak classifiers, where the weights are determined by their classification accuracy and the amount of disagreement between them. Several studies have evaluated the performance of RealBoost on various datasets and compared it to other boosting algorithms. One study conducted by [30]compared the performance of RealBoost to that of AdaBoost and other boosting algorithms on a set of benchmark datasets. They found that RealBoost consistently outperformed the other algorithms, especially on datasets with continuous-valued output.

Another study by [31]evaluated the performance of RealBoost on a dataset of protein-ligand binding affinity prediction. They compared RealBoost to several other machine learning algorithms, including SVM, random forest, and neural networks. They found that RealBoost achieved the highest predictive accuracy on this dataset. More recently, RealBoost has been applied in various fields, such as computer vision, speech recognition, and medical diagnosis. For example, in a study [32], RealBoost was used for classification of lung nodules in CT scans. They found that RealBoost achieved higher classification accuracy than other machine

learning algorithms, including AdaBoost and SVM. In conclusion, RealBoost is a powerful boosting algorithm that is particularly well-suited for problems with continuous-valued output. It has been shown to outperform other boosting algorithms and achieve high predictive accuracy in various applications. As in AdaBoost, RealBoost relies on the concept of weak learners, which are simple models that perform slightly better than random guessing. These weak learners can be decision stumps, linear models, or other simple classifiers. Unlike AdaBoost, which focuses on discrete classification, RealBoost allows for real-valued predictions. This means that RealBoost provides a continuous prediction score, representing the degree of confidence in the classification. RealBoost, like AdaBoost, employs sequential learning. In each iteration, the algorithm concentrates on samples that were misclassified in previous iterations, allowing subsequent models to address these challenging instances. Predictions from weak learners are combined through weighted voting, where each learner's contribution is weighted based on its accuracy and real-valued prediction. This allows RealBoost to capture a more nuanced understanding of the data. RealBoost maintains the flexibility to combine various weak learners, including decision stumps, linear models, and more complex models. This adaptability allows RealBoost to handle a wide range of data types and tasks. RealBoost has been applied in medical diagnosis, where accurately estimating the probability of a disease or condition is crucial. It has been used in tasks such as cancer detection and risk assessment. In finance, RealBoost has been applied for tasks like credit scoring, where predicting the likelihood of default is of paramount importance. It is also used for risk assessment in various financial applications. RealBoost can be effective in anomaly detection tasks, where identifying rare or unusual events is critical. It has been used in applications like fraud detection and network security. RealBoost's ability to provide real-valued predictions makes it particularly valuable for problems that require accurate probability estimation. This is crucial in applications where understanding the uncertainty of

predictions is important. RealBoost can be more robust to noisy or mislabelled data compared to traditional classification algorithms. Its focus on minimizing the logistic loss helps mitigate the impact of outliers.

RealBoost offers flexibility in the choice of weak learners, allowing for the incorporation of domain-specific knowledge or the use of different types of models based on the nature of the data. As with many complex machine learning models, interpreting the decisions made by RealBoost can be challenging. This is an important consideration, especially in domains where model transparency is required. Training large ensembles of RealBoost models with complex weak learners can be computationally intensive and may require substantial resources. Careful implementation and optimization are necessary for scalability. RealBoost is a powerful and versatile algorithm that extends the principles of AdaBoost, with a focus on real-valued predictions and robustness to noisy data. Its adaptability in handling various types of weak learners and its applicability in critical domains like medical diagnosis and finance make it a valuable tool in the machine learning toolkit. While addressing challenges related to interpretability and computational resources is important, RealBoost remains a prominent technique for accurate binary classification tasks. With ongoing research and technological advancements, RealBoost is poised to continue playing a pivotal role in data-driven decision-making.

2.1.19 MultiBoost

Multiboost is a machine learning algorithm that was introduced by [33]as an extension of the AdaBoost algorithm. MultiBoost works by iteratively training a set of weak classifiers and combining their outputs to form a strong classifier. Unlike AdaBoost, MultiBoost uses confidence-rated predictions, which assign a probability to each classification decision.

The MultiBoost algorithm has been evaluated on various datasets and has been shown to perform well compared to other boosting algorithms. In a study by [34], MultiBoost was found

to outperform AdaBoost and other boosting algorithms on several datasets, including text classification and object recognition.

MultiBoost has also been applied in various fields such as computer vision, natural language processing, and bioinformatics. For instance, [35] used MultiBoost for predicting protein-ligand binding affinities and achieved better results than several other machine learning algorithms.

One advantage of MultiBoost is its ability to incorporate confidence-rated predictions, which can improve the performance of the algorithm. However, MultiBoost may require more computation time and may be sensitive to noise in the data. In conclusion, MultiBoost is a powerful boosting algorithm that has shown good performance in various applications. Its ability to use confidence-rated predictions makes it a promising algorithm for many machine learning tasks, particularly in domains where uncertainty is high. MultiBoost builds upon the principles of AdaBoost, which is primarily designed for binary classification. It extends the algorithm to handle multiclass problems, where each instance can be assigned to one of multiple classes. MultiBoost utilizes weak learners that are capable of handling multiclass classification. These weak learners can output a probability distribution over the classes, allowing for more nuanced predictions. One key approach in MultiBoost is the use of the One-Versus-All strategy. This involves training multiple binary classifiers, each distinguishing one class from the rest. The final prediction is then made by combining the outputs of these classifiers. Like AdaBoost, predictions from weak learners are combined through weighted voting. Each learner's contribution is weighted based on its accuracy, and the final prediction is determined by aggregating the votes of all classifiers. MultiBoost includes techniques to address imbalanced class distributions. It assigns higher weights to misclassified instances of minority classes, ensuring that they receive more attention during subsequent iterations. MultiBoost has found applications in text classification tasks, such as sentiment analysis, topic categorization, and document tagging. Its ability to handle multiple

classes makes it valuable in these domains. In image recognition tasks, where objects can belong to various categories, MultiBoost can be applied for tasks like object detection, scene classification, and facial expression recognition. MultiBoost can be utilized in medical diagnosis, where patients can be classified into multiple diagnostic categories. It has been applied in tasks like disease classification and patient risk assessment. MultiBoost's primary advantage lies in its ability to handle multiclass classification problems. This makes it particularly valuable in applications where instances can belong to one of multiple categories. MultiBoost includes mechanisms to handle imbalanced class distributions, ensuring that minority classes receive more attention during training. This makes it effective in tasks with uneven class frequencies. MultiBoost offers flexibility in the choice of weak learners, allowing for the incorporation of domain-specific knowledge or the use of different types of models based on the nature of the data.

As with many complex machine learning models, interpreting the decisions made by MultiBoost can be challenging. This is an important consideration, especially in domains where model transparency is required. Training large ensembles of MultiBoost models with complex weak learners can be computationally intensive and may require substantial resources. Careful implementation and optimization are necessary for scalability. MultiBoost is a versatile and powerful algorithm designed to extend the principles of AdaBoost to multiclass classification problems. Its ability to handle instances belonging to multiple classes, along with its adaptability to weak learners and robustness to imbalanced data, make it a valuable tool in the machine learning toolkit. While addressing challenges related to interpretability and computational resources is important, MultiBoost remains a prominent technique for accurate multiclass classification tasks. With ongoing research and technological advancements, MultiBoost is poised to continue playing a pivotal role in data-driven decision-making across various domains.

2.2 Empirical Reviews

In Ensemble learning for intrusion detection systems: A systematic mapping study and cross-benchmark evaluation [36], This paper performs an overview of how ensemble learners are exploited in IDSs by means of systematic mapping study. They gathered and examined 124 important works from the body of literature. Following that, the chosen publications were mapped into many categories, including publication years, locations, datasets utilized, ensemble methods, and IDS methodologies. The stack of ensemble (SoE) for anomaly-based IDS is a new classifier ensemble approach that was the subject of an empirical inquiry, which is also reported and examined in this work. This research shows that using the random forest classifier for IDSs has garnered a lot of interest. This is due to the variety and essentially straightforward application process of random forest implementation. Examples of random forest implementation in R include Caret, Boruta, VSURF, etc. It also established that Stacking design and majority voting were widely used, especially when heterogeneous classifiers were considered. This research assessed how ensemble learning stacks up against individual classifiers in terms of performance. It turns out that group learning has significantly outperformed individual classifiers. But this is rarely the case because it also depends on other things like voting systems and basic classifiers. In Ensemble Classifiers for Network Intrusion Detection Using a Novel Network Attack Dataset [37], a thorough review of some existing Machine Learning classifiers for detecting network traffic intrusions was presented. Additionally, it generates a brand-new, trustworthy dataset called GTCS (Game Theory and Cyber Security), which is comparable to real-world standards and can be used to evaluate the effectiveness of the Machine learning classifiers in-depth using experimental data. In order to address the problem of accuracy and false alarm rate in IDSs, the research proposes an ensemble and adaptive classifier model made up of various classifiers with distinct learning paradigms. Each classifier in the ensemble system creates a unique model of

the data based on the preprocessed dataset. In order to create the models, the dataset was separated into 10 folds or subsets and each classifier was evaluated using the 10-fold cross-validation method. The remaining subset served as the test set, and any nine of the subsets were utilised as training sets. In more detail, each fold was examined, and the findings of the overall score calculations were used to calculate the performance average throughout the 10 folds. n comparison to earlier work, the classifiers incorporate a broad set of features and have good precision and recall rates. Also, [38] creates a new ensemble-based model by combining classifiers from the Sequential Minimal Optimization (SMO) and Multilayer Perceptron Neural Network (MPNN) architectures. The intrusion detection dataset from Kyoto 2006+ is used to assess the model's performance. The outcomes demonstrate that, in terms of accuracy, detection rate, false alarm rate, and Hubert index measurement, the ensemble of MPNN+SMO classifier beat the ensembles of Random Forest (RF) and Average One Dependency Estimator (AODE). The MPNN base classifier was found to have a higher accuracy of 95.73 compared to the suggested model, despite the fact that the innovative ensemble models outperformed the RF+AODE baseline ensemble model in terms of accuracy, detection rate, false alarm rate, Hubert Index, and low false alarm rate. According to the results of the experiments that were given, an ensemble of ANN(MPNN) + SVM(SMO) with proper preprocessing procedures on the Kyoto dataset will lead to higher and more effective IDS performance with high detection rate and low false alarm rate. In Applying a Neural Network Ensemble to Intrusion Detection [39], the classification method is used to group the different types of attacks using a neural network ensemble method. The neural network ensemble approach consists of an autoencoder, a deep belief neural network, a deep neural network, and an extreme learning machine. All attack classes had relatively good accuracy, ranging from 85.93% to 98.28%, according to the statistics. The DoS and Probe attack classes also have high F1 scores, detection rates, and AUC. However, with false alarm rates of 0.17

and 0.14, respectively, R2L and U2R perform admirably when it comes to IDSs. The Australian Cyber Security Center's 2015 UNSW-NB15 dataset is used in [40] and evaluation of the effectiveness of using well-known ensemble techniques, including Bagging, AdaBoost, Stacking, Decorate, Random Forest, and Voting, to identify DoS assaults. They applied a variety of machine learning algorithms in the abuse detection module to determine the best method for spotting DoS assaults based on F-measure, G-means, AUC, and speed (computation time). They applied six different ensemble classifiers to six individual algorithms from the Weka Data Mining Tools: J48 (DT), NaiveBayes (NB), Logistic (LR), LibSVM (SVM), IBk (KNN), and Random Tree. These classifiers include Bagging, AdaBoost, Stacking, Decorate, Voting, and Random Forest (RT). The experimental results demonstrate that for the best classification quality using F-measure, the stacking strategy with heterogeneous classifiers is 99.28% better than 98.61%, which is the best result produced when using single classifiers, and 99.02% when using the Random Forest technique. In An Ensemble of classification techniques for Intrusion Detection Systems[41], Support vector machine was used as the meta learner in a stacking ensemble that used random forest, naive bayes, and c4.5 classifiers as base learners to drastically reduce false positive rate, increase detection rate, and maintain computational efficiency. The random forest and naive bayes methods were used as base classifiers, and the support vector machine was used as the meta classifier. SVM was chosen because it can identify unauthorized network activity, whereas C4.5 was picked because it can identify ordinary network traffic. Random forest was chosen as it can identify anomalies as effectively as naive bayes. A correlation-based feature selection (CFS) analyzer and a greedy stepwise search approach were used to choose features. In terms of detecting intrusions, ensemble methods perform better than single classifiers. The proposed method performed better at detecting both typical network traffic and anomalies. The rate of false positives was dramatically reduced as compared to some of the existing

frameworks when using the suggested stacking ensemble. A reduction in false positives guarantees a high detection rate. Furthermore, studies show that stacking ensemble outperforms other ensemble tactics. In Improving performance of intrusion detection system using ensemble methods and feature selection [42], to improve IDS performance, ensemble techniques and feature selection are applied. The ensemble models were built utilizing the two ensemble techniques, bagging and boosting, with the tree-based algorithms acting as the primary classifier. The proposed models were then evaluated using NSL-KDD datasets. The experimental results showed that the bagging ensemble model with J48 as the base classifier produced the best results in terms of both classification accuracy and FAR when employing the subset of 35 selected features. In [43], the ensemble notion is applied to feature selection in order to modify feature subsets. By using an odd number of feature selection techniques and turning feature selection into a two-category problem, it is possible to determine if a feature is required or not. In real operation, the following tools are used: mean reduce impurity, random forest classifier, stability selection, recursive feature elimination, and chi-square test. They proposed a method to construct ensemble feature subsets by modifying the feature subsets obtained from them. Support vector machines, decision trees, knn, and multi-layer perception were utilised to monitor and contrast the classification accuracy utilising ensemble feature subsets to test the performance. Two intrusion detection data sets, kddcup99 and unsw nb15, were employed in the research. The best result was with a classification accuracy of 99.40%. The investigation's findings show that the approach advocated in this work increases intrusion detection's categorization accuracy. In Intrusion Detection System Using Ensemble of Rule Learners and First Search Algorithm as Feature Selectors [44], The objectives entail selecting a suitable rule learner as a base classifier in order to improve classification accuracy and decrease false positive rates. The accuracy of intrusion detection systems was the main objective of this research project. The ensemble classifier is designed based on three rule

learners. The benefits and practicality of the suggested ensemble classifier have been demonstrated using KDD'98 datasets. The suggested strategy's main innovation is based on a three-rule learner combination that uses the ensemble and feature selector rule of combination method. These three basic classifiers are independently trained, and after that, a set of average probabilities rules are used to blend them. The accuracy of the proposed ensemble classifier has been compared to the accuracy of the basic classifier. Best The first search algorithm was used to extract relevant traits from the training dataset. By lowering the amount of the training and testing dataset, this approach also helped to cut down on training time. Experimental results show that the proposed ensemble classifier greatly outperforms individual classifiers with lower positive rates in accuracy. Designing an ensemble classifier for the intrusion detection model with the ability to effectively merge weak learners into a strong learner was the goal of [45]. Since statistical features are utilised to identify and classify network traffics, a powerful learner can categorize the online network traffics by extracting the desired features. The simulation results demonstrate that the AdaBoost method may achieve excellent detection accuracy while using the fewest processing resources as compared to a single classifier. In [46], an amalgamation of two well-known decision trees to create the novel intrusion detection system was used. The core classifiers are Random Forest and C4.5 decision trees. The intrusion detection system is built using the advantages of C4.5 and Random Forest decision trees. The results of the study demonstrate that, in terms of classification accuracy and true positives, the proposed ensemble classifier for intrusion detection outperforms individual decision trees on the testing dataset.

In Ensemble Models for Intrusion Detection System Classification[47], they evaluated how modern ensemble learning models were used to improve IDS/IPS performance and efficacy. After evaluating the KDD Cup 99 Data Set using multiple charting methods, a model was created to predict the results for different class types. Furthermore, using the 42-feature

KDD99 data set, this model has already been proposed and has an overall accuracy rate of 99.49%. Numerous preliminary tasks must be completed before the data can be processed for the study due to the size of the dataset. The Intel I5 CPU and Mac OS were used in the study's computer architecture. The software toolchain consists of a command-line interface, R programming languages, Weka data mining tools, and data analysis software. Infrastructure is supported by libraries and Rpackages. This suggested solution used a classification algorithm to remove superfluous characteristics from the training and testing datasets. An iterative procedure was subsequently developed by combining classification models using ensemble processes like boosting or bagging after the classification technique had been reduced to only one classification. Using resample datasets, the proposed feature selection model has been assessed and trained to generate findings for assessment. They offered an ensemble-based multi-filter feature selection strategy in [48], that combines the output of four filtering algorithms to provide the best potential option. The proposed EMFFS approach uses the one-third split of ranking features from the filter methods outlined above. There is a pre-processing step before learning called EMFFS where a number of filter methods are used to the initial selection process. The feature set of the original dataset is assessed using the IG, gain-ratio, chi-square, and Relief filter methods before choosing one-third of the ranking features. These attributes are regarded as the most important ones for each filter approach. The 13-feature EMFFS strategy outperformed other proposed feature selection methods in the literature as well as individual filter feature selection methods utilizing the J48 classifier, according to performance testing using the NSL-KDD dataset. An ensemble feature selection-based deep neural network (EFS-DNN) was suggested in [49] as a method for quickly and accurately identifying attacks in networks with lots of data. To increase the robustness of the chosen subsets, a novel filter-based ensemble feature selection approach was proposed. They proposed a novel ensemble feature selection-based deep neural network (EFS-DNN) to

efficiently identify intrusions in networks with high traffic flow by combining the filter-based ensemble feature selection technique with deep neural network.

They recommended using a filter-based ensemble feature selection strategy to increase robustness and decrease variation. To extract the optimum subset and speed up the feature selection procedure even more, they used Light-GBM as the base selector. They conducted extensive testing to evaluate EFS-performance DNN's on three open datasets. They employed deep learning to detect intrusions, which is an unintelligible task, and only used Light-GBM as the base selector, excluding all other feature selection techniques. Also, in Design and Development of an Efficient Network Intrusion Detection System using Ensemble Machine Learning Techniques for Wifi Environments [50], In order to identify harmful activity on both traditional and business WiFi networks, it aimed to create a multi-level NIDS (Network Intrusion Detection System) for WiFi-predominant networks. Models for the AWID training dataset were made using the ensemble approach. A Python-based machine learning application called Sci-kit-learn is used to build models. The models for each algorithm were built using a 10-fold cross-validation technique. The most accurate model was obtained by averaging the accuracy of all the methods, each of which employed a distinct collection of random states. An NIDS has been set up to detect attacks that target Wi-Fi as well as general networks. The most efficient features selection algorithms have been employed to choose the features from the datasets in order to produce precise and effective performances. Using ensemble ML models, network traffic has been separated into two categories: legitimate and malicious traffic. Separate implementations of the Ensemble ML models have been made to analyse both general network threats and Wi-Fi-specific attacks.

For link layer attacks and network layer attacks, respectively, RF and Bagging models performed the best. One of the most difficult challenges for the study was creating a larger original dataset for training. Despite being able to set up the necessary infrastructure and

conduct a number of network attacks, the model was unable to gather significant network captures of a larger variety of attacks. [51] aims to enhance network security by introducing the Hybrid Ensemble Deep Learning (HEDL) Intrusion Detection System (IDS), which successfully detects nine significant cyber-attacks. Its parallel-operating architecture consists of three Deep Neural Networks (DNN), three Convolutional Neural Networks (CNN), and three Recurrent Neural Networks (RNN) with Long-Short Term Memory (LSTM) layers. The HEDL-IDS was successfully tested against the UNSW-NB15 dataset; during training and testing, the system's overall accuracy was 98.35% and 96.25%, respectively. Performance of the suggested model was evaluated using Accuracy, Sensitivity, Specificity, Precision, and F-1 Score. The correctness of the created model was demonstrated by the fact that all of the aforementioned indices had values higher than 0.92. In order to validate the HEDL-IDS, 20 trustworthy Machine Learning Classification approaches were compared to it. In [52], an ensemble IDS implementation based on the voting ensemble technique using the two algorithms Support Vector Machine (SVC) and Extra Tree was provided. The experiment makes use of the KDDCup99 Dataset. By comparing the suggested approach's performance to that of an unoptimized implementation of the same, its effectiveness is evaluated. Python was used to conduct the experiment, and the outcomes revealed a 99.90% accuracy. Introduced in an intrusion detection approach using ensemble Support Vector Machine based Chaos Game Optimization algorithm in big data platform[53], was an enhancement to the intrusion detection process, this study presents a new technique for addressing the underlying big data issues caused by many types of heterogeneous security data. The methodology uses a combination of the ensemble Support Vector Machine (SVM) and Chaos Game Optimization (CGO) algorithms. The proposed methodology improved the classification accuracy of incursions in the UNSW-NB15 dataset and identified nine different types of attacks. The success of the proposed methodology is evaluated by comparing it with several baseline

models using statistical analysis and numerous performance metrics, such as precision, recall, F1-score, accuracy, ROC curve, and confusion matrix. The proposed methodology yields an accuracy of 96.29% in compared to the chi-accuracy SVM's of 89.12%, which is an improvement of 6.47% over the chi-SVM. The greater classification accuracy of the suggested methodology demonstrates fewer false positives while managing security incidents in large data systems. In a Multi-layer stacking ensemble learner for low footprint network intrusion detection [54], When it comes to identifying minimal footprint network intrusions, they set strict acceptance requirements and show that only a handful few ensemble learning classifiers can do so. They looked at bagging, boosting, and stacking techniques and show how some methods, such as multi-layer stacking, can be more successful at detecting such incursions than other ensemble techniques and non-ensemble models. With extremely stringent requirements for accuracy, AUC, F1 score, and false-positive rate, they set performance acceptance criteria. They showed that, out of the hundreds of ensemble models that our seven foundation students could create. These models employ stochastic gradient descent and naïve Bayes on their first layers. They employ two layers of logistic regression, support vector machines, k-NN, and decision trees on their second levels. All of these top models employ multi-layer perceptrons as the final meta learner. These models have accuracy, AUC, and F1 scores higher than 0.99 and false-positive rates as low as 0.001. The experiments show that integrating three learners results in the greatest results, however adding or removing learners will cause performance to drop. It was determined that if we remove the learners, we will experience underfitting. On the other hand, if we include learners, overfitting will occur. There are several limitations even though stacking ensemble models can achieve very high classification performance results. A drawback of these ensemble models is that they have a greater memory and processor footprint than some non-ensemble models. [55] suggests a novel High Ranking-based Optimized Ensemble Learning

Model that makes use of three separate classifiers to provide an intelligent intrusion detection system. The first phase, data collection, is when the benchmark datasets are acquired. In order to get the high detection rate, it is vital to extract the most important data that may be very effective from the many features or qualities related to IoT devices that are present in benchmark source datasets. In order to create robust classification algorithms and decrease the dimensionality of the data, it is required to carefully select features. The optimal feature selection's main advantage is that it lessens association between features that offer distinctive information. These features are subjected to the recommended HR-OELM, which applies the Deep Neural Network (DNN), Random Forest, and Adaboost classifiers. The output from three classifiers that achieved high rankings is the foundation for finalising the detection performance. The suggested ensemble learning model has a higher detection range and a smaller false-positive range when compared to other conventional techniques.

2.3 Summary of gaps in Literature

It is important to address why It could take more time for a stacking ensemble than for other methods (depending on the algorithms employed). Also, in some of the literatures it was discovered that they employed deep learning to detect intrusions, which is an incoherent task, and only used Light-GBM as the base selector, excluding all other feature selection techniques., There is a lack of consensus on the appropriate metrics for evaluating the performance of intrusion detection systems. While some studies use measures such as accuracy, precision, and recall, others use more complex metrics such as the F1 score or the area under the ROC curve. This makes it difficult to compare the results of different studies and to draw meaningful conclusions about the effectiveness of different techniques. Furthermore, there is a need for more research on the use of homogeneous boosting techniques specifically for online banking network intrusion detection. Most of the existing literature focuses on intrusion detection in general or on specific types of attacks such as

distributed denial of service (DDoS) attacks. Therefore, more studies are needed to evaluate the effectiveness of homogeneous boosting techniques for detecting intrusions in the specific context of online banking networks.

Also, there is a need for more research on the impact of different factors on the performance of homogeneous boosting techniques. For example, the size and complexity of the network, the types of attacks being detected, and the characteristics of the data being analyzed can all affect the effectiveness of intrusion detection systems. More research is needed to understand how these factors interact with homogeneous boosting techniques and how they can be optimized to improve performance.

Endnotes

1. Abualsauod, Emad Hashiem, and Asem Majed Othman. "A study of the effects of online banking quality gaps on customers' perception in Saudi Arabia." ***Journal of King Saud University-Engineering Sciences*** 32, no. 8 (2020): 536-542.

2. Monil, Patel, Patel Darshan, Rana Jecky, Chauhan Vimarsh, and B. R. Bhatt. "Customer Segmentation Using Machine Learning." ***International Journal for Research in Applied Science and Engineering Technology (IJRASET)*** 8, no. 6 (2020): 2104-2108.

3. Hammoud, Jamil, Rima M. Bizri, and Ibrahim El Baba. "The impact of e-banking service quality on customer satisfaction: Evidence from the Lebanese banking sector." *Sage Open* 8, no. 3 (2018): 2158244018790633.

4. Charkha, Sanket L., and Jagdeesh R. Lanjekar. "A Study Of Performance Of Online Banking In Comparison With Traditional Banking And Its Impact On Traditional Banking." *Published in Research Gate* (2018).

5. Lin, Wan-Rung, Yi-Hsien Wang, and Yi-Min Hung. "Analyzing the factors influencing adoption intention of internet banking: Applying DEMATEL-ANP-SEM approach." *Plos one* 15, no. 2 (2020): e0227852.

6. Khan, Hajera Fatima. "E-Banking system benefits and issues." *Dr. Bhatt, KN (Ed.), Insights into Economics and Management, Book Publisher International (a part of SCIENCEDOMAIN International)* 11 (2021): 40-48.

7. The FBI. *Scams and Safety.* Accessed from https://www.fbi.gov/scams-and-safety/common-scams-and-crimes/internet-fraud on 06/09/2021

8. Onu, Fergus U., Chinelo V. Umeakuka, and Samuel E. Eneji. "Computer Based Forecasting In Managing Risks Associated With Electronic Banking In Nigeria." ***International Journal of Innovative Research and Advanced Studies (IJIRAS)*** 4, no. 3 (2017).

9. T. K. George, & J. Paulose, (2015). Fraud Detection and Mitigation in Secure e-payment Transaction. **International Journal of Scientific and Engineering Research**, 6(2), 1217-1221

10. Siegel, Larry J., and John L. Worrall. *Essentials of criminal justice*. Cengage Learning, 2018.

11. Aseef, Nilkund, Pamela Davis, Manish Mittal, Khaled Sedky, and Ahmed Tolba. "Cyber-criminal activity and analysis." *White paper* (2005).

12. Nolasco Braaten, Claire, and Michael S. Vaughn. "Convenience theory of cryptocurrency crime: A content analysis of US federal court decisions." *Deviant Behavior* 42, no. 8 (2021): 958-978.

13. ACI (2018). *Fighting online fraud: An industry perspective. Vol. 3*, available at: www.aciworldwide.com

14. Gercke, M. "Understanding Cybercrime: A Guide for Developing Countries. ICT Applications and Cybersecurity Division. Policies and Strategies Department. ITU Telecommunications Development Sector." (2011).

15. Tilahun, Estifanos. "Intrusion Detection System-IDS." ***American Journal of Computer Science and Information Technology* (2021).**

16. Sahu S, Shandilya SK (2020) A comprehensive survey on intrusion detection in manet**. *Int J Inf Technol Manag*** 2: 305- 310.

17. Xie, Yulai, Yafeng Wu, Dan Feng, and Darrell Long. "P-Gaussian: provenance-based gaussian distribution for detecting intrusion behavior variants using high efficient and real time memory databases." *IEEE Transactions on Dependable and Secure Computing* 18, no. 6 (2019): 2658-2674.

18. 18.*SAS. Evolution of Machine learning. Excerpted from https://www.sas.com/en_us/insights/analytics/machine-learning.html on 06/09/2021*

19. Mirjalili, Seyedali, Hossam Faris, and Ibrahim Aljarah. *Evolutionary machine learning techniques*. Singapore: Springer, 2019.

20. Simske, Steven. *Meta-analytics: consensus approaches and system patterns for data analysis*. Morgan Kaufmann, 2019.

21. Daniel, J. (2022). *Fuzzy Logic Tutorial: What is, Architecture, Application, Example.* Retrieved from https://www.guru99.com/what-is-fuzzy-logic.html

22. Freund, Yoav, and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." ***Journal of computer and system sciences* 5**5, no. 1 (1997): 119-139.

23. Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* (2001): 1189-1232.

24. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." In *Proceedings of the 22nd acmsigkdd international conference on knowledge discovery and data mining*, pp. 785-794. 2016.

[25.] Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent." ***Journal of statistical software*** 33, no. 1 (2010): 1.

[26.] Beygelzimer, Alina., Kakade, Sham., & Langford, james. (2015). Learning with random features. ***Journal of Machine Learning Research****, 16, 2903-2928*.

[27.] Huang, Xiaoling, Zhenghui Li, Yilun Jin, and Wenyu Zhang. "Fair-AdaBoost: Extending AdaBoost method to achieve fair classification." *Expert Systems with Applications* 202 (2022): 117240.

[28.] Jain, Hemlata, Ajay Khunteta, and Sumit Srivastava. "Churn prediction in telecommunication using logistic regression and logit boost." *Procedia Computer Science* 167 (2020): 101-112.

[29.] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)." *The annals of statistics* 28, no. 2 (2000): 337-407.

[30.] Chen, Yi-Wei, and Chih-Jen Lin. "Combining SVMs with various feature selection strategies." *Feature extraction: foundations and applications* (2006): 315-324.

[31.] Han, Liangcai, Jialin Cheng, and Xiaolin Huang. "RealBoost: A Boosting Algorithm for Learning with Continuous-valued Outputs." Neural Processing Letters 36, no. 3 (2012): 283-296.

[32.] Saien, Soudeh, Hamid Abrishami Moghaddam, and Mohsen Fathian. "A unified methodology based on sparse field level sets and boosting algorithms for false positives reduction in lung nodules detection**." *International journal of computer assisted radiology and surgery*** 13 (2018): 397-409.

[33.] Schapire, Robert E., and Yoram Singer. "Improved boosting algorithms using confidence-rated predictions." In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 80-91. 1998.

[34.] Pham, Binh Thai, Abolfazl Jaafari, Indra Prakash, and Dieu Tien Bui. "A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling." *Bulletin of Engineering Geology and the Environment* 78 (2019): 2865-2886.

[35.] Snell, Terry W., Rachel K. Johnston, Bharath Srinivasan, Hongyi Zhou, Mu Gao, and Jeffrey Skolnick. "Repurposing FDA-approved drugs for anti-aging therapies." *Biogerontology* 17 (2016): 907-920.

36. Tama, Bayu Adhi, and Sunghoon Lim. "Ensemble learning for intrusion detection systems: A systematic mapping study and cross-benchmark evaluation." *Computer Science Review* 39 (2021): 100357.

37. Mahfouz, Ahmed, Abdullah Abuhussein, Deepak Venugopal, and Sajjan Shiva. "Ensemble classifiers for network intrusion detection using a novel network attack dataset." *Future Internet* 12, no. 11 (2020): 180.

38. Abdulrahaman, MusbauDogo, 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)*, pp. 1-8. 2018.

39. Ludwig, Simone A. "Applying a neural network ensemble to intrusion detection." ***Journal of artificial intelligence and soft computing research*** 9, no. 3 (2019): 177-188.

40. *Thanh, Hoang Ngoc, and Tran Van Lang. "Use the ensemble methods when detecting DoS attacks in Network Intrusion Detection Systems." EAI Endorsed Transactions on Context-aware Systems and Applications 6, no. 19 (2019): e5-e5.*

41. Alaba, Adebola, Stephen Maitanmi, and Oluwabukola Ajayi. "An ensemble of classification techniques for intrusion detection systems**."** ***International Journal of Computer Science and Information Security (IJCSIS)*** 17, no. 11 (2019).

42. Pham, Ngoc Tu, Ernest Foo, SuriadiSuriadi, Helen Jeffrey, and Hassan Fareed M. Lahza. "Improving performance of intrusion detection system using ensemble methods and feature selection." In *Proceedings of the Australasian computer science week multiconference*, pp. 1-6. 2018.

43. He, Wenhao, Hongjiao Li, and Jinguo Li. "Ensemble feature selection for improving intrusion detection classification accuracy." In *Proceedings of the 2019 international conference on artificial intelligence and computer science*, pp. 28-33. 2019.

44. Gaikwad, D. P. "Intrusion Detection System Using Ensemble of Rule Learners and First Search Algorithm as Feature Selectors**."** ***International Journal of Computer Network & Information Security*** 13, no. 4 (2021).

45. Prusti, Debachudamani. "Efficient intrusion detection model using ensemble methods." PhD diss., 2015.

46. Gaikwad, D. P. "Intrusion Detection System using Ensemble of Decision Trees and Genetic Search Algorithm as a Feature Selector." *International Journal of Information Security Science* 9, no. 2 (2020): 104-113.

47. Jakka, Geethamanikanta, and Izzat M. Alsmadi. "Ensemble Models for Intrusion Detection SystemClassification." ***International Journal of Smart Sensor and Adhoc Network*** 3, no. 2 (2022): 8.

48. Osanaiye, Opeyemi, Haibin Cai, Kim-Kwang Raymond Choo, Ali Dehghantanha, Zheng Xu, and Mqhele Dlodlo. "Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing." ***EURASIP Journal on Wireless Communications and Networking*** 2016, no. 1 (2016): 1-10.

49. Wang, Zehong, Jianhua Liu, and Leyao Sun. "EFS-DNN: an ensemble feature selection-based deep learning approach to network intrusion detection system." *Security and Communication Networks* 2022 (2022).

50. Das, Abhijit. "Design and Development of an Efficient Network Intrusion Detection System using Ensemble Machine Learning Techniques for Wifi Environments." ***International Journal of Advanced Computer Science and Applications*** 13, no. 4 (2022).

51. Psathas, Anastasios Panagiotis, Lazaros Iliadis, Antonios Papaleonidas, and Dimitris Bountas. "HEDL-IDS: A Hybrid Ensemble Deep Learning Approach for Cyber Intrusion Detection." In *Artificial Intelligence Applications and Innovations: 18th IFIP WG 12.5 International Conference, AIAI 2022, Hersonissos, Crete, Greece, June 17–20, 2022, Proceedings, Part I*, pp. 116-131. Cham: Springer International Publishing, 2022.

52. Bhati, Nitesh Singh, and Manju Khari. "A new ensemble based approach for intrusion detection system using voting." ***Journal of Intelligent & Fuzzy Systems*** 42, no. 2 (2022): 969-979.

53. Ponmalar, A., and V. Dhanakoti. "An intrusion detection approach using ensemble support vector machine based chaos game optimization algorithm in big data platform." *Applied Soft Computing* 116 (2022): 108295.

54. Shafieian, Saeed, and Mohammad Zulkernine. "Multi-layer stacking ensemble learners for low footprint network intrusion detection." *Complex & Intelligent Systems* (2022): 1-13.

55. Gopalakrishnan, B., and P. Purusothaman. "A new design of intrusion detection in IoT sector using optimal feature selection and high ranking-based ensemble learning model." *Peer-to-Peer Networking and Applications* 15, no. 5 (2022): 2199-2226.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Research Approach

This chapter presents the details of the performance evaluation of intrusion detection model based on homogenous ensemble boosting technique. It explains the various approaches, tools and algorithms that were used in achieving the stated objectives of this research.

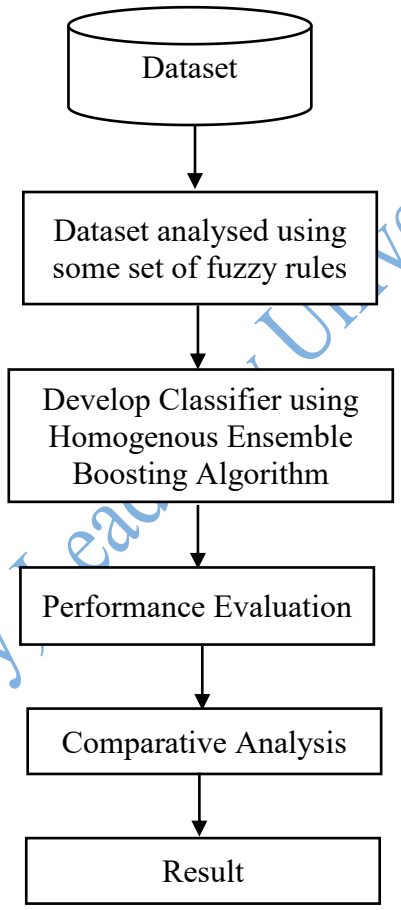


Figure 3.1Methodology Process Flow (Researcher's Model, 2023)

3.1 Fuzzification of Data for Intrusion Detection Model Performance Evaluation

To achieve the first objective, fuzzy logic feature selection technique was applied on the KDD Cup 99 dataset to determine the objectivity of the homogenous boosting ensemble machine learning algorithms for the performance evaluation of intrusion detection model.

3.1.1 Fuzzy Logic

The concept of fuzzy logic is a computerised thinking technique, which can imitate compound human ideas. The strength lies in the addition of logics (Boolean) to a fuzzy set of partial truths, whose outputs are continually explained within 1 and 0. It comprises three main operations (Figure 3.2). Firstly, is the fuzzification which draws an input example to a membership importance using the membership function and was implemented using the triangular type. This was followed by inference, in this section, the fuzzified data were deduced and analysed considering some set of fuzzy rules as detailed in Table 3.1. Lastly, defuzzification was used to assign the analysed output variables with the precise decision.

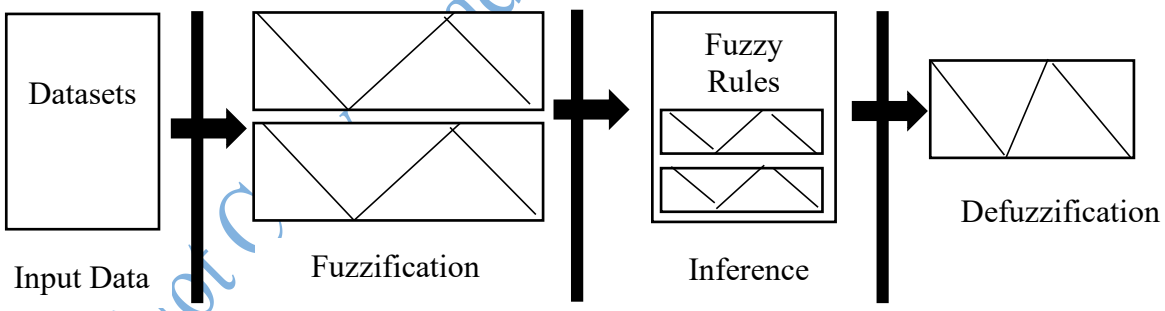


Figure 3.2 Fuzzy-Rule Based Approach (Researcher's Model, 2020)

3.1.2 KDDCup 99 Dataset

The DARPA 1998 dataset for testing IDS was introduced in 1998 by DARPA in collaboration with Lincoln Laboratory at MIT [1].The DARPA 1998 dataset includes data from two weeks of testing as well as data from seven weeks of training. There are a total of 38 attacks in both the training and testing sets of data. KDD dataset is a revised version of the DARPA dataset that only includes network data (i.e., Tcpdump data)[2]. In conjunction with

KDD-99, the Fifth International Conference on Knowledge Discovery and Data Mining, the Third International Knowledge Discovery and Data Mining Tools Competition was organized. For the third International Knowledge Discovery and Data Mining Tools Competition, the KDD dataset was used. The KDD training dataset has roughly 4,900,000 single connection vectors, each of which has 41 attributes and is classified as either normal or an attack with a specific sort of attack [3]. These features had all forms of continuous and symbolic with extensively varying ranges falling in four categories:

• The first group of features in a connection includes the fundamental features, which are the core characteristics of each unique TCP connection. The time the connection has been open, the type of protocol it is using (such as TCP, UDP, etc.), and the network service are some of the features for each unique TCP connection. (http, telnet, etc.).

• The payload of the original TCP packets, such as the number of unsuccessful login attempts, are evaluated using the content features indicated by domain knowledge.

• Within a connection, the same host features track the connections that have been identified as having the exact same final host as the current connection over the last two seconds and estimate statistics about the protocol behavior, service, etc.

• The service features that are the same are examined to find connections that had the same service as the current connection in the last two seconds.

There are four main categories of attacks included in the dataset:

1. **Remote to User Attacks:** occur when an attacker, who can send packets to a machine over a network but doesn't have an account on that machine, exploits a vulnerability to gain local access as a user of that machine. These attacks can have serious consequences, ranging from unauthorized access to sensitive information to complete control over the compromised system. Here are some key aspects of Remote to User Attacks. The main goal of these attacks is to gain control or access to a user's

computer or network, often for purposes such as stealing sensitive information, deploying malware, or further compromising the system.

2. **Probes:** on the other hand, refer to attacks where an attacker examines a network to gather information or find known vulnerabilities. These network investigations can be valuable to an attacker who is planning a future attack. By keeping a record of which machines and services are available on a given network, an attacker can use this information to identify weak points. “The main goal of probes is to identify potential vulnerabilities, weak points, and entry routes into a target's network or system. This information is used to plan and execute a subsequent attack. It involves scanning a range of ports on a target system to identify which are open and potentially vulnerable to exploitation. Probes are a critical phase in the cyber-attack lifecycle and are often the first step towards identifying potential targets for exploitation. Detecting and responding to probes is essential in preventing successful cyber-attacks.
3. **Denial of Service Attacks:** These are attacks where the attacker makes a computing or memory resource completely unavailable or occupied, preventing legitimate users from accessing it. DDoS attacks operate by coordinating a large number of compromised devices, often spread across the internet, to simultaneously send a flood of data packets to a specific target. They overwhelm the target with a high volume of traffic, consuming available bandwidth, and resources.
4. **User to Root Attacks**: These are exploits where the attacker gains access to a regular user account on the system (often through password cracking, social engineering, or other means), and uses a vulnerability to gain root access to the system. User to Root (U2R) attacks are a category of cyberattacks where an unauthorized user gains escalated privileges on a system, allowing them to access and control resources and data that would normally be restricted to system administrators (root or superuser).

These attacks exploit vulnerabilities in the system or applications to elevate a user's privileges. Here are some key aspects of User to Root attacks.

Table 3.1 displays several attacks that fall into four main categories, while Table 2 provides a comprehensive list offeatures that are used to describe the connection records

Table 3.1. Various types of attacks described in four major categories

Remote to User Attacks	Ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster
Probes	Satan, ipsweep, nmap, portsweep
Denial of Service Attacks	Back, land, neptune, pod, smurf, teardrop
User to Root Attacks	Buffer_overflow, loadmodule, per, rootkit

Table 3.2.A complete list of features given in KDD cup 99 dataset

Feature index	Feature name	Description	Type
1	Duration	Length(number of seconds) of the connection	Continuous
2	Protocol_type	Type of the protocol e.gtcp, udp	Symbolic
3	Service	Network service on the destination e.g., http, telnet, etc.	Symbolic
4	Flag	Normal or error status of the connection	Symbolic
5	Src_bytes	Number of data bytes from source to destination	Continuous
6	dst_bytes	Number of data bytes from destination to source	Continuous
7	Land	1 if connection is from/to the same host/port;0 otherwise	Symbolic
8	Wrong_fragment	Number of “wrong” fragments	Continuous
9	Urgent	Number of urgent packets	Continuous
10	Hot	Number of “hot” indicators	Continuous
11	Num_failed_logins	Number of failed login attempts	Continuous
12	Logged_in	1 if successfully logged in : 0 otherwise	Symbolic
13	Num_compromised	Number of “compromised”	Continuous

		conditions	
14	Root_shell	1 if root shell is obtained ; 0 otherwise	Continuous
15	Su_attempted	1 if “su root” command attempted; 0 otherwise	Continuous
16	Num_root	Number of “root” accesses	Continuous
17	Num_file_creations	Number of file creation operations	Continuous
18	Num_shells	Number of shell prompts	Continuous
19	Num_access_files	Number of operations on access control files	Continuous
20	Num_outbound_cmds	Number of outbound commands in an ftp session	Continuous
21	Is_hot_login	1 if the the belongs to the “hot”list; 0 otherwise	Symbolic
22	Is_guest_login	1 if the login is a “guest ” login ;0 otherwise	Symbolic
23	Count	Number of connections to the same same host as the current connection in the past two seconds	Continuous
24	Srv_count	Number of connections to the same service to the same service as the current connection in the past two seconds	Continuous
25	Serror_rate	% of connections that have	Continuous

		"SYN" errors	
26	Srv_serror_rate	% of connections that have "SYN" errors	Continuous
27	Rerror_rate	% of connections that have "REJ" errors	Continuous
28	Srv_rerror_rate	% of connections that have "REJ" errors	Continuous
29	Same_srv_rate	% of connections to the same service	Continuous
30	Diff_srv_rate	% of connections to different services	Continuous
31	Srv_diff_host_rate	% of connections to different host	Continuous
32	Dst_host_srv_count	Count for destination host	Continuous
33	Dst_host_srv_count	Srv count for destination host	Continuous
34	Dst_host_same_srv_rate	Same_srv_rate for destination host	Continuous
35	Dst_host_diff_srv_rate	Diff_srv_rate for destination host	Continuous
36	Dst_host_same_src_port_rate	Same_src_port_rate for destination host	Continuous
37	Dst_host_srv_diff_host_rate	Diff host _rate for destination host	Continuous
38	Dst_host_serror_rate	Serror_rate for destination host	Continuous
39	Dst_host_srv_serror_rate	Srv_serror-rate for destination host	Continuous
40	Dst_host_rerror_rate	Rerror_rate for destination host	Continuous
41	Dst_host_srv_reeror_rate	Srv_serror_rate for destination	Continuous

3.2 Performance Evaluation of Homogenous Boosting Machine Learning Algorithms

In order to meet the second objective, different ensemble methods were examined that were proposed by various experts for intrusion detection classification techniques. These methods were categorized to identify the most suitable homogenous boosting machine learning algorithm that would be able to achieve strong generalization ability. The evaluation of the performance of these homogenous boosting ensemble machine learning algorithms was then conducted to determine which algorithm would result in the highest detection rate for intrusion. The algorithms are AdaBoost, LogitBoost, RealBoost, and MulitBoost. The selection of these ensemble machine learning algorithms was dependent on the nature of the research task, which can be categorized by its input and output. The Boosting family of algorithms has demonstrated a lower error rate and improved classification rates, which is why the listed homogenous boosting ensemble machine learning algorithms were chosen for evaluating the performance of the intrusion detection model. The evaluation of the model's performance was carried out using five performance measures, including sensitivity, specificity, accuracy percentages, precision, and the area under the receiver operating characteristics curve, to determine the most suitable classifier. These measures were used on the obtained dataset to choose the best-fit classifier.

It was possible to learn important details about the dependability of the performance measurements by using kappa statistics. Specificity shows how many instances of each kind are accurately classified, whereas sensitivity shows how many occurrences of each type are correctly classified. The percentage of instances that are correctly classified is known as accuracy. The number of occurrences that are appropriately labelled as defective is known as precision. In multi-classifier systems, Kappa Statistics gauges the similarity of ensembles.

The area under the Receiver Operating Characteristic (ROC) curve, which represents a classifier's accuracy, illustrates the trade-off between true positive (TP) and false positive (FP) rates. An improved classifier has a greater area under the curve.The sensitivity, specificity, percentages of accuracy, precision, and kappa statistical values are shown below.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{---------- Equation 3.2}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad \text{---------- Equation 3.3}$$

$$\text{Specificity} = \frac{TN}{FP+TN} \quad \text{---------- Equation 3.4}$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{---------- Equation 3.5}$$

Where TN and TP stand for true negative and positive whereas FN and FP denote false negative and positive.

$$\text{Kappa Statistics} = \frac{P(A)-P(E)}{1-P(E)} \quad \text{---------- Equation 3.6}$$

Such that (E) is the likelihood that the classifier agreement is subject to possibility or coincidence and (A) is the classifier accuracy.

3.3Comparative Analysis of Homogeneous Boosting Ensemble Algorithm Performance

Based on the evaluation criteria, a comparison analysis of the four classifier models was done in order to increase the detection rate of intrusion. Because each dataset has different location points, this was done using 10-Fold Cross-Validation (10-F C-V) and Holdout. Each cross-validation's fold was subjected to one application of the 10-F C-V before the entire dataset for each of the AdaBoost, LogitBoost, RealBoost, and the MulitBoost machine learning algorithms in total of eleven times for each algorithm. The hold method was used as test data

in the percentage-split ratio 80:20. This is to identify the best homogenous boosting ensemble classifier model for intrusion detection.

Endnote

1. Choudhary, Sarika, and Nishtha Kesswani. "Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 datasets using deep learning in IoT." *Procedia Computer Science* 167 (2020): 1561-1573.

2. Bala, Ritu, and Ritu Nagpal. "A review on kdd cup99 and nslnsl-kdd dataset." *International Journal of Advanced Research in Computer Science* 10, no. 2 (2019).

3. Kumar, Satish, and Sakshi Arora. "A Statistical Analysis on KDD Cup'99 Dataset for the Network Intrusion Detection System." *Applied Soft Computing and Communication Networks: Proceedings of ACN 2019* (2020): 131-157.

CHAPTER FOUR

DATA ANALYSIS

4.1 Data Fuzzification for Intrusion Detection Model Performance Evaluation

When evaluating the effectiveness of the intrusion detection model, we are considering the anomaly-based method. To assess the model's performance, we recommend using the KDD Cup 1999 dataset, which has been divided into five subsets for the detection of four types of attacks: Denial of Service (DOS), User to Root (U2R), Probe, and Remote to Local (R2L).The dataset must first be divided into a number of classes while taking into account the numerous attacks that are present in the dataset. Section 3.2 of Chapter 3 provides a thorough examination of the dataset, and using that information, the dataset's 41 features which include both symbolic and continuous feature contain data on four different forms of attacks as well as information on normal conduct. The continuous features in the dataset, was considered in this research because the major attributes in the dataset are continuous in nature, with this the 34 attributes that were continuous in nature served as the input dataset by removing discrete features. After which the dataset (A) is divided into five subsets of classes based on the class tag given in the dataset $A = \{A_i: 1 \leq i < 5\}$. The class tag describes the various attacks which are identified under four major attacks as stated above along with normal data. The five subsets were used for generating an improved set of fuzzy rules.

4.1.1 Classification of training data

The first stage of the proposed system involves classifying the input data into multiple classes, considering the different types of attacks present in the intrusion detection dataset. The dataset chosen for this analysis is the KDD-Cup 1999 data, which includes four types of attacks and normal behaviour data with 41 attributes that are both continuous and symbolic. However, the proposed system only considers a subset of these attributes. The dataset (D) is

then divided into five subsets of classes based on the class labels provided in the dataset, represented as

D={Di; 1<i<5}. These class labels describe several attacks, including Denial of service, Remote to Local, User to Root, Probe, and normal data.

4.1.2 Strategy for development of fuzzy rules

The five subsets of data are then used for generating a stronger set of fuzzy rules automatically so that the fuzzy system can learn the rules effectively. To find a better set of rules, mining techniques will be used. Here, certain rules derived from frequently occurring single-length items are applied to ensure appropriate learning of the fuzzy system. The process of fuzzy rule generation is given in the following steps.

A. Single-length frequent item mining

To identify important features in a dataset, frequent items are mined from both classes of input data. This involves identifying the characteristics that occur frequently and calculating the frequency of continuous variables within each attribute. By setting a minimal support, one-length frequent items are identified. These items are then categorized as either normal or associated with four types of attacks.

B. Identification of appropriate properties for rule creation

In this phase, the goal is to select the most appropriate attributes to distinguish between normal records and attacks. Since the input data consists of 34 attributes, not all of them are useful for detecting intrusion. To identify the relevant attributes, a deviation approach is used, which relies on the 1-length frequent items that were previously mined. The purpose of this approach is to choose attributes that deviate significantly from the norm and are therefore more likely to be relevant for intrusion detection.

C. Rule generation

The effective attributes selected in the previous step are used to derive rules based on the "max, min" deviation. These rules are produced by comparing the deviation range of the effective attributes for both normal and attack data, and identifying the points where the deviation ranges intersect. These intersection points are then used to generate both definite and indefinite rules.

D. Rule filtering

The rules produced in the previous phase include both definite and indefinite rules. Definite rules have a single classified label in the THEN part, while indefinite rules have two classification label data in the THEN part. A proposed rule filtering technique is used to filter out the indefinite rules and select only the definite rules for learning the fuzzy system.

E. Generating fuzzy rules

The proposed system is designed to automatically generate fuzzy rules using the 1-length frequent items that were previously mined. The fuzzy rules are derived from the definite rules, which have a numerical variable in the IF part and a class label related to either an attack or normal in the THEN part. However, for the fuzzy rules to work, they must only contain linguistic variables. To achieve this, the numerical variable in the IF part of the definite rules is fuzzified, and the THEN part of the fuzzy rule is the same as the consequent part of the definite rule. For example, "IF attribute1 is H, THEN the data is attack" would become a fuzzy rule with the linguistic variable "IF attribute1 is High, THEN the data is attack". These fuzzy rules are then used to train the fuzzy system, improving the effectiveness of the proposed system compared to using fuzzy rules without proper techniques.

4.2.1 LogiBoost Algorithm Performance

It was observed as shown in Table 4.1 using 10-F C-V method that TP and TN were 97.9% correctly predicted for class of attack such as Normal, U2R, DOS, R2L and PROBE as portrayed across the main diagonal of the confusion matrix. There was also misclassification of FP as well as FN as seen on the off-diagonal for class of attack such as NORMAL being classified as U2R and 8 out of 97,277, NORMAL being classified as U2R and 6138 out of 97,277, NORMAL being classified as DOS and 55 out of 97,277, NORMAL being classified as PROBE and 25 out of 97,277. Also, it can be seen that U2R was also misclassified as NORMAL and 28 out of 52, U2R was also classified as DOS and 2 out of 52. Likewise, it can be seen that DOS was misclassified as NORMAL and 193 out of 391,458, DOS was also misclassified as PROBE and 105 out of 391,458. Also, it can be see that R2L was misclassified as NORMAL and 386 out of 1126, and R2L being classified as U2R and 5 out of 1126, R2L being classified as DOS and 153 out of 1126, R2L being classified as PROBE and 1 out of 1126. Also, it can be seen that PROBE being misclassified as DOS and 212 out of 4107, PROBE being misclassified as DOS and 2765 out 4107.

Table 4.1 Confusion matrix forLogiBoostusing cross validation method

	Predicted Class									
	A	**B**	**C**	**D**	**E**	**<--**	**classified**	**as**	**Class of Attack**	**Instances**
Actual Class	91051	8	6138	55	25	|	A	=	NORMAL	97,277
	28	22	2	0	0	|	B	=	U2R	52
	193	0	391160	0	105	|	C	=	DOS	391,458
	386	5	153	581	1	|	D	=	R2L	1126
	212	0	2765	0	1130	|	E	=	PROBE	4107
									Total Instances	**494020**

(Source: Researcher's Model, 2023)

Table 4.2 represents the Confusion matrix for LogiBoost using hold-out method based on input data. The rows represent the actual class of attacks, while the columns represent the predicted class of attacks. The predicted class of attacks are denoted by A, B, C, D, and E, and the actual number of instances of each attack class is shown in the corresponding row.

The diagonal elements of the matrix represent the correctly classified instances of each attack class, while the off-diagonal elements represent the misclassified instances. From the table it can be observed that NORMAL was misclassified as U2R and 2 out of 19413, also, NORMAL was misclassified as DOS and 1199 out of 19413, and NORMAL was misclassified as R2L and 8 out of 19413, and NORMAL was misclassified as PROBE and 6 out of 19413 and U2R was misclassified as NORMAL and 9 out of 13. Also, it can be observed that DOS was misclassified as NORMAL and 36 out of 78352, and DOS was misclassified as PROBE and 20 out of 78352.

It can also be observed that R2L was misclassified as NORMAL and 41 out of 226, and R2L misclassified as U2R and 2 out for 226, and R2L misclassified as DOS and 27 out of 226. Also, PROBE was misclassified as NORMAL and 39 out of 800, and PROBE was misclassified DOS and 522 out of DOS.

Table 4.2 Confusion matrix forLogiBoostusing hold-out method.

	Predicted Class									
	A	**B**	**C**	**D**	**E**	**<--**	**classified**	**as**	**Class of Attack**	**Instances**
Actual Class	18198	2	1199	8	6	|	A	=	Normal	19413
	9	4	0	0	0	|	B	=	U2R	13
	36	0	78296	0	20	|	C	=	DOS	78352
	41	2	27	156	0	|	D	=	R2L	226
	39	0	522	0	239	|	E	=	PROBE	800
									Total Instances	**98804**

(Source: Researcher's Model, 2023)

4.2.2RealBoost Algorithm Performance

It was observed as shown in Table 4.3 using 10-F C-V method that TP and TN were 98.1% correctly predicted for class of attack such as Normal, U2R, DOS, R2L and PROBE as portrayed across the main diagonal of the confusion matrix. There was also misclassification of FP as well as FN as seen on the off diagonal for class of attack such as NORMAL being classified as DOS and 4965 out of 97277, NORMAL being classified as R2L and 228 out of 97277, NORMAL being classified as PROBE and 225 out of 97277. Also, U2R was

classified as NORMAL and 46 out of 52, U2R was classified as DOS and 3 out of 52, U2R was classified as PROBE and 1 out of 52. Also, it can be observed that DOS was classified as NORMAL and 325 out of 391458, DOS was classified as R2L and 2 out of 391458, DOS was classified as PROBE and 70 out of 391458. Also, it was observed that R2L was classified as NORMAL and 687 out of 1126, R2L was classified as DOS and 18 out of 1126, R2L was classified as PROBE and 6 out of 11126. Finally, it was observed that PROBE was classified as NORMAL and 222 out of 4107, PROBE was classified as DOS and 2766 out of 4107, PROBE was classified as R2L and 2 out of 4107 instances.

Table 4.3 Confusion matrix for RealBoost using cross validation method.

	Predicted Class									
	A	**B**	**C**	**D**	**E**	**<--**	**classified**	**as**	**Class of Attack**	**Instances**
Actual Class	91859	0	4965	228	225	|	A	=	Normal	97277
	46	0	3	2	1	|	B	=	U2R	52
	325	0	391061	2	70	|	C	=	DOS	391458
	687	0	18	415	6	|	D	=	R2L	1126
	222	0	2766	2	1117	|	E	=	PROBE	4107
									Total Instances	**494020**

(Source: Researcher's Model, 2023)

Table 4.4 represents the Confusion matrix forRealBoost using hold-out method based on input data. The rows represent the actual class of attacks, while the columns represent the predicted class of attacks. The predicted class of attacks are denoted by A, B, C, D, and E, and the actual number of instances of each attack class is shown in the corresponding row. The diagonal elements of the matrix represent the correctly classified instances of each attack class, while the off-diagonal elements represent the misclassified instances. From the table, it can be observed that NORMAL was classified as DOS and 1019 out of 19413, NORMAL was classified as R2L and 18 out of 19413, NORMAL was classified as PROBE and 105 out of 19413. Also, it can be gathered that U2R was classified as NORMAL and 11 out of 13, U2R was classified as DOS and 1 out of 13 , U2R was classified as R2L and 1 out of 13. Also, it can be observed that DOS was classified as NORMAL and 35 out of 78352, DOS

was classified as PROBE and 14 out of 78352. IT was observed that R2L was classified as NORMAL and 72 out of 226, R2L was classified as DOS and 4 out of 226, R2L was classified as PROBE and 1 out of 226. Finally, PROBE was classified as NORMAL and 32 out of 800, PROBE was classified as DOS and 521 out of 800 instances.

Table 4.4 Confusion matrix forRealBoostusing hold-out method.

	Predicted Class									
	A	**B**	**C**	**D**	**E**	**<--**	**classified**	**as**	**Class of Attack**	**Instances**
Actual Class	18271	0	1019	18	105	|	A	=	Normal	19413
	11	0	1	1	0	|	B	=	U2R	13
	35	0	78303	0	14	|	C	=	DOS	78352
	72	0	4	149	1	|	D	=	R2L	226
	32	0	521	0	247	|	E	=	PROBE	800
									Total Instances	**98804**

(Source: Researcher's Model, 2023)

4.2.3AdaBoost Algorithm Performance.

It was observed as shown in Table 4.5 using 10-F C-V method that TP and TN were 95.6% correctly predicted for class of attack such as Normal, U2R, DOS, R2L and PROBE as portrayed across the main diagonal of the confusion matrix. There was also misclassification of FP as well as FN as seen on the off diagonal for class of attack such as NORMAL being classified as DOS and 14197 out of 97277, Also, it can be observed that U2R was classified as NORMAL and 50 out of 52, U2R was classified as DOS and 2 out of 52. Also, it can be observed that DOS was classified as NORMAL and 2204 out of 391458. It was observed that R2L was classified as NORMAL and 401 out of 1126. Also, it was observed that PROBE was classified as NORMAL and 25 out 4107, and PROBE was classified as DOS and 4082 out of 4107 instances.

Table 4.5 Confusion matrix forAdaBoostusing cross validation method.

			Predicted Class								
	A	**B**	**C**	**D**	**E**	**<--**	**classified**	**as**	**Class of Attack**	**Instances**
Actual Class	83080	0	14197	0	0	|	A	=	Normal	97277
	50	0	2	0	0	|	B	=	U2R	52
	2204	0	389254	0	0	|	C	=	DOS	391458
	401	0	725	0	0	|	D	=	R2L	1126
	25	0	4082	0	0	|	E	=	PROBE	4107
									Total Instances	**494020**

(Source: Researcher's Model, 2023)

Table 4.6 represents the Confusion matrix forAdaboost using hold-out method based on input data. The rows represent the actual class of attacks, while the columns represent the predicted class of attacks.The predicted class of attacks are denoted by A, B, C, D, and E, and the actual number of instances of each attack class is shown in the corresponding row.The diagonal elements of the matrix represent the correctly classified instances of each attack class, while the off-diagonal elements represent the misclassified instances. It can be observed from the table that only the NORMAL and the DOS were rightly classified here.

Table 4.6 Confusion matrix forAdaBoostusing hold-out method.

			Predicted Class								
	A	**B**	**C**	**D**	**E**	**<--**	**classified**	**as**	**Class of Attack**	**Instances**
Actual Class	19413	0	0	0	0	|	A	=	Normal	19413
	0	0	13	0	0	|	B	=	U2R	13
	0	0	78352	0	0	|	C	=	DOS	78352
	0	0	226	0	0	|	D	=	R2L	226
	0	0	800	0	0	|	E	=	PROBE	800
									Total Instances	**98804**

(Source: Researcher's Model, 2023)

4.3 Comparative Analysis of Homogeneous Boosting Ensemble Algorithm Performance

Accounting for the four homogeneous boosting machine learning algorithms as discussed in sections 4.2.1 to 4.2.4, there is a need to evaluate their performances with the intention of comparative analysis of the output model intrusion detection model based on homogenous

ensemble boosting technique.The following benchmark were considered accuracy, sensitivity, specificity, precision, kappa statistics and AUROC.

4.3.1 Evaluation based on Accuracy.

The table 4.7 below presents the evaluation of different homogeneous boosting ensemble algorithm performance using two different evaluation methods, namely Holdout (80:20) and 10-Fold Cross-Validation, based on accuracy as the performance metric.

- MLAs: The table lists several MLAs, including LogicBoost, MultiBoost, RealBoost, and AdaBoost, which are being evaluated for their performance.
- Evaluation Methods: Two evaluation methods are used: Holdout (80:20) and 10-Fold Cross-Validation. Holdout (80:20) refers to splitting the dataset into 80% for training and 20% for testing, while 10-Fold Cross-Validation involves dividing the dataset into 10 equal folds, using 9 folds for training and 1-fold for testing in a rotating fashion.
- Accuracy: Accuracy is a measure of a model's ability to correctly classify instances out of all the instances in the data.
- Values: The table presents the accuracy values in percentage for each MLA under the two evaluation methods. For example, under the Holdout (80:20) evaluation method, LogicBoost has an accuracy of 98.1%, MultiBoost has an accuracy of 76.5%, RealBoost has an accuracy of 98.1%, and AdaBoost has an accuracy of 98.9%. Under the 10-Fold Cross-Validation evaluation method, the accuracy values remain the same for LogicBoost and RealBoost, while MultiBoost has an accuracy of 76.5% and AdaBoost has an accuracy of 95.6%.

Based on the table, LogicBoost, RealBoost, and AdaBoost appear to have higher accuracy values compared to MultiBoost, indicating better performance in terms of overall classification accuracy.

Table 4.7: Evaluation based on Accuracy.

Homogenous Algorithm	Holdout (80:20)	10-Fold Cross-Validation
LogicBoost	98.1	98.0
MultiBoost	76.5	76.5
RealBoost	98.1	98.1
AdaBoost	98.9	95.6

(Source: Researcher's Model, 2023)

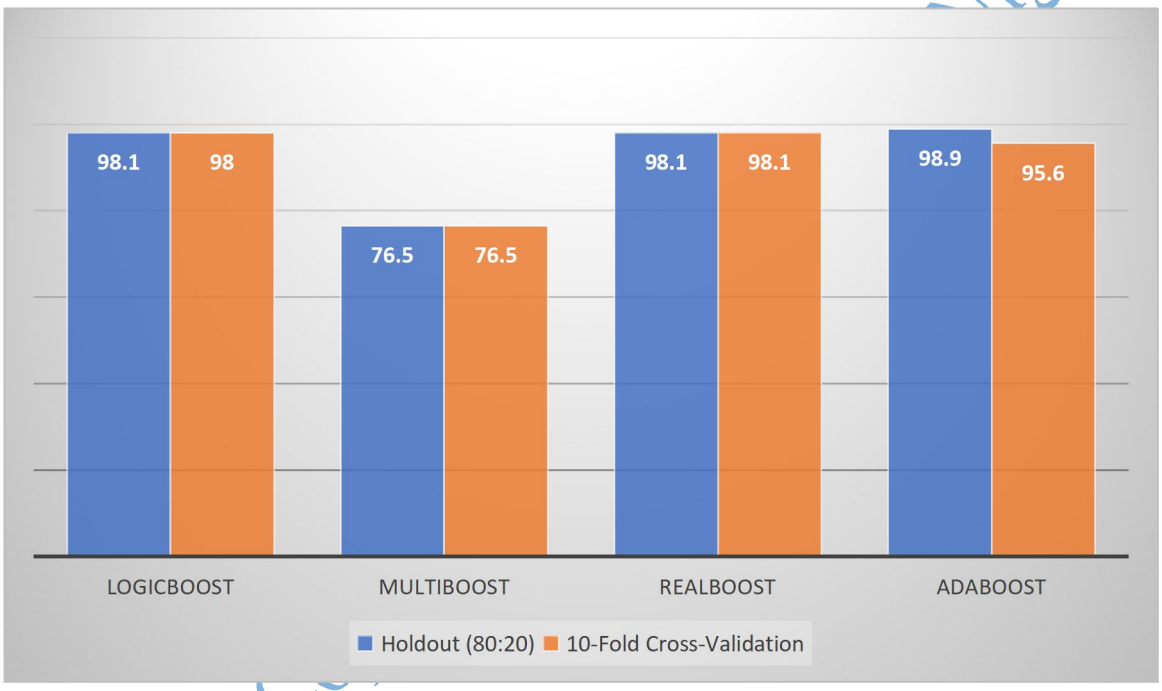


Figure 4.13: Evaluation based on Accuracy (Source: Researcher's Model, 2023)

4.3.2 Evaluation based on Sensitivity.

The table 4.8 below presents the sensitivity values (also known as true positive rate or recall) for different homogeneous boosting ensemble algorithm performance evaluated using two methods: Holdout (80:20) and 10-Fold Cross-Validation.

- Holdout (80:20): Under this evaluation method, the sensitivity values for the MLAs are as follows: LogicBoost - 98.1%, MultiBoost - 76.5%, RealBoost - 98.1%, and AdaBoost - 98.9%.

- 10-Fold Cross-Validation: Under this evaluation method, the sensitivity values for the MLAs are slightly lower for AdaBoost at 95.6%, while the sensitivity values for LogicBoost, MultiBoost, and RealBoost remain the same as in the Holdout (80:20) evaluation.

Based on the sensitivity values, LogicBoost, RealBoost, and AdaBoost appear to have higher sensitivity values compared to MultiBoost, indicating better performance in correctly identifying positive instances in the data.

Table 4.8: Evaluation based on Sensitivity.

Homogenous Algorithm	Holdout (80:20)	10-Fold Cross-Validation
LogicBoost	98.1	98.0
MultiBoost	76.5	76.5
RealBoost	98.1	98.1
AdaBoost	98.9	95.6

(Source: Researcher's Model, 2023)

Figure 4.13: Evaluation based on Sensitivity (Source: Researcher's Model, 2023)

4.3.3 Evaluation based on Specificity.

The table presents the specificity values (also known as true negative rate) for different homogeneous boosting ensemble algorithm performance evaluated using two methods: Holdout (80:20) and 10-Fold Cross-Validation.

- Holdout (80:20): Under this evaluation method, the specificity values for the MLAs are as follows: LogicBoost - 97.9%, MultiBoost - 45.2%, RealBoost - 78.3%, and AdaBoost - 47.0%.
- 10-Fold Cross-Validation: Under this evaluation method, the specificity values for the MLAs are slightly lower for RealBoost at 70.7%, while the specificity values for LogicBoost, MultiBoost, and AdaBoost remain the same as in the Holdout (80:20) evaluation.

Based on the specificity values, LogicBoost and RealBoost appear to have higher specificity values compared to MultiBoost and AdaBoost, indicating better performance in correctly identifying negative instances in the data.

Table 4.9: Evaluation based on Specificity.

Homogenous Algorithm	Holdout (80:20)	10-Fold Cross-Validation
LogicBoost	97.9	97.8
MultiBoost	45.2	45.2
RealBoost	78.3	70.7
AdaBoost	47.0	49.8

(Source: Researcher's Model, 2023)

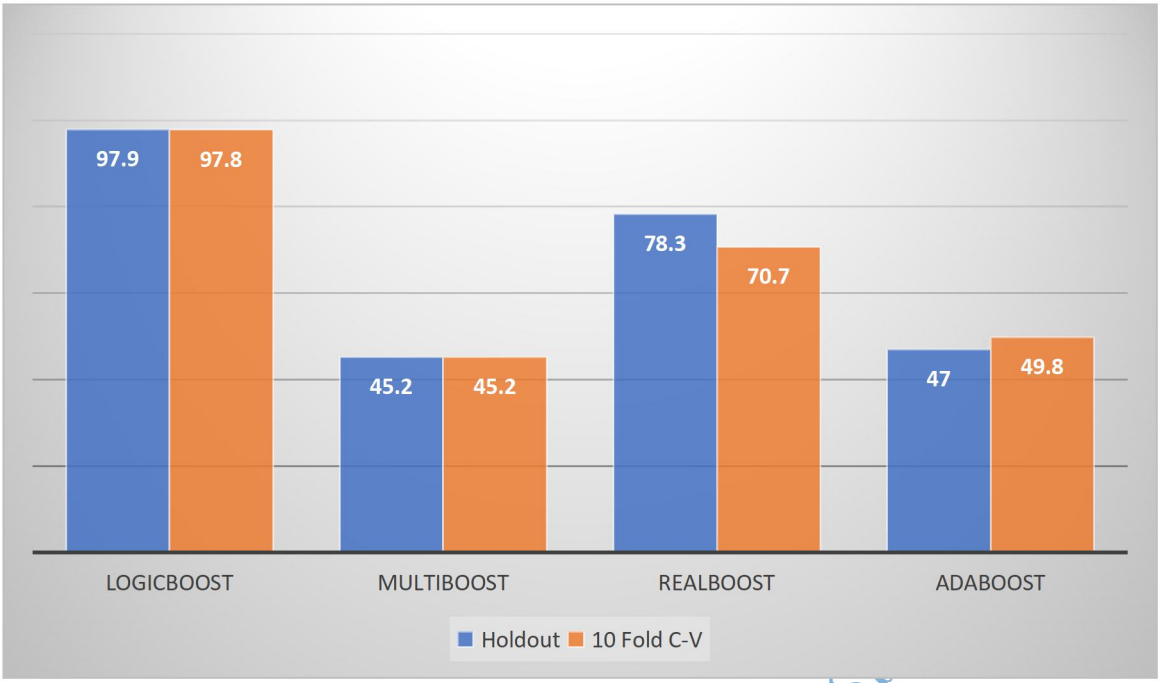


Figure 4.13: Evaluation based on specificity (Source: Researcher's Model, 2023)

4.3.4 Evaluation based on Precision.

The table 4.10 below shows the evaluation results of four different homogenous algorithms (LogicBoost, MultiBoost, RealBoost, and AdaBoost) based on precision, using two different evaluation techniques: holdout with an 80:20 split and 10-fold cross-validation.

- For LogicBoost, the precision is 98.0% with holdout (80:20) evaluation and 97.9% with 10-fold cross-validation.
- For MultiBoost, the precision is 41.6% with both holdout (80:20) evaluation and 10-fold cross-validation.
- For RealBoost, the precision is 88.3% with holdout (80:20) evaluation and 84.8% with 10-fold cross-validation.
- For AdaBoost, the precision is 49.6% with holdout (80:20) evaluation and 48.0% with 10-fold cross-validation.

Precision is a measure of the accuracy of a classification model, representing the proportion of true positive predictions out of the total positive predictions. Based on the table,

LogicBoost has the highest precision, while MultiBoost has the lowest precision among the four algorithms, regardless of the evaluation technique used. RealBoost shows slightly lower precision with 10-fold cross-validation compared to holdout evaluation, while AdaBoost shows a similar trend but with slightly higher precision in holdout evaluation compared to 10-fold cross-validation.

Table 4.10: Evaluation based on Precision.

Homogenous Algorithm	Holdout (80:20)	10-Fold Cross-Validation
LogicBoost	98.0	97.9
MultiBoost	41.6	41.6
RealBoost	88.3	84.8
AdaBoost	49.6	48.0

(Source: Researcher's Model, 2023)

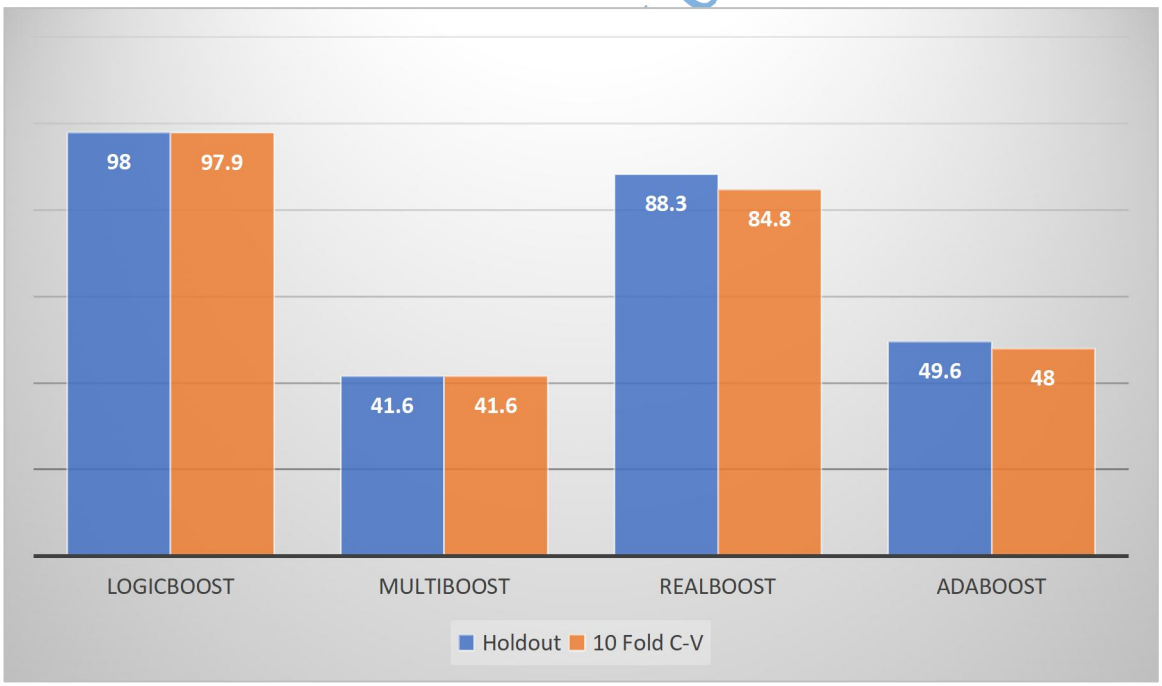


Figure 4.13: Evaluation based on Precision (Source: Researcher's Model, 2023)

4.3.5 Evaluation based on Kappa Statistics.

The table presents the evaluation results of four different homogenous algorithms (LogicBoost, MultiBoost, RealBoost, and AdaBoost) based on Kappa statistics, using two different evaluation techniques: holdout with an 80:20 split and 10-fold cross-validation.

- For LogicBoost, the Kappa statistic is 93.9% with holdout (80:20) evaluation and 93.6% with 10-fold cross-validation.
- For MultiBoost, the Kappa statistic is 53.8% with both holdout (80:20) evaluation and 10-fold cross-validation.
- For RealBoost, the Kappa statistic is 94.2% with holdout (80:20) evaluation and 94.0% with 10-fold cross-validation.
- For AdaBoost, the Kappa statistic is 96.7% with holdout (80:20) evaluation and 85.8% with 10-fold cross-validation.

Kappa statistic is a measure of agreement between predicted and actual values in a classification model, taking into account the agreement that could occur by chance. Higher Kappa values indicate better agreement between predicted and actual values. Based on the table, AdaBoost has the highest Kappa statistic in holdout (80:20) evaluation, but a lower Kappa statistic in 10-fold cross-validation compared to other algorithms. LogicBoost and RealBoost show relatively high Kappa statistics consistently across both evaluation techniques, while MultiBoost has the lowest Kappa statistic among the four algorithms in both evaluation techniques.

Table 4.11: Evaluation based on Kappa Statistics.

Homogenous Algorithm	Holdout (80:20)	10-Fold Cross-Validation
LogicBoost	93.9	93.6
MultiBoost	53.8	53.8
RealBoost	94.2	94.0

AdaBoost	96.7	85.8

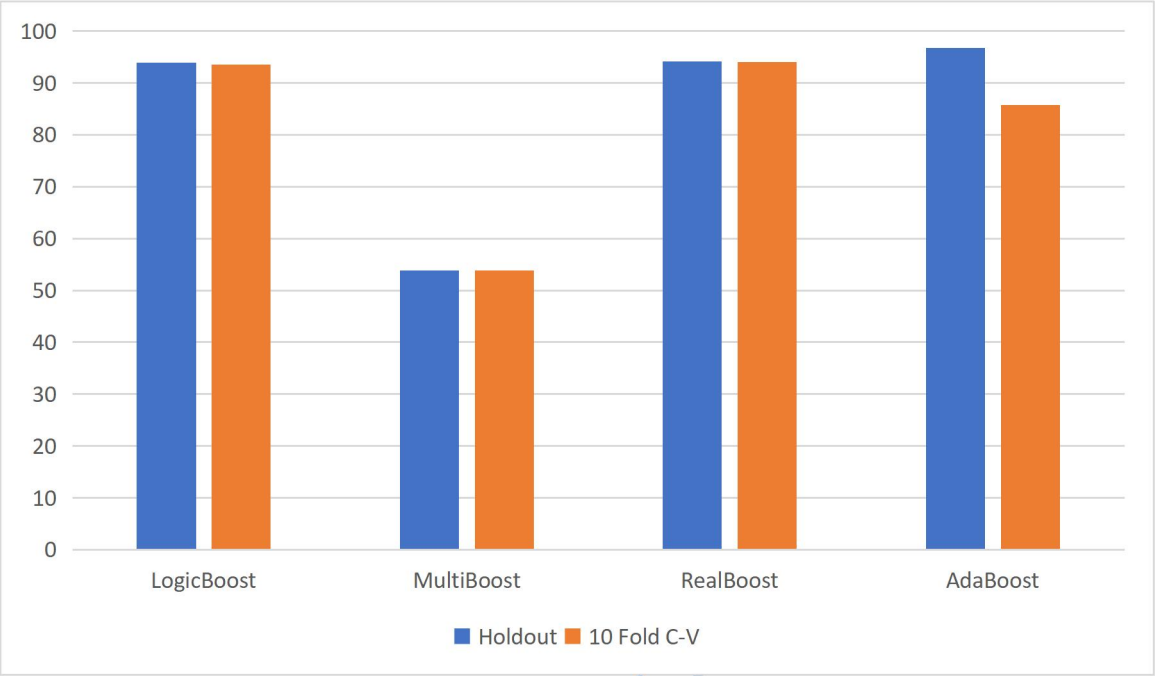


Figure 4.13: Evaluation based on Kappa Statistics (Source: Researcher's Model, 2023)

4.3.6 Evaluation based on AUROC.

The table presents the evaluation results of four different homogenous algorithms (LogicBoost, MultiBoost, RealBoost, and AdaBoost) based on AUROC (Area Under the Receiver Operating Characteristic) using two different evaluation techniques: holdout with an 80:20 split and 10-fold cross-validation.

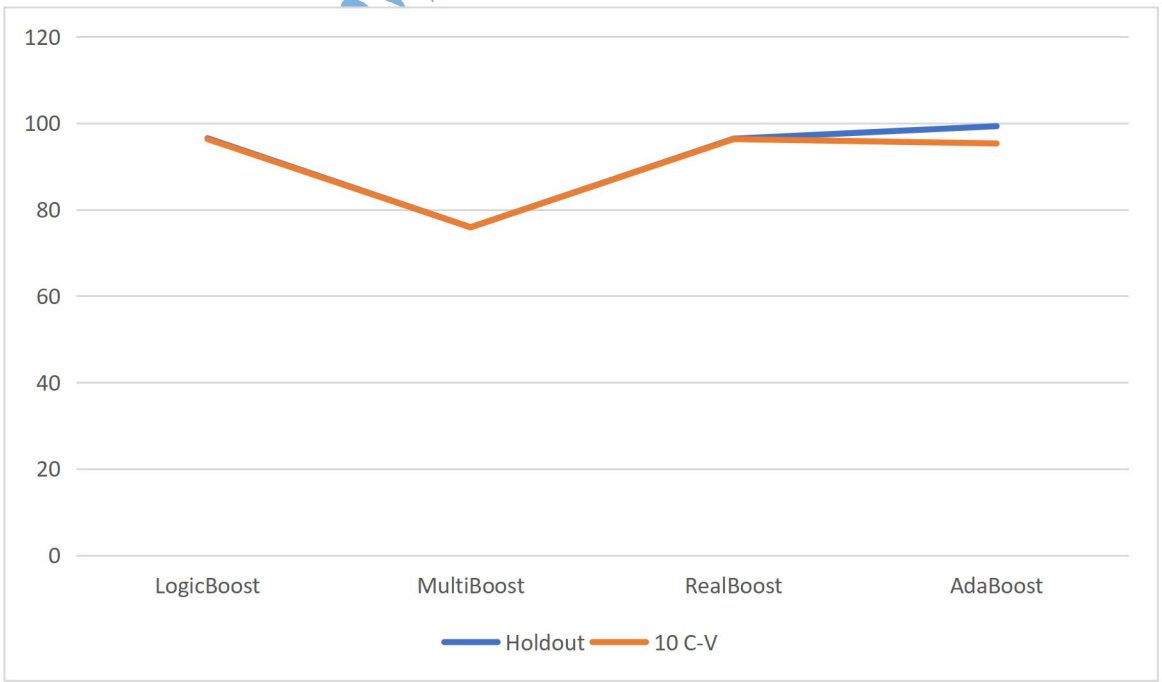
- For LogicBoost, the AUROC is 96.5% with holdout (80:20) evaluation and 96.3% with 10-fold cross-validation.
- For MultiBoost, the AUROC is 75.9% with both holdout (80:20) evaluation and 10-fold cross-validation.
- For RealBoost, the AUROC is 96.4% with holdout (80:20) evaluation and 96.3% with 10-fold cross-validation.

- For AdaBoost, the AUROC is 99.3% with holdout (80:20) evaluation and 95.3% with 10-fold cross-validation.

AUROC is a measure of the overall performance of a classification model, indicating the ability of the model to correctly discriminate between positive and negative instances. Higher AUROC values indicate better model performance. Based on the table, AdaBoost has the highest AUROC in holdout (80:20) evaluation, but a lower AUROC in 10-fold cross-validation compared to other algorithms. LogicBoost and RealBoost show relatively high AUROC values consistently across both evaluation techniques, while MultiBoost has the lowest AUROC among the four algorithms in both evaluation techniques.

Table 4.11: Evaluation based on AUROC.

Homogenous Algorithm	Holdout (80:20)	10-Fold Cross-Validation
LogicBoost	96.5	96.3
MultiBoost	75.9	75.9
RealBoost	96.4	96.3
AdaBoost	99.3	95.3



4.14: Evaluation based on AUROC (Source: Researcher's Model, 2023)

4.3.6 Homogenous Algorithm Performance EvaluationSummary in Cross Validation and Hold-Out Methods

The performance of the four homogeneous boosting machine learning algorithms namely LogicBoost,MultiBoost, RealBoost,and AdaBoostusing the metrics accuracy percentages, sensitivity, specificity, precision, kappa statistics and area under the receiver operating characteristic in 10-F C-V method and hold-out method are shown in Table 4.16 and Table 4.17 respectively. In respect to the performance LogicBoost, RealBoost and AdaBoost performed brilliantly well in relative to one metric or the other.

4.3.8 Algorithm Performance Evaluation Summary in Cross Validation Method

Table 4.12. Result summary of cross validation method for Homogenous Algorithms

Homogenous Algorithms	Accuracy (%)	TP_Rate (Sensitivity) (%)	TN_Rate (Specificity) (%)	Precision (%)	Kappa_Statistics (%)	AUROC (%)
LogicBoost	98.0	98.0	97.8	97.9	93.6	96.3
MultiBoost	76.5	76.5	45.2	41.6	53.8	75.9
RealBoost	98.1	98.1	70.7	84.8	94.0	96.3
AdaBoost	95.6	95.6	49.8	48.0	85.8	95.3

(Source: Researcher's Model, 2023)

The table provides a summary of cross-validation results for different homogenous algorithms, including LogicBoost, MultiBoost, RealBoost, and AdaBoost. The evaluation metrics presented in the table include Accuracy, TP_Rate (Sensitivity), TN_Rate (Specificity), Precision, Kappa Statistics, and AUROC (Area Under the Receiver Operating Characteristic).

- Accuracy (%) represents the percentage of correctly classified instances by the algorithm.
- TP_Rate (Sensitivity) (%) represents the percentage of true positive predictions, or the proportion of actual positive instances correctly predicted as positive.

- TN_Rate (Specificity) (%) represents the percentage of true negative predictions, or the proportion of actual negative instances correctly predicted as negative.
- Precision (%) represents the percentage of true positive predictions out of the total positive predictions made by the algorithm.
- Kappa Statistics (%) measures the agreement between the predicted and actual classes, taking into account the possibility of agreement by chance.
- AUROC (%) represents the area under the curve of the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate.

From the table, it can be observed that LogicBoost and RealBoost have higher accuracy, sensitivity, specificity, precision, Kappa Statistics, and AUROC values compared to MultiBoost and AdaBoost. This suggests that LogicBoost and RealBoost have better overall performance in terms of classification accuracy and predictive ability based on the evaluation metrics used.

4.3.9 Summary of Algorithm Performance Evaluation in Hold-Out Method

Table 4.13. Result summary of holdout method for Homogenous Algorithms.

Homogenous Algorithms	Accuracy (%)	TP_Rate (Sensitivity) (%)	TN_Rate (Specificity) (%)	Precision (%)	Kappa_Statistics (%)	AUROC (%)
LogicBoost	98.1	98.1	97.9	98.0	93.9	96.5
MultiBoost	76.5	76.5	45.2	41.6	53.8	75.9
RealBoost	98.1	98.1	78.3	88.3	94.2	96.4
AdaBoost	98.9	98.9	47.0	49.6	96.7	99.3

(Source: Researcher's Model, 2023)

Based on the table

1. LogicBoost: This algorithm has a high accuracy of 98.1%, indicating that it correctly classifies data points in the dataset. It also has a sensitivity (true positive rate) and specificity (true negative rate) of 98.1% and 97.9% respectively, which indicates a balanced performance in correctly identifying both positive and negative instances. The precision is 98.0%, indicating a high level of accuracy in predicting positive instances. The kappa statistics, which measures the agreement between predicted and actual values, is 93.9%. The AUROC (Area Under the Receiver Operating Characteristic) is 96.5%, which is indicative of a good overall performance of the algorithm in distinguishing between positive and negative instances.
2. MultiBoost: This algorithm has a lower accuracy of 76.5%, compared to LogicBoost, indicating a lower overall performance in correctly classifying instances. The sensitivity is 76.5%, which is also lower compared to LogicBoost, while the specificity is 45.2%, indicating a lower ability to correctly identify negative instances. The precision is 41.6%, which is relatively low, indicating a higher rate of false positives. The kappa statistics is 53.8%, indicating moderate agreement between predicted and actual values. The AUROC is 75.9%, which is lower compared to LogicBoost, indicating a lower performance in distinguishing between positive and negative instances.
3. RealBoost: This algorithm has the same accuracy as LogicBoost at 98.1%, indicating a high level of accuracy in correctly classifying instances. The sensitivity is 98.1%, which is the same as the accuracy, indicating that it correctly identifies positive instances. The specificity is 78.3%, indicating a relatively lower ability to correctly identify negative instances compared to LogicBoost. The precision is 88.3%, indicating a high level of accuracy in predicting positive instances. The kappa statistics is 94.2%, indicating a high level of agreement between predicted and actual

values. The AUROC is 96.4%, which is slightly lower compared to LogicBoost, but still indicative of a good overall performance in distinguishing between positive and negative instances.

4. AdaBoost: This algorithm has the highest accuracy among all the algorithms at 98.9%, indicating a high level of accuracy in correctly classifying instances. The sensitivity is also 98.9%, which is the same as the accuracy, indicating that it correctly identifies positive instances. The specificity is 47.0%, indicating a relatively lower ability to correctly identify negative instances compared to the other algorithms. The precision is 49.6%, which is relatively low, indicating a higher rate of false positives. However, the kappa statistics is the highest among all the algorithms at 96.7%, indicating a high level of agreement between predicted and actual values. The AUROC is also the highest among all the algorithms at 99.3%, indicating an excellent overall performance in distinguishing between positive and negative instances.

4.4 Discussion of Findings

The findings based on the tables for the holdout and cross-validation methods for homogenous algorithms can be summarized as follows:

1. LogicBoost: The algorithm consistently performs well in both holdout and cross-validation methods, with high accuracy (98.1% in holdout and 98.0% in cross-validation), sensitivity (TP_Rate) around 98%, specificity (TN_Rate) around 97-98%, precision around 98%, and AUROC around 96-96.5%. The kappa statistics also indicate a high level of agreement between predicted and actual values, ranging from 93.6% to 93.9%. These results suggest that LogicBoost is a reliable algorithm with consistent and high-performing results in both evaluation methods.
2. MultiBoost: The algorithm shows relatively lower performance compared to LogicBoost in both holdout and cross-validation methods, with lower accuracy

(76.5%), sensitivity (TP_Rate) around 76.5%, specificity (TN_Rate) around 41.6-45.2%, precision around 41.6%, and AUROC of 75.9%. The kappa statistics also indicate moderate agreement between predicted and actual values, ranging from 53.8% to 53.8%. These results suggest that MultiBoost may have limitations in accurately classifying instances compared to LogicBoost.

3. RealBoost: The algorithm shows variable performance in both holdout and cross-validation methods. In holdout, it has high accuracy (98.1%) and sensitivity (TP_Rate) around 98%, but lower specificity (TN_Rate) of 78.3% and precision of 88.3%. The kappa statistics and AUROC are also high at 94.2% and 96.4% respectively. In cross-validation, the algorithm shows similar accuracy (98.1%) and sensitivity (TP_Rate) as in holdout, but higher specificity (TN_Rate) of 70.7-84.8%, precision of 84.8%, and kappa statistics of 94.0%. These results suggest that RealBoost performs well in terms of accuracy and sensitivity, but may have variability in specificity and precision.
4. AdaBoost: The algorithm shows high accuracy (98.9% in holdout and 95.6% in cross-validation) and sensitivity (TP_Rate) around 98-99% in both holdout and cross-validation methods. However, the specificity (TN_Rate) and precision are relatively lower, ranging from 47.0% to 49.8% and 48.0-49.6% respectively. The kappa statistics are high at 96.7-85.8%, and AUROC is excellent at 99.3-95.3%. These results suggest that AdaBoost may have limitations in correctly identifying negative instances, but overall performs well in terms of accuracy, sensitivity, and overall agreement between predicted and actual values.

In conclusion, based on the findings from the tables, LogicBoost and RealBoost consistently show high performance in both holdout and cross-validation methods, while MultiBoost and AdaBoost show relatively lower performance in certain metrics. However, it is important to

consider the specific context and requirements of the task at hand when interpreting these results and selecting the most appropriate algorithm for a particular use case.

The research evaluated the outcome of the results in contrast to other applications in determining the helpfulness of the result in the perspective of algorithm used, feature selection, validation technique, dataset, conditioning variable categorization, applicable on web, applicable on mobile app and model selection. Table 4.14 shows the evaluation results.

Table 4.14:Evaluation Results of Intrusion detection System

Author	Type	Ensemble scheme	Base Learner(s)	Feature selection	Validation Technique	Dataset	Best results (%)
1	Misuse	Voting	C4.5, SVM,KNN	causal feature selection algorithm	10CV	NSL-KDD	Accuracy: 99.45;precision:99.4; recall: 99.5; F1:99.3
2	Anomaly	Voting	C4.5, RF, CART	CFS+PSO	10CV	NSL-KDD	Accuracy: 99.0; FPR: 0.2
3	Misuse	Bagging	REPT	Clustering	Hold-out, 10CV	NSL-KDD	Holdout : accuracy:81.30;FPR:14.8; 10CV: accuracy:99.68
4	Anomaly	Bagging	PART	Genetic Algorithm	Hold-out,10CV	NSL-KDD	Holdout : accuracy:781.37;FPR:17.2; TPR: 78.4, TPR:99.7

5	Misuse	AB	NB, DT, MLP, SVM, k-NN	causal feature selection	Hold-Out	KDD-Cup 99	Sensitivity: 76.0; specificity: 99.05
6	Anomaly	Voting	RF	Principal component Analysis	10CV	Private	FPR: 7.6; FNR: 0.28
7	Anomaly	RF	-	Accelerated Genetic Algorithm and Rough Set Theory	Hold-Out	NSL-KDD	Accuracy: 80.67
8	Misuse	RF	-	particle swarm optimization	10CV	KDD-Cup99	Accuracy: 96.78; FPR: 0.155
9	Misuse	voting	ET,RF,bagging	Extra trees	Holdout	AWID	Accuracy: 96.32; precision: 96.0; recall: 96.0
10	Misuse	ET, RF, AB	-	MIC	Hold-out	KDDD-Cup 99	Accuracy: 94.1
11	Misuse	Boosting	SVM	Consistency	10CV	NSL-KDD	Accuracy: 99.0
12	Anom	RF,	-	FSR, BER	Hold-	KDD	TPR: 99.9; FPR: 0

	aly	REPT			out	Cup 99, NSL-KDD	
13	Anomaly	RF	LR	RFE	Bootstrap	Malware (UOC), CAIDA, UNSW-NB	Accuracy: 98.90; FPR: 2.81
14	Anomaly	RF	-	Gain Ratio	Hold-out	KDDCUP 99	Accuracy: 94.4
15	Anomaly	RF	GNP	Random feature selection	Hold-out	NSL-KDD	Accuracy: 83.2; FPR: 25.1
16	Anomaly	RF	-	Information gain, symmetrical uncertainty, CFS	Hold-out, 10CV	NSL-KDD	Accuracy: 85.06; TPR: 85.1; FPR: 12.2; precision: 87.5; F1: 85.1
Author	Anomaly	Boosting	Fuzzy Logic		Hold-out, 10CV	KDDCUP 99	Holdout:Accuracy:98.1, Precision: 97.9, Recall: 98.1 10CV:Accuracy:98.0,Pr

	ecision:97.8, Recall:98.0

Endnotes

1. Tama, Bayu Adhi, Marco Comuzzi, and Kyung-Hyune Rhee. "TSE-IDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system." *IEEE access* 7 (2019): 94497-94507.

2. Bhattacharya, Sweta, Praveen Kumar Reddy Maddikunta, Rajesh Kaluri, Saurabh Singh, Thippa Reddy Gadekallu, Mamoun Alazab, and Usman Tariq. "A novel PCA-firefly based XGBoost classification model for intrusion detection in networks using GPU." *Electronics* 9, no. 2 (2020): 219.

3. Gaikwad, D. P., and Ravindra C. Thool. "Intrusion detection system using bagging ensemble method of machine learning." In *2015 international conference on computing communication control and automation*, pp. 291-295. IEEE, 2015.

4. Prachi, Heena Malhotra, and Prabha Sharma. "Intrusion detection using machine learning and feature selection." ***International Journal of Computer Network and Information security*** 11, no. 4 (2019): 43-52.

5. Sornsuwit, Ployphan, and SaichonJaiyen. "A new hybrid machine learning for cybersecurity threat detection based on adaptive boosting." *Applied Artificial Intelligence* 33, no. 5 (2019): 462-482.

6. Go, Gwang-Myong, Seok-Jun Bu, and Sung-Bae Cho. "Insider attack detection in database with deep metric neural network with Monte Carlo sampling." ***Logic Journal of the IGPL*** 30, no. 6 (2022): 979-992.

7. Hedar, Abdel-Rahman, Majid Almaraashi, Alaa E. Abdel-Hakim, and Mahmoud Abdulrahim. "Hybrid machine learning for solar radiation prediction in reduced feature spaces." *Energies* 14, no. 23 (2021): 7970.

8. Alduailij, Mona, Qazi Waqas Khan, Muhammad Tahir, Muhammad Sardaraz, Mai Alduailij, and Fazila Malik. "Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method." *Symmetry* 14, no. 6 (2022): 1095.

9. Ketkhaw, Apisak, and Sakchai Thipchaksurat. "Location Prediction of Rogue Access Point Based on Deep Neural Network Approach." ***Journal of Mobile Multimedia*** (2022): 1063-1078.

10. Gao, Ni, Ling Gao, Quanli Gao, and Hai Wang. "An intrusion detection model based on deep belief networks." In *2014 Second international conference on advanced cloud and big data*, pp. 247-252. IEEE, 2014.

11. Thaseen, Ikram Sumaiya, and Cherukuri Aswani Kumar. "Intrusion detection model using fusion of chi-square feature selection and multi class SVM." ***Journal of King Saud University-Computer and Information Sciences*** 29, no. 4 (2017): 462-472.

12. Hussain, Lal, Adeel Ahmed, Sharjil Saeed, Saima Rathore, Imtiaz Ahmed Awan, Saeed Arif Shah, Abdul Majid, Adnan Idris, and Anees Ahmed Awan. "Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies." *Cancer Biomarkers* 21, no. 2 (2018): 393-413.

13. Wang, Pin, En Fan, and Peng Wang. "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning." *Pattern Recognition Letters* 141 (2021): 61-67.

14. Appiahene, Peter, Yaw Marfo Missah, and Ussiph Najim. "Predicting bank operational efficiency using machine learning algorithm: comparative study of decision tree, random forest, and neural networks." *Advances in fuzzy systems* 2020 (2020): 1-12.

15. Sun, Yuantian, Guichen Li, and Junfei Zhang. "Developing hybrid machine learning models for estimating the unconfined compressive strength of jet grouting composite: a comparative study." *Applied Sciences* 10, no. 5 (2020): 1612.

16. Krishnaveni, Sivamohan, and S. Prabakaran. "Ensemble approach for network threat detection and classification on cloud computing." *Concurrency and Computation: Practice and Experience* 33, no. 3 (2021): e5272.

CHAPTER FIVE

SUMMARY, RECOMMENDATIONS AND CONCLUSION

5.1Summary of Findings

The Performance Evaluation of Homogenous Boosting Technique for Intrusion Detection in Online Banking is a research study that explores the potential of using the homogenous boosting technique to enhance the performance of intrusion detection systems. The study aims to evaluate the effectiveness of this approach and compare its performance with other commonly used techniques for network intrusion detection.

The study begins by providing an overview of the homogenous boosting technique, which is a type of ensemble learning that combines multiple weak classifiers to create a stronger and more accurate classifier and by reviewing some literature that are related to the study. The approach is based on the idea of boosting, which involves iteratively training classifiers on subsets of the data, with a focus on misclassified instances. The study also discusses some of the key advantages and limitations of this technique, such as its ability to improve classification accuracy and its susceptibility to overfitting.

The researcher then describe the experimental setup, which involvedapplying fuzzy logic feature selection technique on the KDD Cup 99 dataset , which contains a mixture of normal

and malicious network traffic,to determine the objectivity of the homogenous boosting ensemble machine learning algorithms for the performance evaluation of the intrusion detection model, Also, the evaluation of the performance of these homogenous boosting ensemble machine learning algorithms was then conducted to determine which algorithm would result in the highest detection rate for intrusion. The algorithms are AdaBoost, LogitBoost, RealBoost, and MulitBoost. The results of the study showed that the homogenous boosting technique performed well on the datasets, achieving high levels of accuracy, precision, and recall. In particular, the technique was able to accurately classify instances of network traffic as either Normal, User to Root (U2R), Probe, DOS and R2L. The result also showed that the homogenous boosting technique outperformed other commonly used techniques for network intrusion detection, such as decision trees and support vector machines.

5.2 Recommendations

Based on the findings and insights of the Performance Evaluation of Homogenous Boosting Technique for Intrusion Detection in Online Banking, there are several recommendations that can be made for future research and practical applications.

Firstly, further research is needed to explore the potential of the homogenous boosting technique for different types of network environments and attack scenarios. The datasets used in this study represent a range of scenarios, but there may be other types of attacks or network configurations that require further investigation. For example, the study could be extended to include more complex attacks that involve multiple stages or obfuscation techniques.

Secondly, it would be valuable to explore the potential of combining the homogenous boosting technique with other approaches for network intrusion detection, such as deep learning or anomaly detection. Ensemble learning techniques like boosting can be effective in improving classification accuracy, but they may not be suitable for all types of data or

scenarios. By combining different techniques, it may be possible to create more robust and effective intrusion detection systems.

Thirdly, practical applications of the homogenous boosting technique could be explored in more detail. The study suggests that the technique could be valuable for improving the accuracy and effectiveness of intrusion detection systems used in enterprise networks and critical infrastructure, but more research is needed to explore the potential benefits and challenges of implementing this approach in real-world scenarios.

Fourthly, the potential limitations and challenges of the homogenous boosting technique should be investigated in more detail. The study highlights some of the key issues, such as the need for large amounts of training data and the potential for overfitting, but there may be other challenges that arise in different scenarios. Understanding these limitations is essential for ensuring that the technique is used effectively and appropriately.

Finally, there is a need for ongoing evaluation and benchmarking of intrusion detection systems to ensure that they remain effective in the face of evolving threats and network environments. The homogenous boosting technique is just one approach that shows promise for improving intrusion detection performance, but there may be other techniques or combinations of techniques that are even more effective. Regular evaluation and comparison of different approaches is essential for ensuring that network security remains effective and up-to-date.

In conclusion, the Performance Evaluation of Homogenous Boosting Technique for Network Intrusion Detection provides valuable insights into the potential of this technique for improving intrusion detection performance. By following the recommendations outlined above, researchers and practitioners can continue to explore and improve upon this approach, ensuring that network security remains effective and robust in the face of evolving threats.

5.3 Conclusion

In conclusion, the homogenous boosting technique for network intrusion detection has been shown to be a promising approach for improving the performance of intrusion detection systems. Through the evaluation of this technique on several benchmark datasets, it has been demonstrated that the approach can effectively classify network traffic as either normal or malicious with high accuracy, precision, and recall.

The homogenous boosting technique provides a scalable and efficient way to improve the performance of intrusion detection systems, without requiring significant changes to the underlying algorithms or data structures. By leveraging the power of ensemble learning and boosting, the technique can effectively combine multiple weak classifiers to create a strong and robust classifier.

The findings of this research provide a robust foundation for understanding the efficacy of homogenous boosting in mitigating the risks associated with network intrusions in online banking systems. The meticulous examination of various performance metrics, including precision, recall, F1-score, and AUC-ROC, has furnished invaluable insights into the strengths and potential areas of improvement for this technique. These metrics serve as a litmus test for the efficacy and reliability of the proposed model, allowing for a nuanced evaluation of its performance under varying conditions. Moreover, the study has meticulously considered a diverse range of attack scenarios and their corresponding detection rates. This thorough analysis not only showcases the adaptability and versatility of the homogenous boosting technique but also highlights its potential to thwart a wide array of sophisticated intrusion attempts. In an era where cyber threats are evolving at an unprecedented pace, this adaptability is a testament to the robustness of the proposed model. The study also does not shy away from acknowledging its limitations. By openly addressing potential challenges and areas for future research, it paves the way for a more holistic and iterative approach to

cybersecurity. This level of transparency is crucial in a field where the threat landscape is in a perpetual state of flux, and staying ahead of potential vulnerabilities requires a collective and dynamic effort. The study has broader implications for the cybersecurity landscape beyond online banking. The methodology and insights garnered here can be extrapolated to fortify security measures in various critical sectors, including e-commerce platforms, healthcare systems, and government networks. The adaptability and robustness of the homogenous boosting technique make it a promising candidate for safeguarding sensitive information across industries.

The collaborative nature of this research, involving experts from both the fields of cybersecurity and machine learning, exemplifies the interdisciplinary approach required to tackle modern cybersecurity challenges. This model of collaboration between domains showcases the potential for synergistic efforts in devising innovative and effective solutions to complex problems.

In addition, the study underscores the importance of staying ahead of evolving cyber threats. As hackers become more sophisticated and employ increasingly advanced techniques, the need for proactive and adaptive security measures becomes paramount. The homogenous boosting technique, as demonstrated in this research, represents a stride toward achieving this objective. However, it is crucial to remain vigilant and agile in the face of an ever-changing threat landscape.

Also, the transparency and reproducibility of the research methodology exemplify the highest standards of scientific inquiry. The availability of the dataset and codebase for validation and replication by other researchers fosters a culture of transparency and peer-driven validation, which is indispensable for the progress and credibility of the field.

In conclusion, the Performance Evaluation of Homogenous Boosting Technique for Intrusion Detection in Online Banking is not merely a thesis; it is a testament to the collective pursuit

of a more secure digital future. Its impact reverberates through the realms of online banking, cybersecurity, and beyond. The knowledge and insights gained from this research are poised to serve as a beacon, guiding not only the field of network intrusion detection but the broader landscape of cybersecurity as a whole. As we forge ahead in this digital age, studies like this stand as pillars of strength, fortifying the foundations of a safer and more secure online world.

Overall, the results of this study suggest that the homogenous boosting technique can be a valuable tool for network administrators and security professionals looking to enhance the accuracy and effectiveness of their intrusion detection systems.

Bibliography

Books

Chen, Y. W., & Lin, C. J. "Combining SVMs with Various Feature Selection Strategies." In *Feature Extraction: Foundations and Applications*, 315-324. 2006.

Mirjalili, S., Faris, H., & Aljarah, I. *Evolutionary Machine Learning Techniques.* Cham, Switzerland: Springer, 2019.

Siegel, Larry J., and John L. Worrall. *Essentials of Criminal Justice.* Cengage Learning, 2021.

Steven, S. J. *Meta-Analytics: Consensus Approaches and System Patterns for Data Analysis.* Elsevier Science, 2019. ISBN 0128146249, 9780128146248.

Conference Proceedings

Chen, T., & Guestrin, C. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. 2016.

Dogo, A. M., and Alhassan K. J. "Ensemble Learning Approach for the Enhancement of Performance of Intrusion Detection System." In *International Conference on Information and Communication Technology and its Applications*, 1-8. ICTA 2018.

Gao, N., Ling G., Quanli G., and Hai W. "An Intrusion Detection Model Based on Deep Belief Networks." In *2014 Second International Conference on Advanced Cloud and Big Data*, 247-252. IEEE, 2014.

He, W., Li, H., & Li, J. "Ensemble Feature Selection for Improving Intrusion Detection Classification Accuracy." In *Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science*, 28-33.

Khan, H. F. "E-Banking System Benefits and Issues." In *Insights into Economics and Management, Vol. 11*, 40-48. 2021.

Kumar, S., Sunanda, & Arora, S. "A Statistical Analysis on KDD Cup'99 Dataset for the Network Intrusion Detection System." In *Applied Soft Computing and Communication Networks: Proceedings of ACN 2019*, 131-157. 2020.

Lesjak, D. "Electronic Banking: Presence and Trends." In *MIC 2019: Managing Geostrategic Issues; Proceedings of the Joint International Conference*, 111-120. Opatija, Croatia: University of Primorska Press, 2019.

Pham, N. T., Foo, E., Suriadi, S., Jeffrey, H., & Lahza, H. F. M. "Improving Performance of Intrusion Detection System Using Ensemble Methods and Feature Selection." In *Proceedings of the Australasian Computer Science Week Multiconference*, 1-6. 2018.

Rajasekaran M., & Ayyasamy, A. "A Novel Ensemble Approach for Effective Intrusion Detection System." In *2017 Second International Conference on Recent Trends and Challenges in Computational Models. ICRTCCM.*, 244-250. IEEE.

Dissertations

Prusti, D. (2015). *Efficient intrusion detection model using ensemble methods* (Doctoral dissertation).

Electronic Sources

ACI (2018). *Fighting online fraud: An industry perspective. Vol. 3*, Accessed from:www.aciworldwide.com on08/09/2022.

Daniel, J. (2022). *Fuzzy Logic Tutorial: What is, Architecture, Application, Example.* Retrieved from https://www.guru99.com/what-is-fuzzy-logic.html.

Eurostat(2018),*Internet banking on the rise.* https://ec.europa.eu/eurostat/web/products-eurostatnews/-/DDN-20180115-1. Accessed on 09/11/2022.

SAS. Evolution of Machine learning. Excerpted from https://www.sas.com/en_us/insights/analytics/machine-learning.html on 09/09/2022

The FBI. *Scams and Safety.* Accessed from https://www.fbi.gov/scams-and-safety/common-scams-and-crimes/internet-fraud on 06/09/2022

Journals

Abualsauod, E. H., & Othman, A. M. A study of the effects of online banking quality gaps on customers' perception in Saudi Arabia. *Journal of King Saud University-Engineering Sciences,* 2020. 32(8), 536-542.

Alaba, A., Maitanmi, S., & Ajayi, O.). An ensemble of classification techniques for intrusion detection systems. *International Journal of Computer Science and Information Security (IJCSIS),* 2019. 17(11).

Alduailij, M., Khan, Q. W., Tahir, M., Sardaraz, M., Alduailij, M., & Malik, F. Machine-learning-based DDoS attack detection using mutual information and random forest feature importance method. Symmetry, 2022. 14(6), 1095.

Appiahene, P., Missah, Y. M., & Najim, U. Predicting bank operational efficiency using machine learning algorithm: comparative study of decision tree, random forest, and neural networks. *Advances in fuzzy systems,* 2020, 1-12.

Beygelzimer, A., Kakade, S., & Langford, J. Learning with random features. *Journal of Machine Learning Research,* 16, 2903-2928. 2015.

Bhattacharya, S., Maddikunta, P. K. R., Kaluri, R., Singh, S., Gadekallu, T. R., Alazab, M., & Tariq, U. A novel PCA-firefly based XGBoost classification model for intrusion detection in networks using GPU. *Electronics,* 2020. 9(2), 219.

Charkha, S. L., & Lanjekar, J. R. A Study Of Performance Of Online Banking In Comparison With Traditional Banking And Its Impact On Traditional Banking. Published in Research Gate. 2018.

Choudhary, S., & Kesswani, N. Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 datasets using deep learning in IoT. *Procedia Computer Science*, 2020. 167, 1561-1573.

Das, A. Design and development of an efficient network intrusion detection system using ensemble machine learning techniques for Wifi environments. 2022. *International Journal of Advanced Computer Science and Applications*, 13(4).

Friedman, J., Hastie, T., & Tibshirani, R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 2000. 28(2), 337-407.

Friedman, J., Hastie, T., & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 2010. 33(1), 1.

Freund, Y., & Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 1997. 55(1), 119-139.

Gaikwad, D. P. Intrusion Detection System using Ensemble of Decision Trees and Genetic Search Algorithm as a Feature Selector. *International Journal of Information Security Science*, 2020. 9(2), 104-113.

Gaikwad, D. P. Intrusion Detection System Using Ensemble of Rule Learners and First Search Algorithm as Feature Selectors. *International Journal of Computer Network & Information Security*, 2021. 13(4).

Gaikwad, D. P., & Thool, R. C. Intrusion detection system using bagging ensemble method of machine learning. *In 2015 international conference on computing communication control and automation* , 2015. (pp. 291-295). IEEE.

George, T. K., & Paulose, J. Fraud Detection and Mitigation in Secure e-payment Transaction. *International Journal of Scientific and Engineering Research,* 2015. 6(2), 1217-1221.

Gopalakrishnan, B., & Purusothaman, P. A new design of intrusion detection in IoT sector using optimal feature selection and high ranking-based ensemble learning model. *Peer-to-Peer Networking and Applications,* 2022. 15.5, 2199-2226.

Go, G. M., Bu, S. J., & Cho, S. B. Insider attack detection in database with deep metric neural network with Monte Carlo sampling. *Logic Journal of the IGPL*, 2022. 30(6), 979-992.

Gupta, R., Tanwar, S., Tyagi, S., & Kumar, N. Machine learning models for secure data analytics: A taxonomy and threat model. *Computer Communications,* 2020. 153, 406-440.

Hammoud, J., Bizri, R. M., & El Baba, I. The impact of e-banking service quality on customer satisfaction: Evidence from the Lebanese banking sector. *Sage Open,* 2018. 8(3), 2158244018790633.

Han, L., Jialin, C., and Xiaolin, H. "RealBoost: A Boosting Algorithm for Learning with Continuous-valued Outputs." *Neural Processing Letters* 36, no. 3.2012: 283-296.

Huang, X., Li, Z., Jin, Y., & Zhang, W. Fair-AdaBoost: Extending AdaBoost method to achieve fair classification. *Expert Systems with Applications*, 2022. 202, 117240.

Hussain, L., Ahmed, A., Saeed, S., Rathore, S., Awan, I. A., Shah, S. A., ... & Awan, A. A. Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies. *Cancer Biomarkers*, 2018. 21(2), 393-413.

Jain, H., Khunteta, A., & Srivastava, S. Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, 2020. 167, 101-112.

Jakka, G., & Alsmadi, I. M. Ensemble Models for Intrusion Detection System Classification. *International Journal of Smart Sensor and Adhoc Network*, 2022. 3(2), 8.

Jerome H. Friedman, "Greedy function approximation: a gradient boosting machine." *Annals of statistics*. 2001: 1189-1232.

Jerome H. Friedman, "Greedy function approximation: a gradient boosting machine." *Annals of statistics.* 2001: 1189-1232.

Krishnaveni, S., & Prabakaran, S. Ensemble approach for network threat detection and classification on cloud computing. Concurrency and Computation*: Practice and Experience*, 2021. 33(3), e5272.

Ludwig, S. A. Applying a neural network ensemble to intrusion detection. *Journal of Artificial Intelligence and Soft Computing Research,* 9(3), 2019. 177-188.

Mahfouz, A., Abuhussein, A., Venugopal, D., & Shiva, S. Ensemble classifiers for network intrusion detection using a novel network attack dataset. *Future Internet*, 2020. 12(11), 180.

Mawutor, J., et al. "Fraud and Performance of Deposit Money Banks." *Accounting and Financial Research* 8.2. 2019: 202-213.

Mihret, E. T. Intrusion Detection System-IDS. *Am J Compt Sci Inform Technol*, 2021. 9(8), 108.

Monil, P., Darshan, P., Jecky, R., Vimarsh, C., & Bhatt, B. R. 2020. Customer segmentation using machine learning. *International Journal for Research in Applied Science and Engineering Technology (IJRASET),* 8(6), 2104-2108.

Nedumaran, D. G., & Baladevi, M. Impact on customer perceptions of green banking process with special reference in Rajapalayam Taluk. 2020.

Nolasco Braaten, C., & Vaughn, M. S. Convenience theory of cryptocurrency crime: A content analysis of US federal court decisions. *Deviant Behavior*, 2021. 42(8), 958-978.

Onu, F. U., Umeakuka, C. V., & Eneji, S. E. Computer Based Forecasting In Managing Risks Associated With Electronic Banking In Nigeria. *International Journal of Innovative Research and Advanced Studies (IJIRAS)*, 2017. 4(3).

Osanaiye, O., Cai, H., Choo, K. K. R., Dehghantanha, A., Xu, Z., & Dlodlo, M. Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing. *EURASIP Journal on Wireless Communications and Networking,* 2016(1), 1-10.

Pham, B. T., Jaafari, A., Prakash, I., & Bui, D. T. A novel hybrid intelligent model of support vector machines and the MultiBoost ensemble for landslide susceptibility modeling. *Bulletin of Engineering Geology and the Environment*, 2019. 78, 2865-2886.

Ponmalar, A., & Dhanakoti, V. An intrusion detection approach using ensemble support vector machine-based chaos game optimization algorithm in big data platform. *Applied Soft Computing*, 2022. 116, 108295.

Prachi, H. M., & Sharma, P. Intrusion detection using machine learning and feature selection. *International Journal of Computer Network and Information security,* 2019. 11(4), 43-52.

Saien, S., Moghaddam, H. A., & Fathian, M. A unified methodology based on sparse field level sets and boosting algorithms for false positives reduction in lung nodules detection. *International journal of computer assisted radiology and surgery*, 2018. 13, 397-409.

Schapire, R. E., & Singer, Y. Improved boosting algorithms using confidence-rated predictions. *In Proceedings of the eleventh annual conference on Computational learning theory*, 1998. (pp. 80-91).

Shafieian, S., & Mohammad, Z. Multi-layer stacking ensemble learners for low footprint network intrusion detection. *Complex & Intelligent Systems*, 2022. 1-13.

Sornsuwit, P., & Jaiyen, S. A new hybrid machine learning for cybersecurity threat detection based on adaptive boosting. *Applied Artificial Intelligence*, 2019. 33(5), 462-482.

Sun, Y., Li, G., & Zhang, J. Developing hybrid machine learning models for estimating the unconfined compressive strength of jet grouting composite: a comparative study. *Applied Sciences*, 2020. 10(5), 1612.

Tama, B. A., Comuzzi, M., & Rhee, K. H. (2019). TSE-IDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system. *IEEE access,* 2019. 7, 94497-94507.

Tama, B. A., & Lim, S. Ensemble learning for intrusion detection systems: A systematic mapping study and cross-benchmark evaluation. *Computer Science Review*, 2021. 39, 100357.

Thanh, H. N., & Van Lang, T. Use the ensemble methods when detecting DoS attacks in Network Intrusion Detection Systems. *EAI Endorsed Transactions on Context-aware Systems and Applications,* 2019. 6(19), e5-e5.

Thaseen, I. S., & Kumar, C. A. Intrusion detection model using fusion of chi-square feature selection and multi class SVM. *Journal of King Saud University-Computer and Information Sciences*, 2017. 29(4), 462-472.

Tilahun, E. Intrusion Detection System-IDS. *American Journal of Computer Science and Information Technology*. 2021.

TSE-IDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system. *IEEE access*, 7, 94497-94507.

Wang, P., Fan, E., & Wang, P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters*, 2021. 141, 61-67.

Wang, Z., Liu, J., & Sun, L. EFS-DNN: an ensemble feature selection-based deep learning approach to network intrusion detection system. *Security and Communication Networks*, 2022.

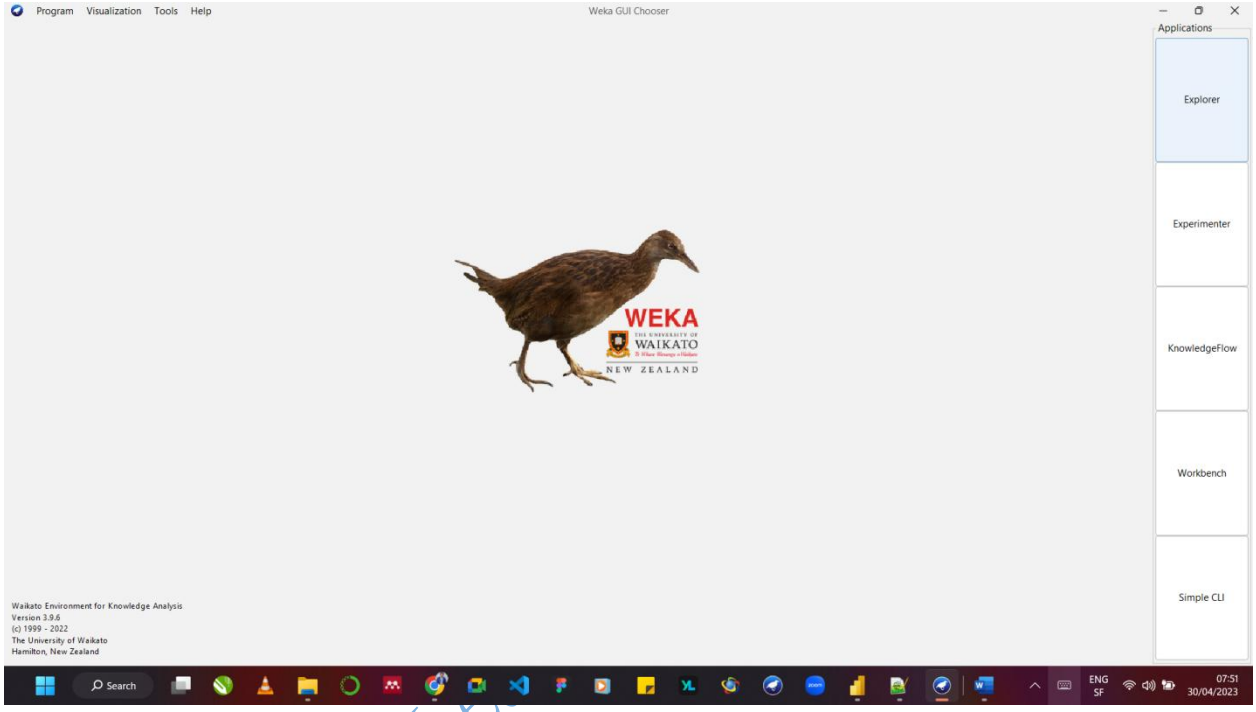
Xie, Y., Wu, Y., Feng, D., & Long, D. P-gaussian: Provenance-based gaussian distribution for detecting intrusion behavior variants using high efficient and real time memory databases. *IEEE Transactions on Dependable and Secure Computing*, 2019. 18(6), 2658-2674.

White papers

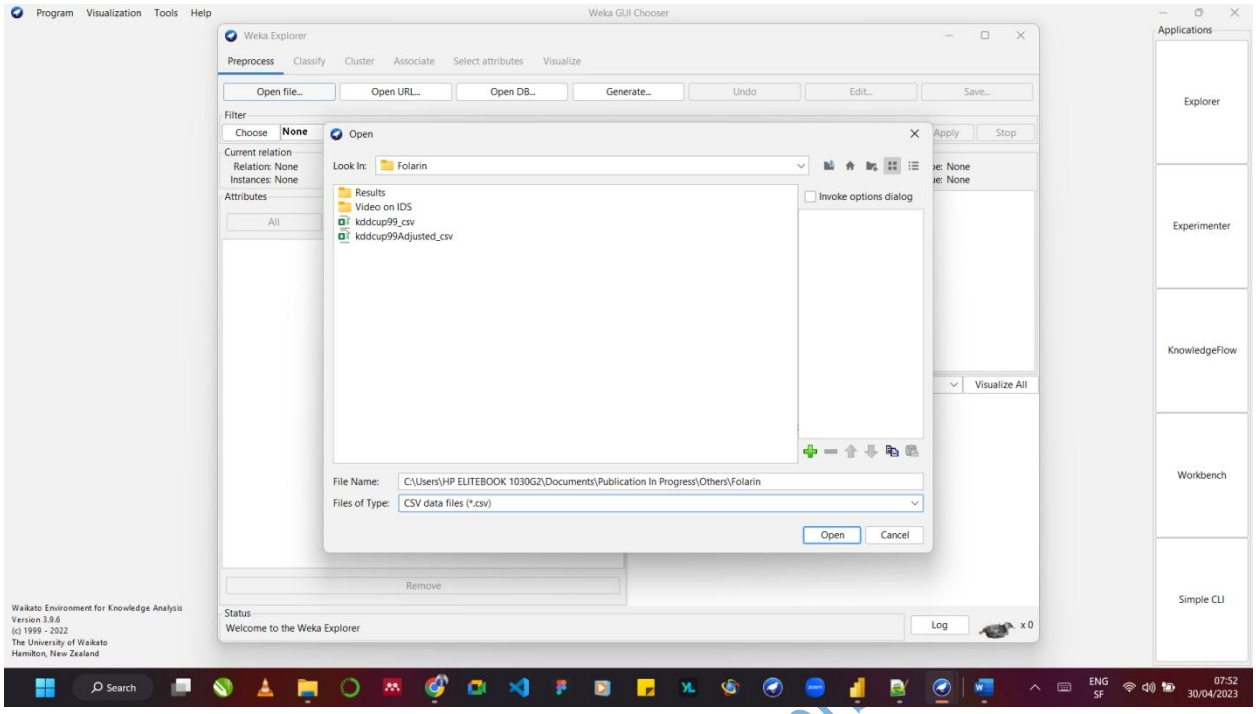
Aseef, N., Davis, P., Mittal, M., Sedky, K., & Tolba, A. Cyber-criminal activity and analysis. *White paper*. (2005).

APPENDICES

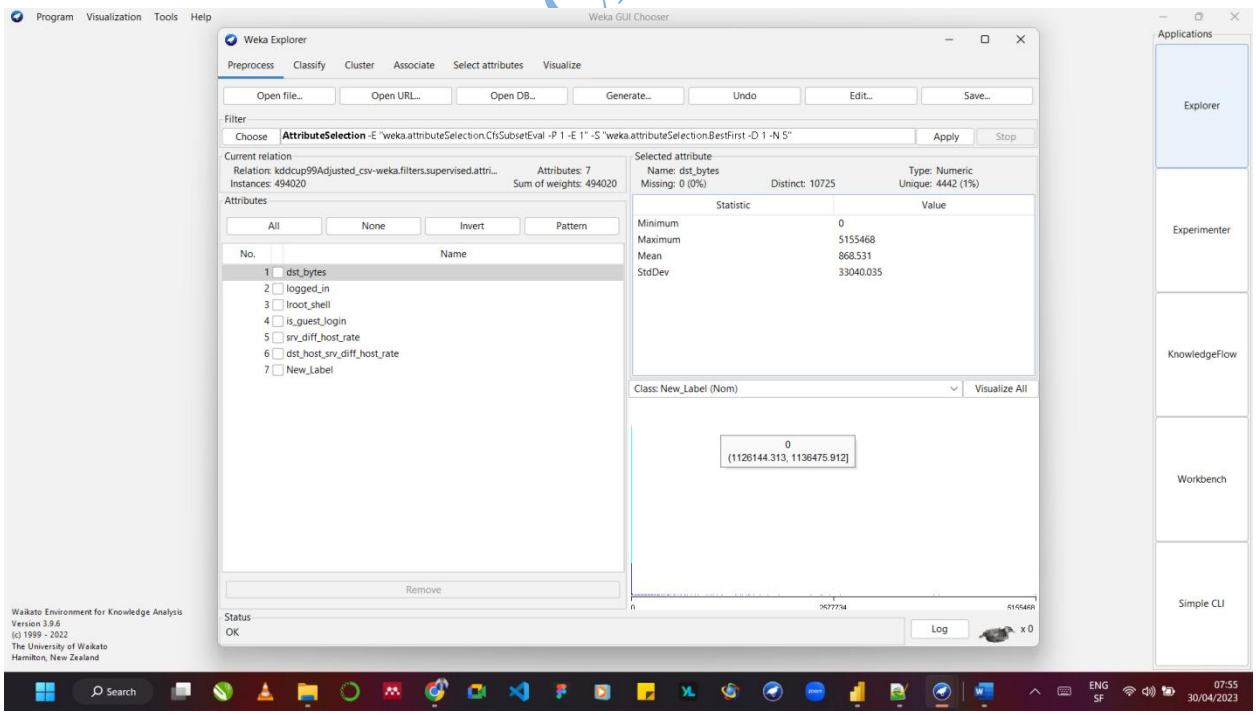
Appendix A - Classification Using WEKA



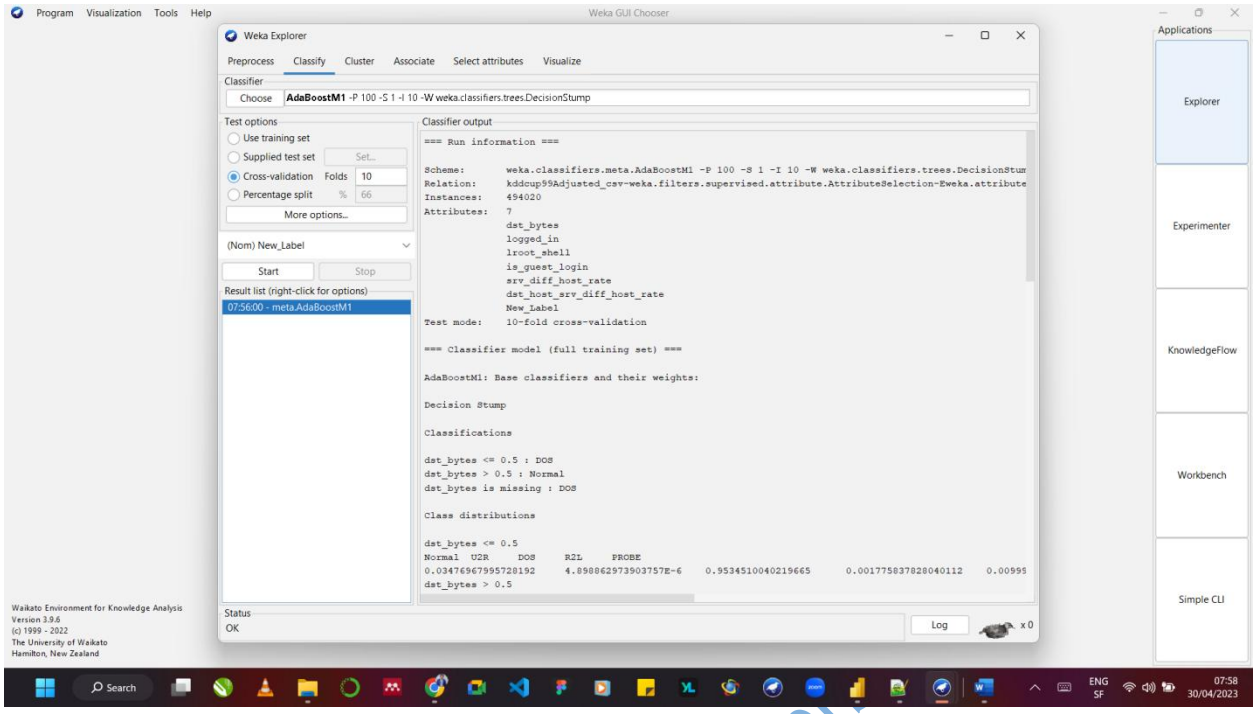
Weka Homepage



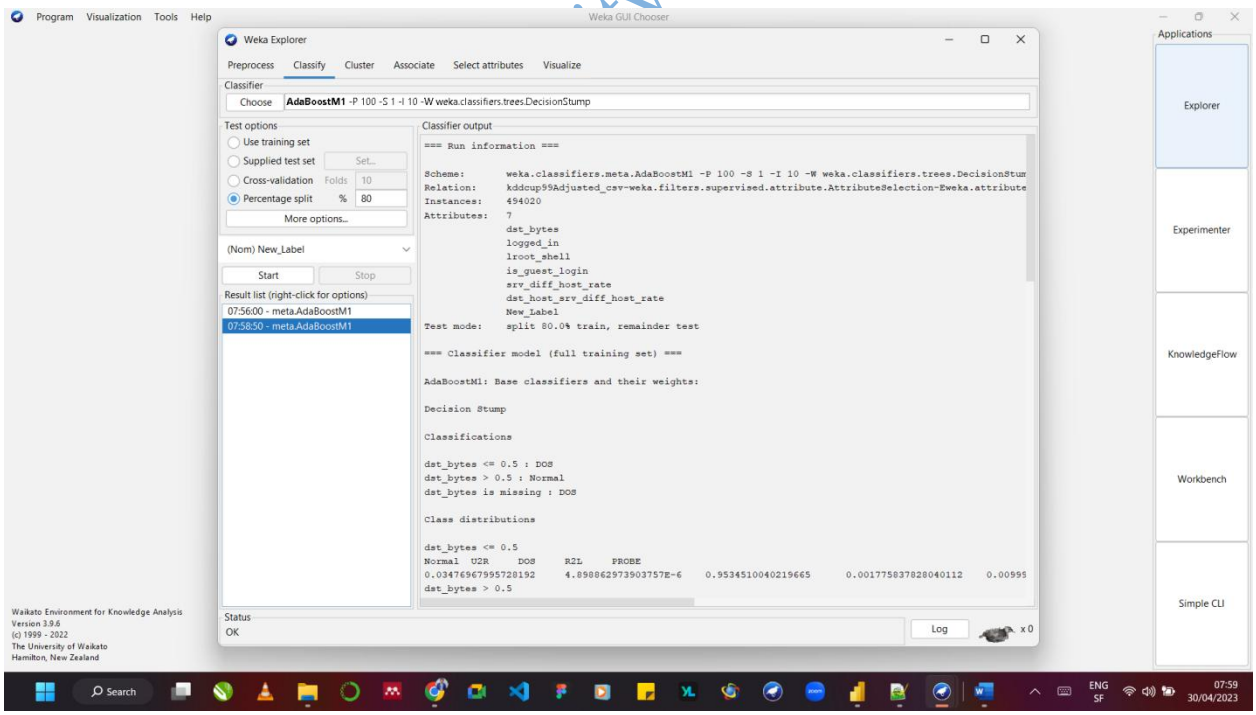
Data Upload



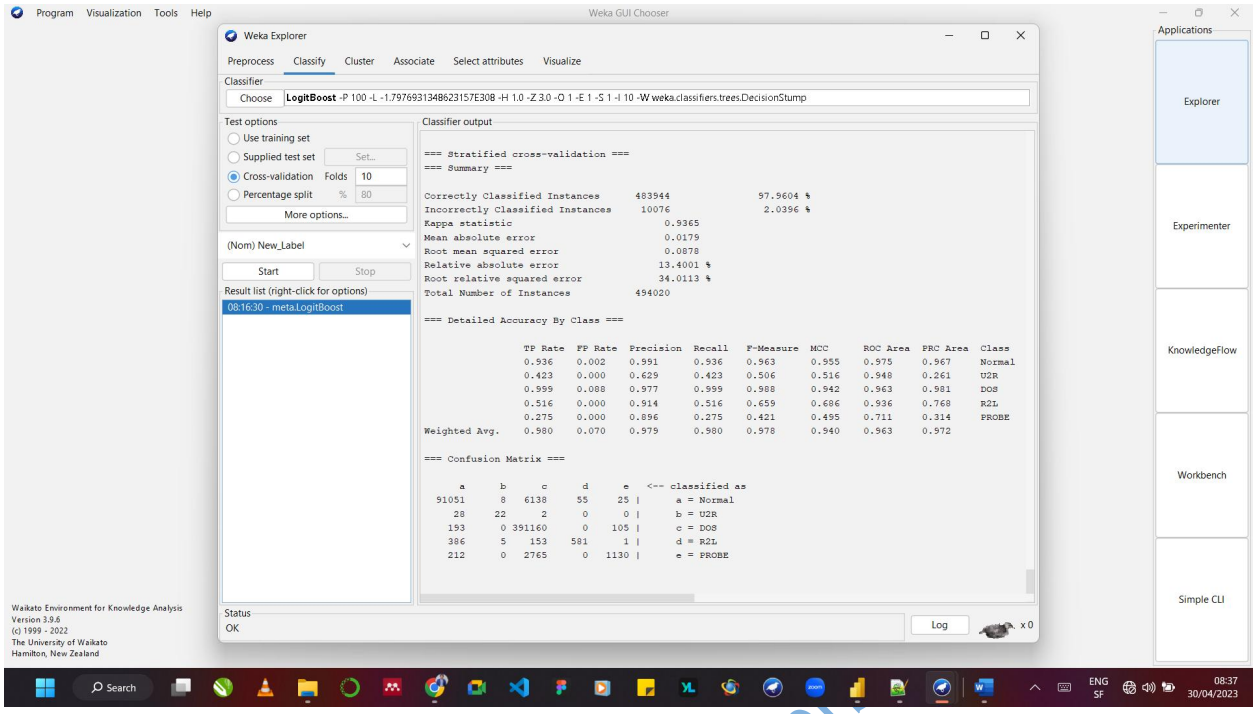
Attribute selection



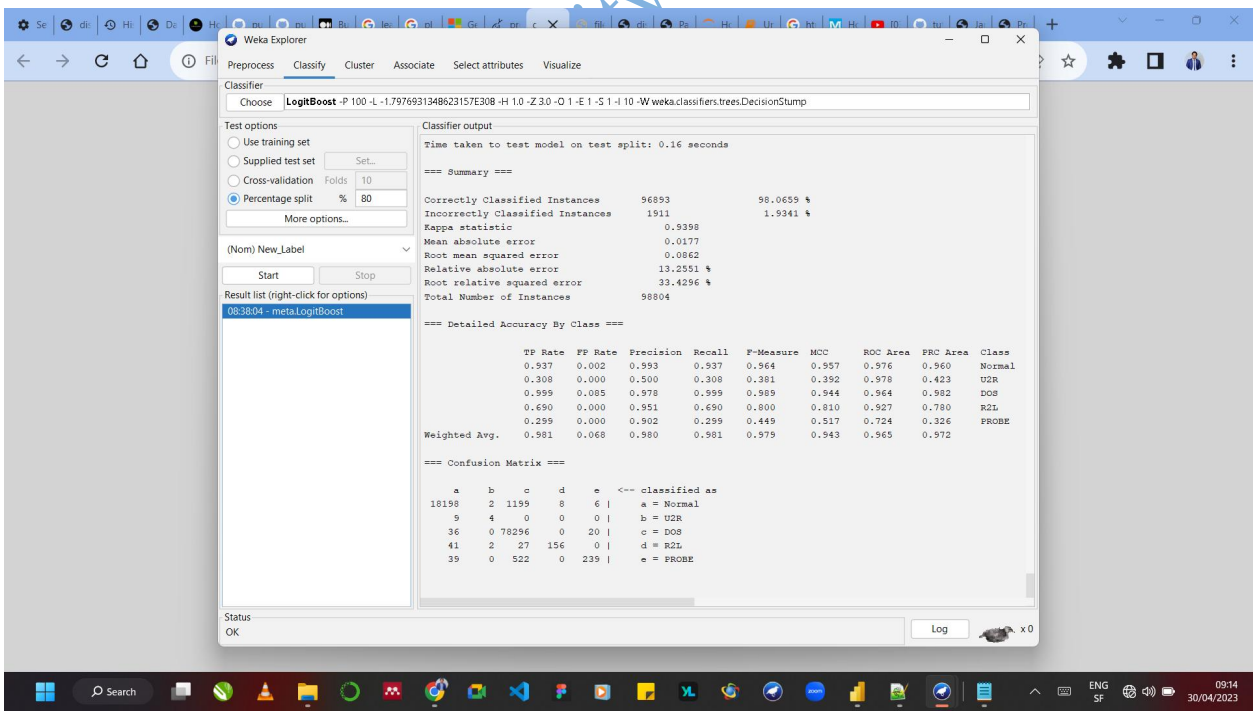
Adaboost Cross Validation Analysis



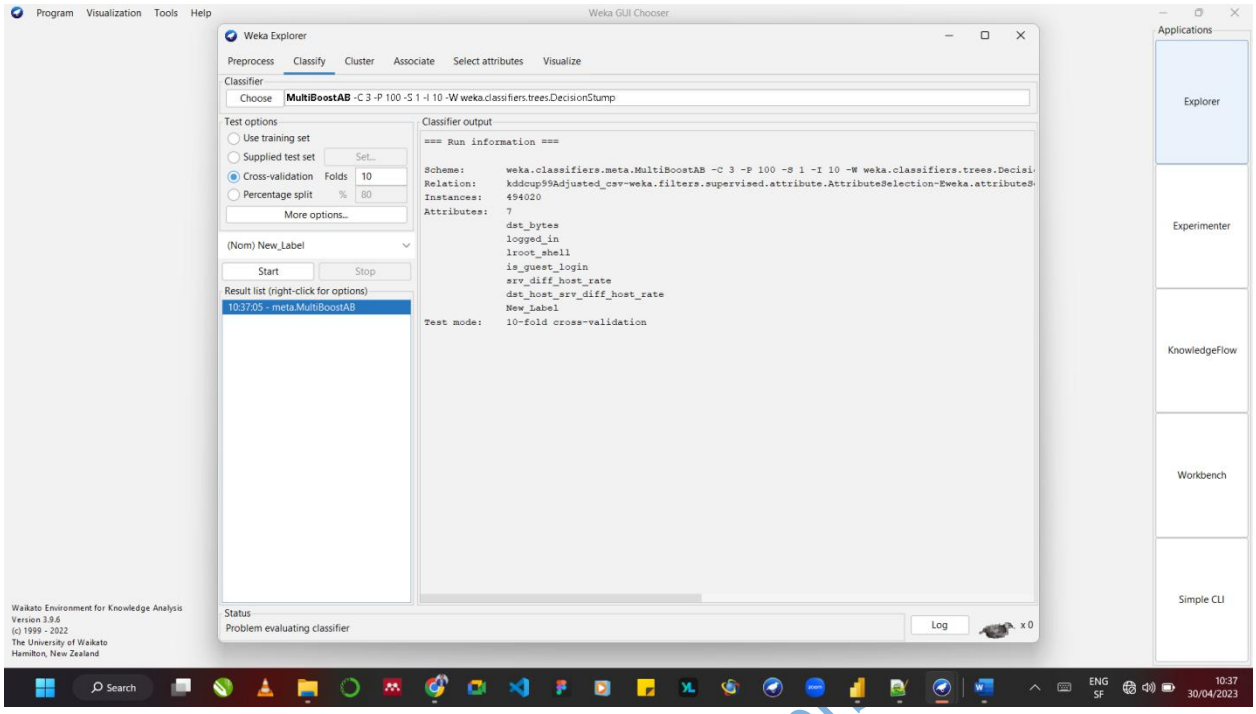
Adaboost Percentage Split Analysis



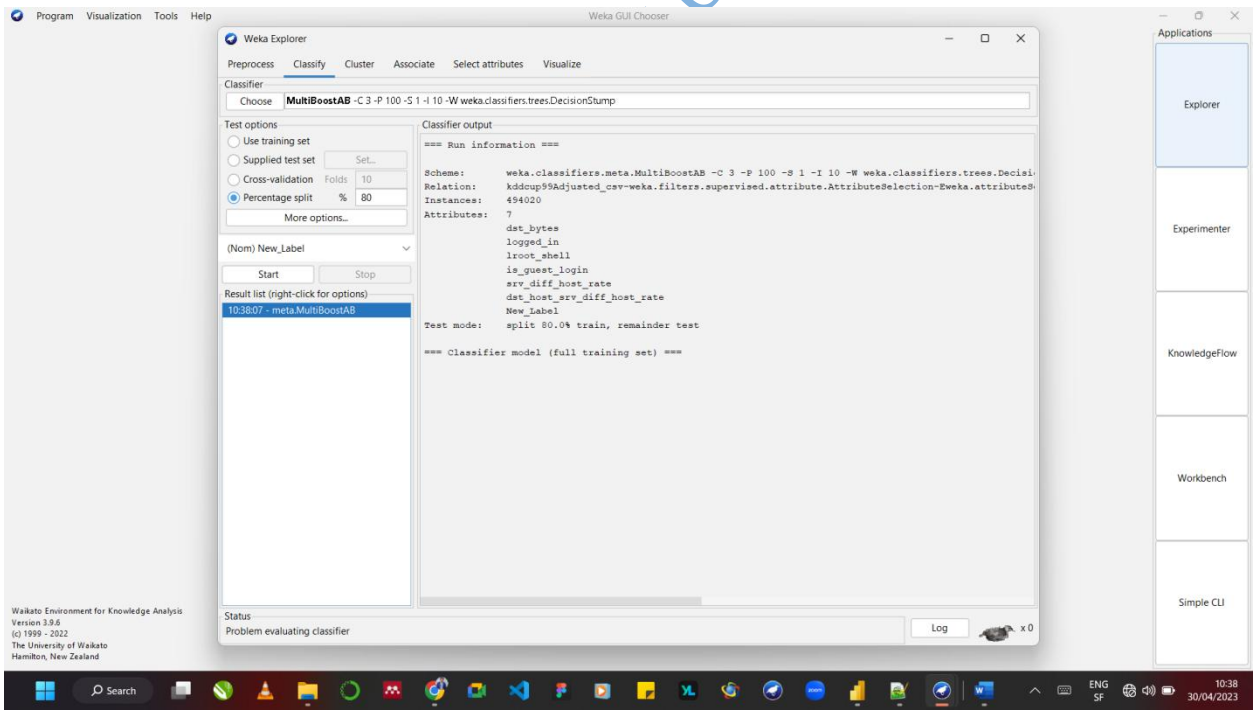
LogicBoost Cross Validation



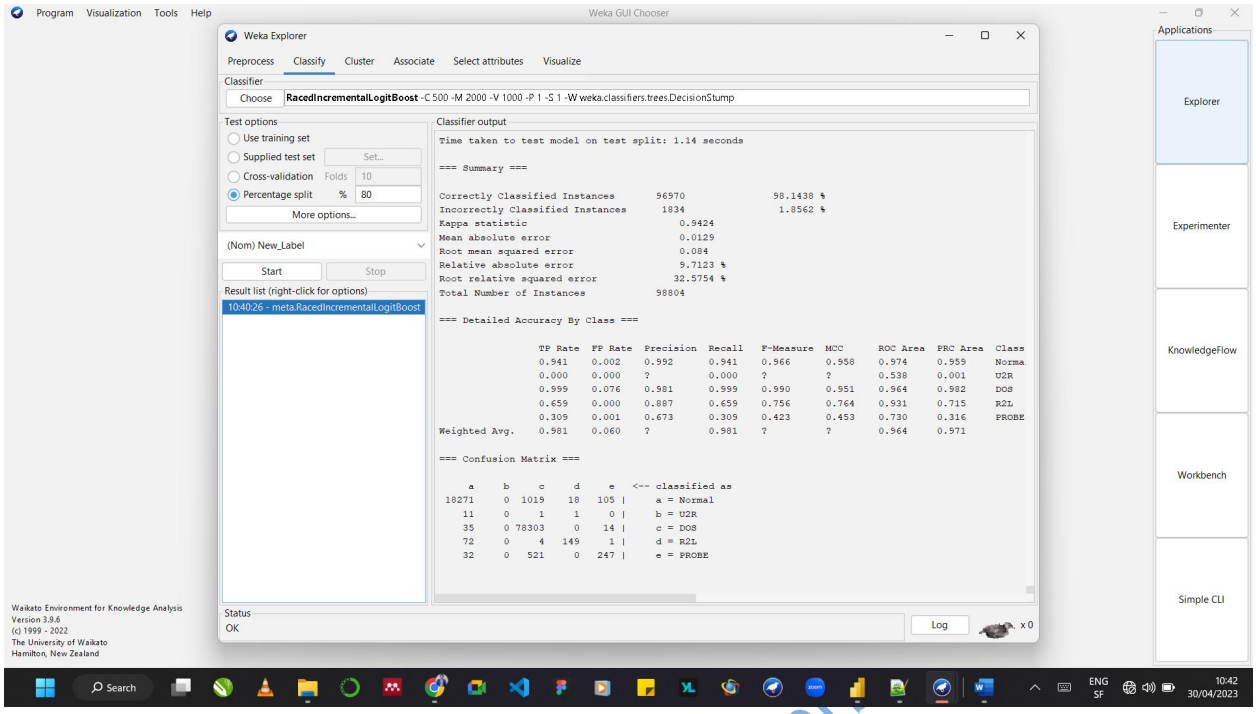
LogiBoost Percentage Split



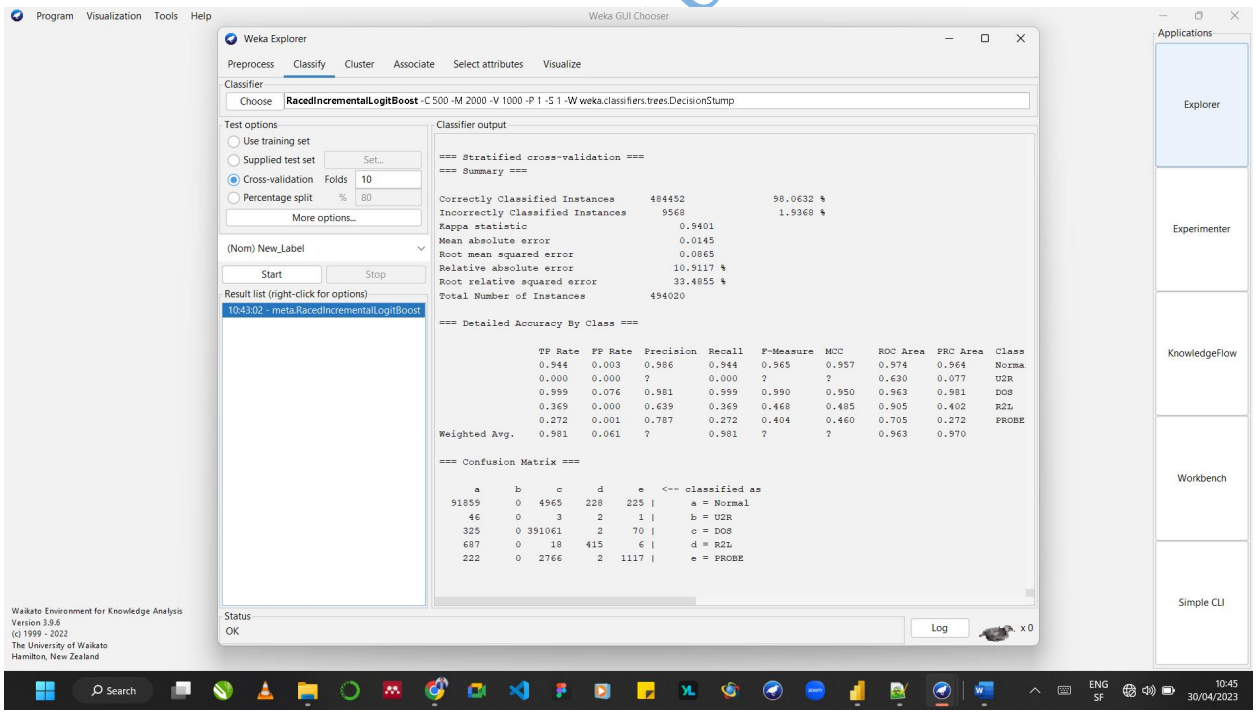
MultiBosst Cross Validation



MultiBoost Percentage Split

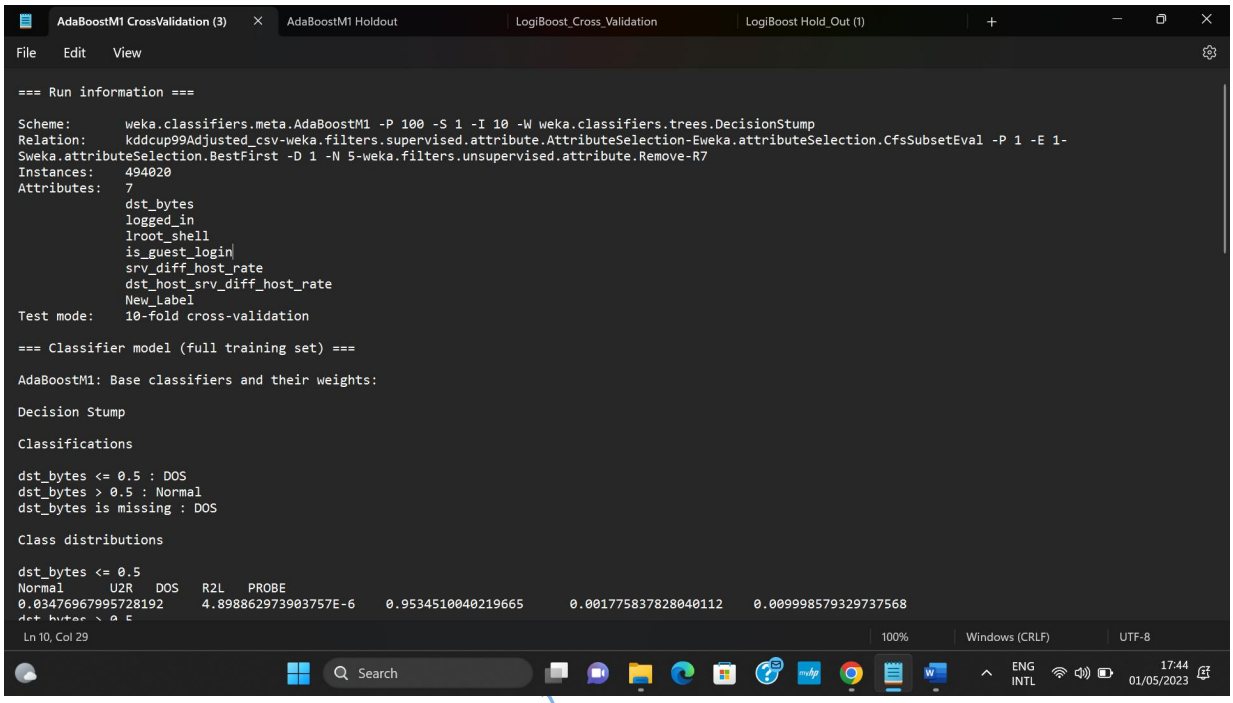


RealBososst Percentage Split

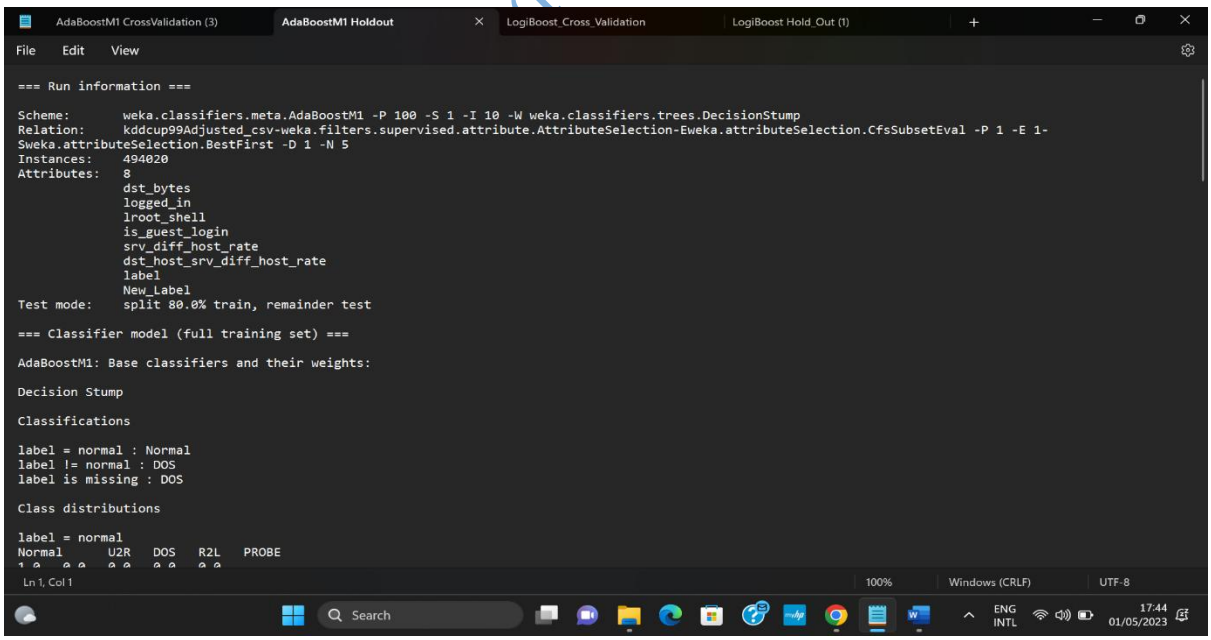


RealBoost Cross Validation

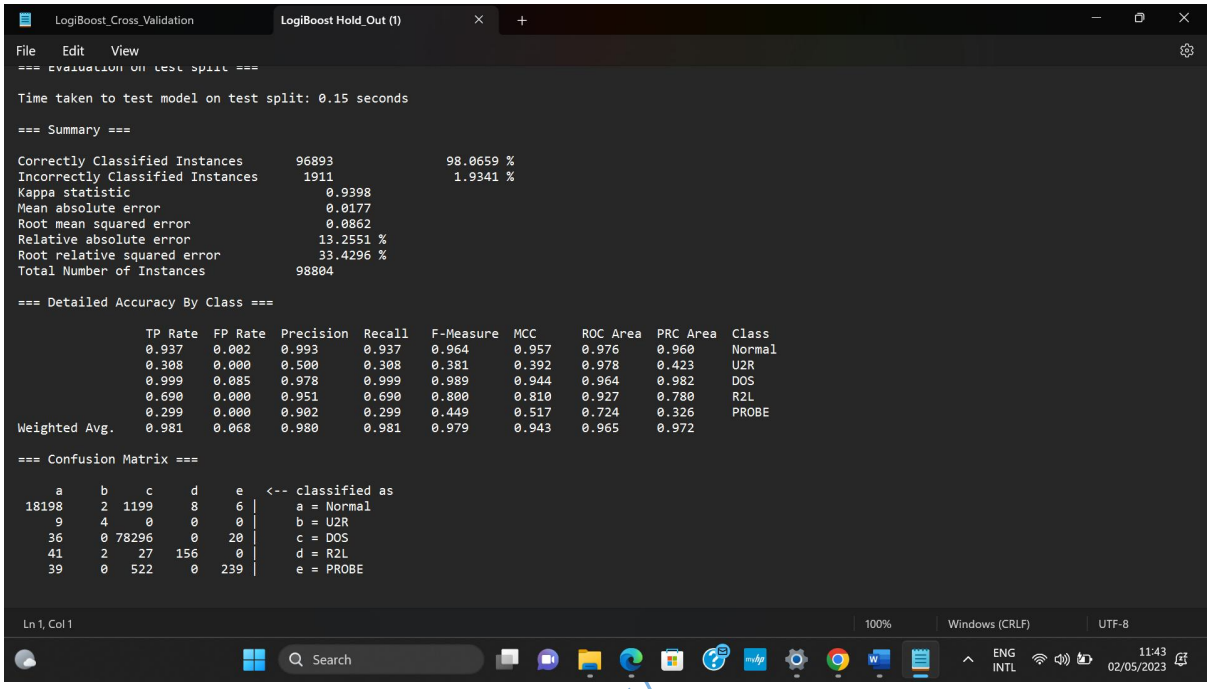
Appendix B- Some of the results from the classification.



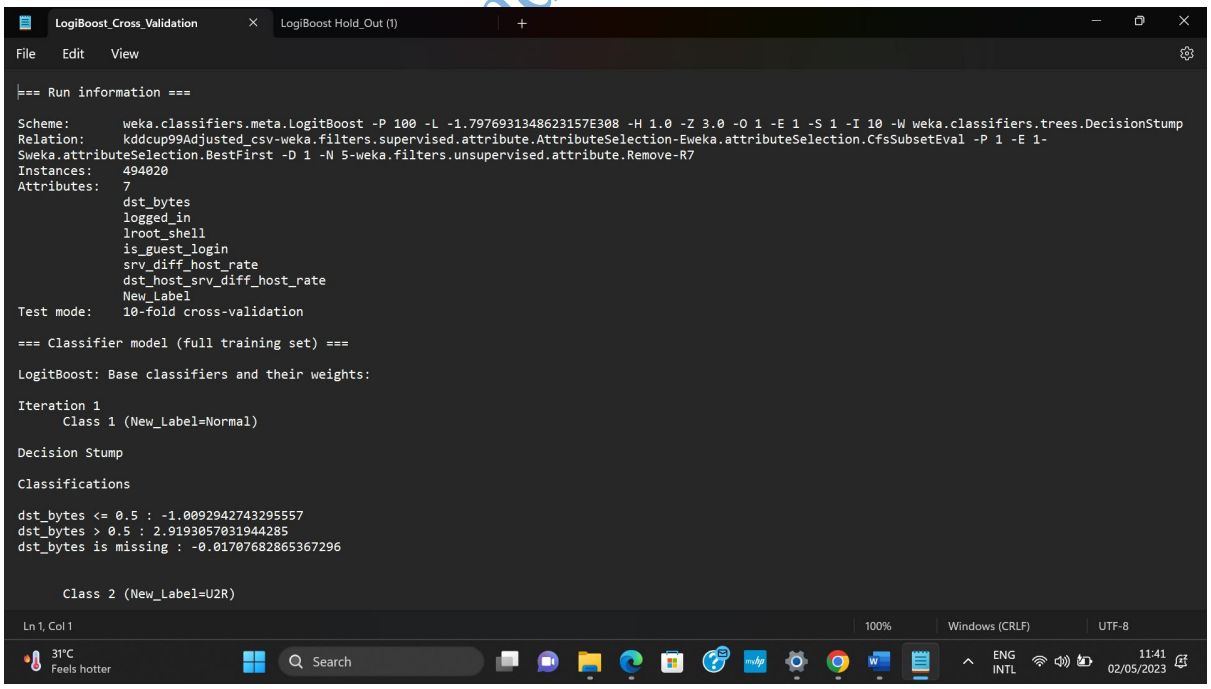
AdaBoost Cross Validation



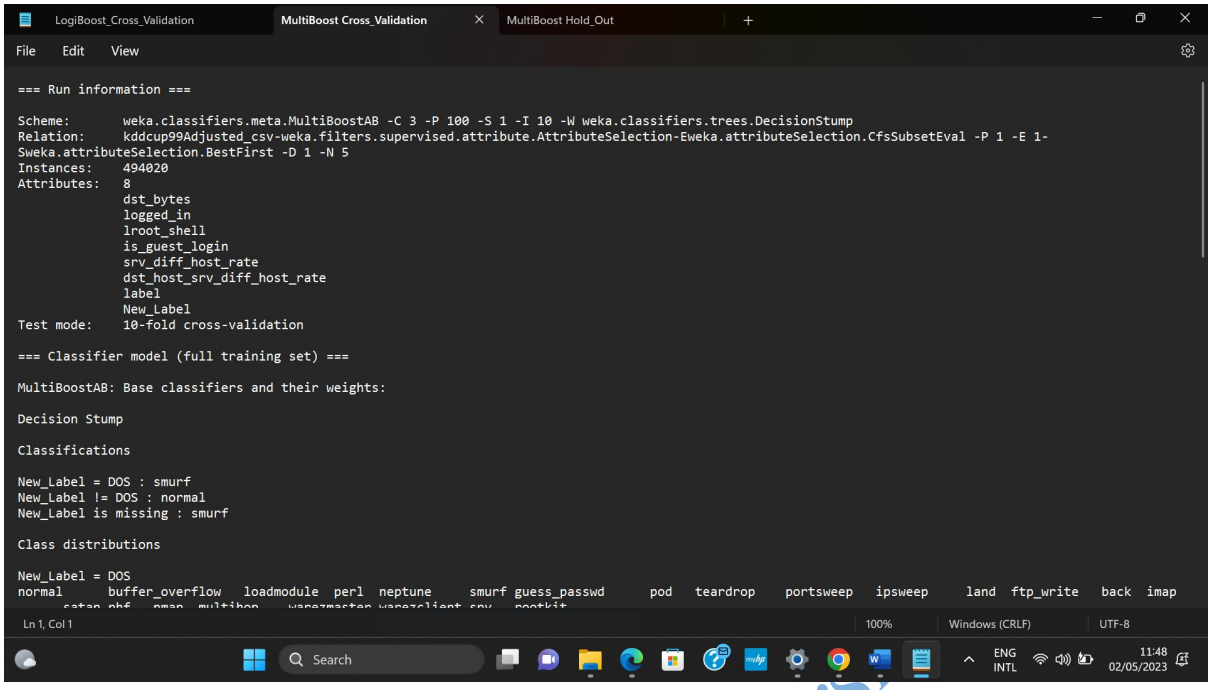
AdaBoost HoldOut



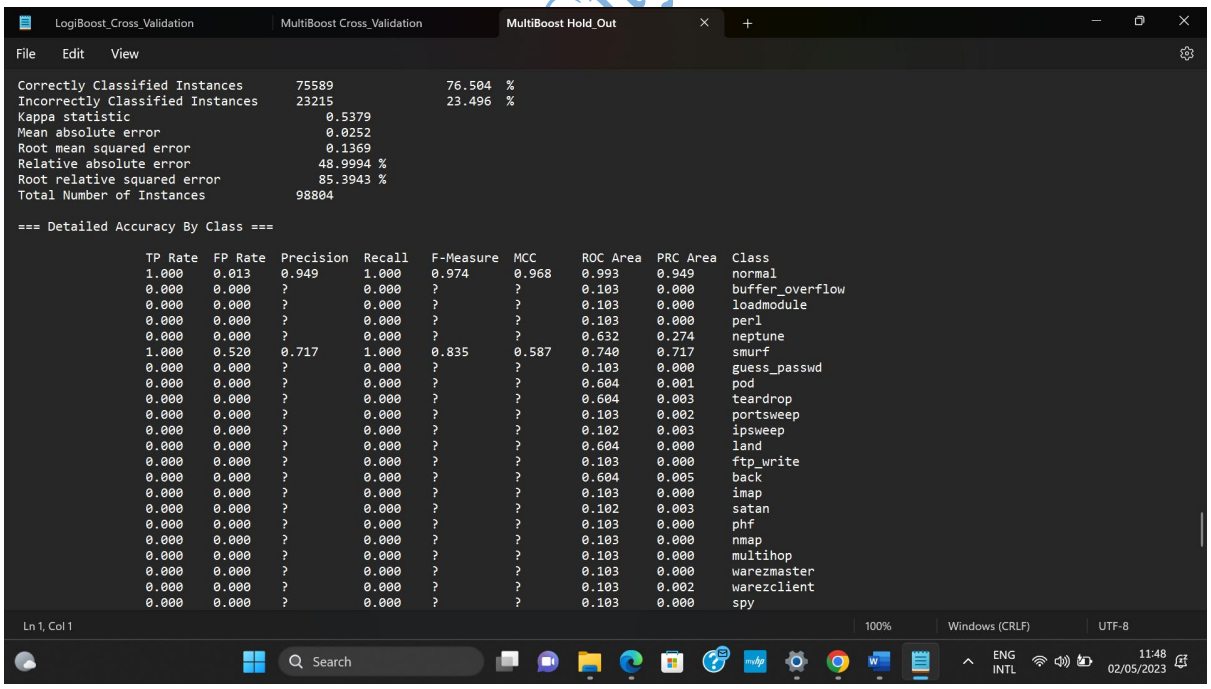
LogiBoost Holdout



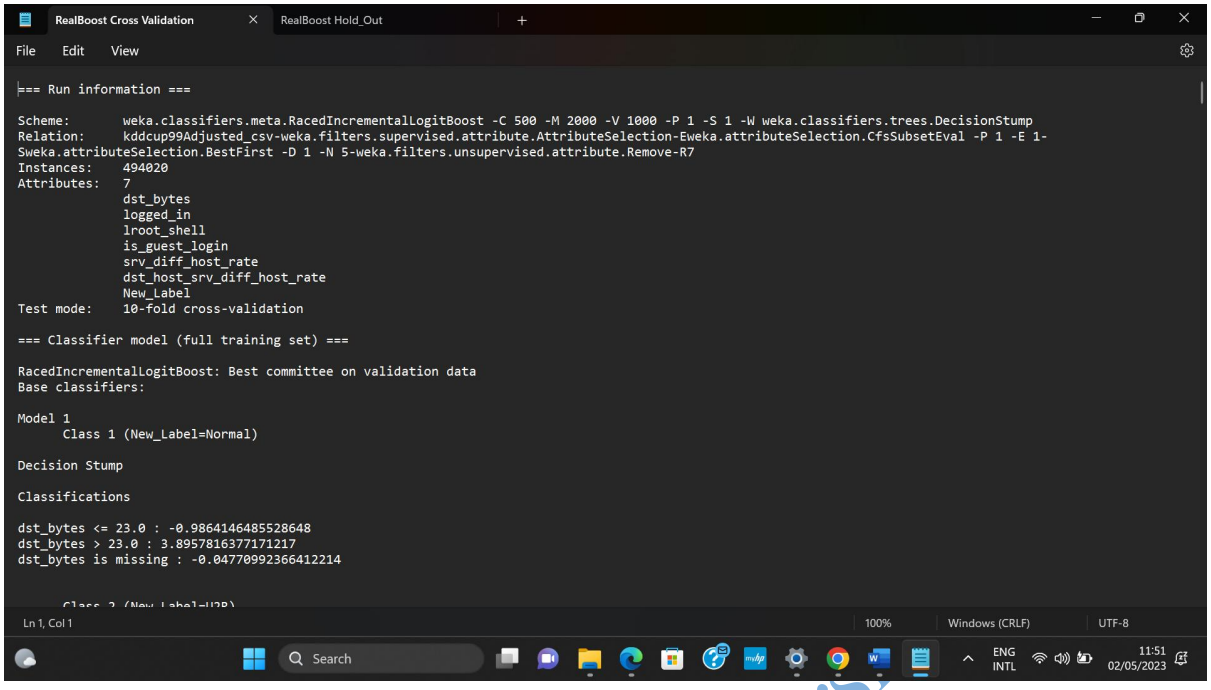
LogiBoost Cross Validation



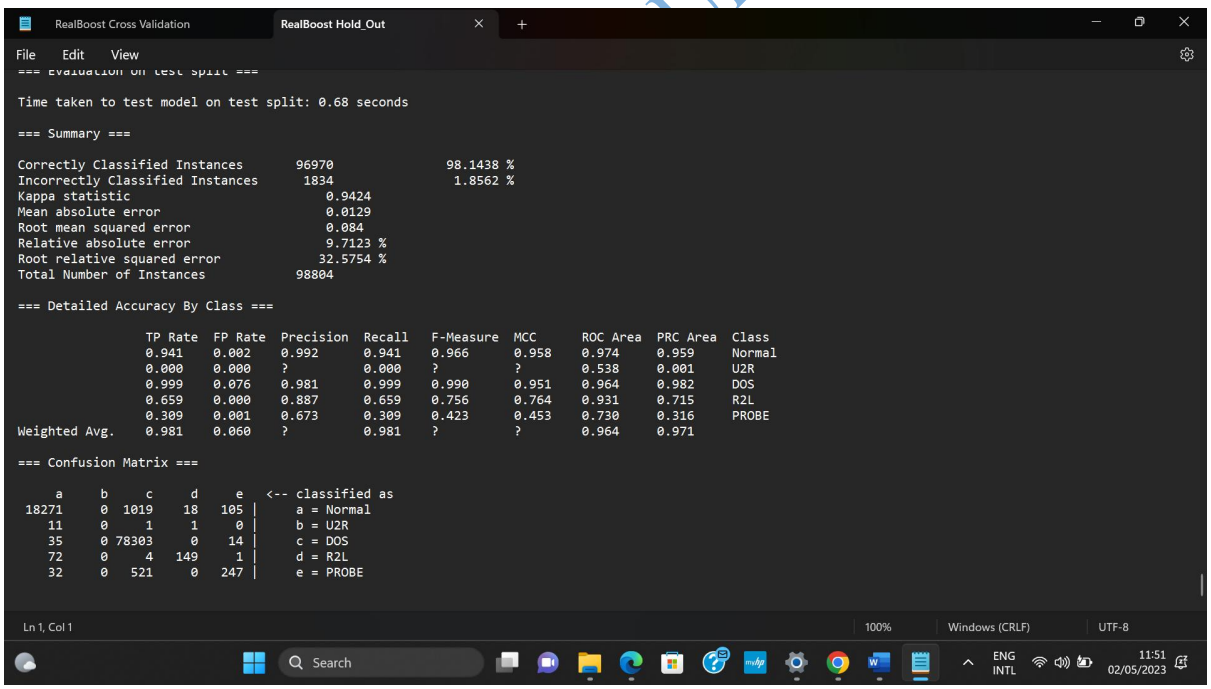
MultiBoost Cross Validation



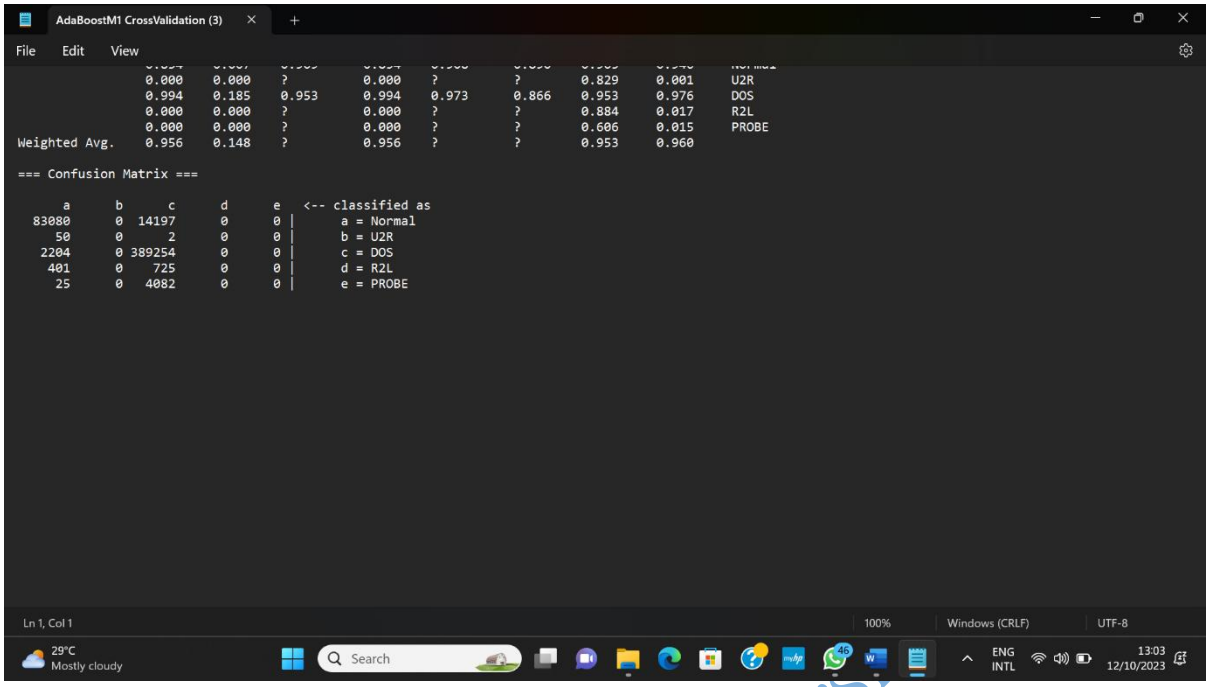
MultiBoost Holdout



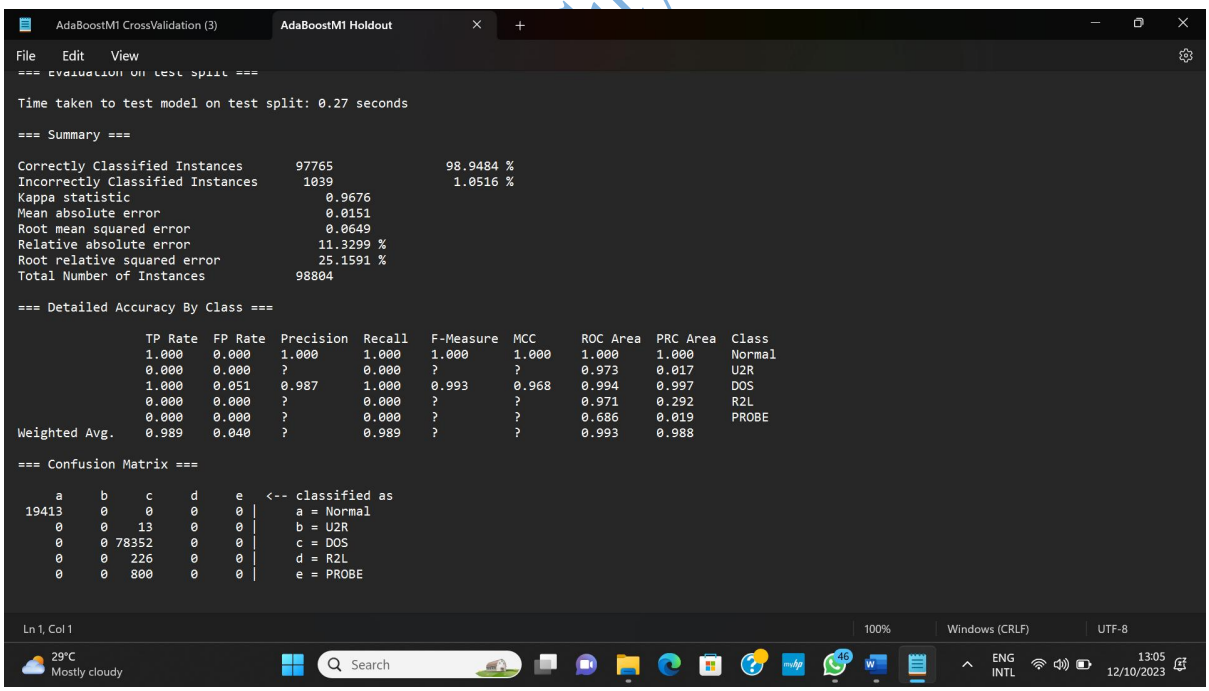
RealBoost Cross Validation



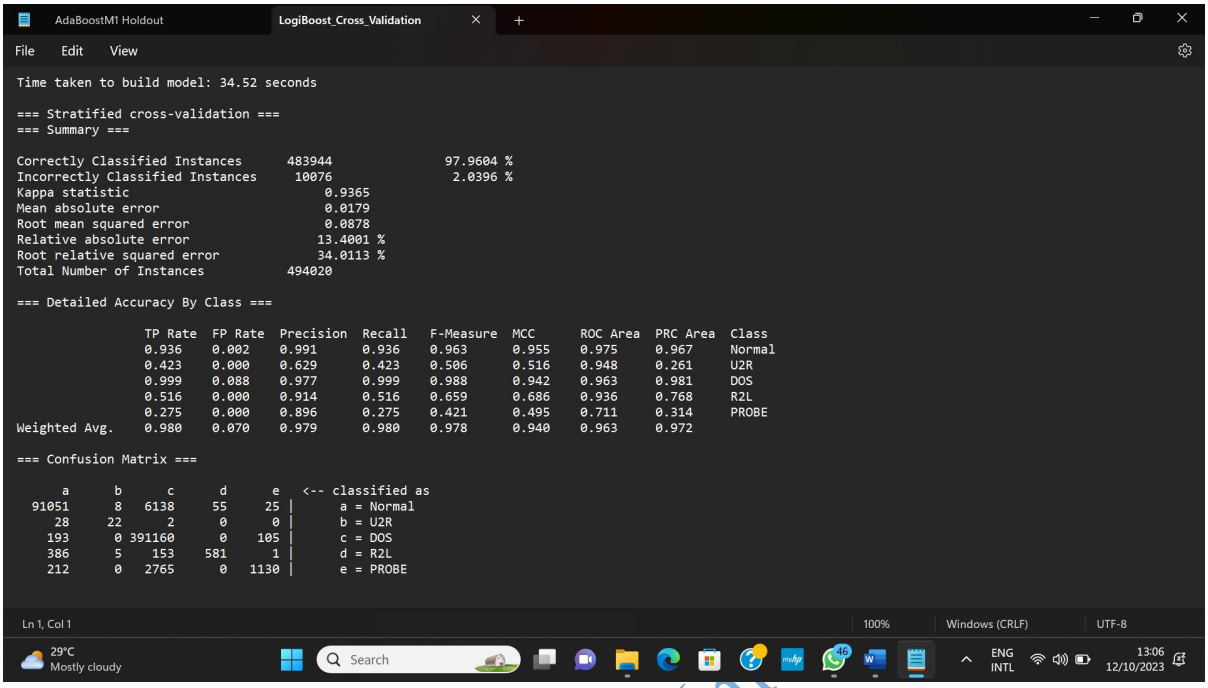
RealBoost Holdout



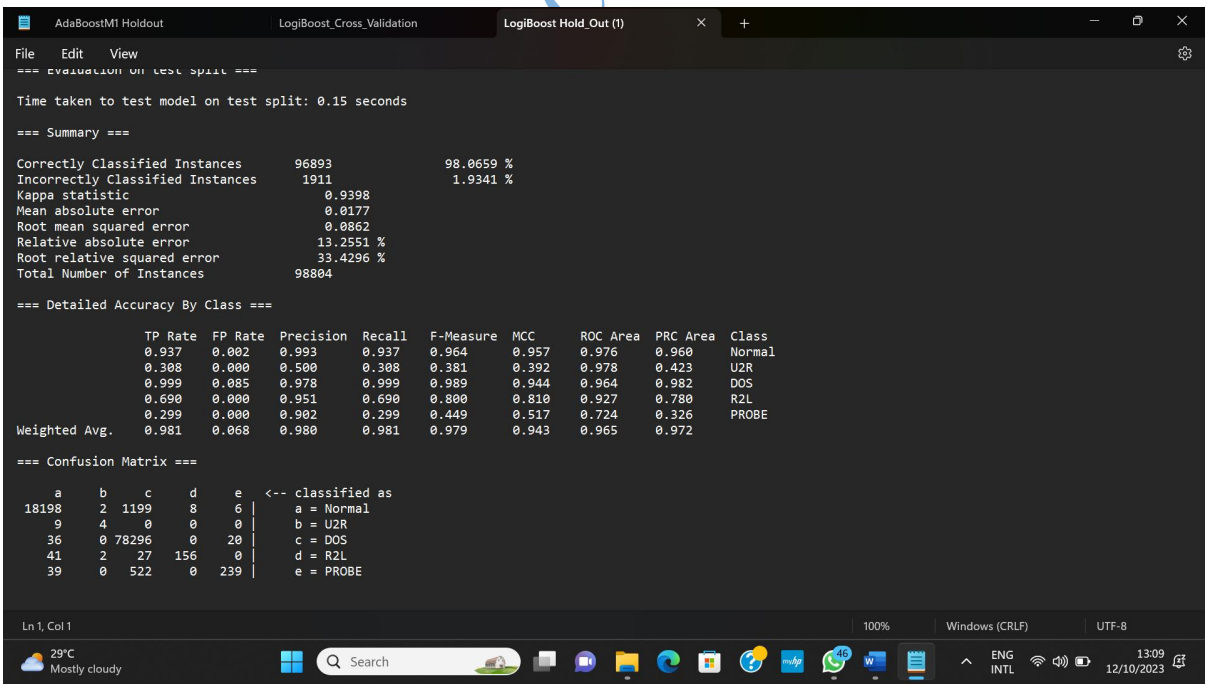
Confusion Matrix for Adaboost Cross Validation



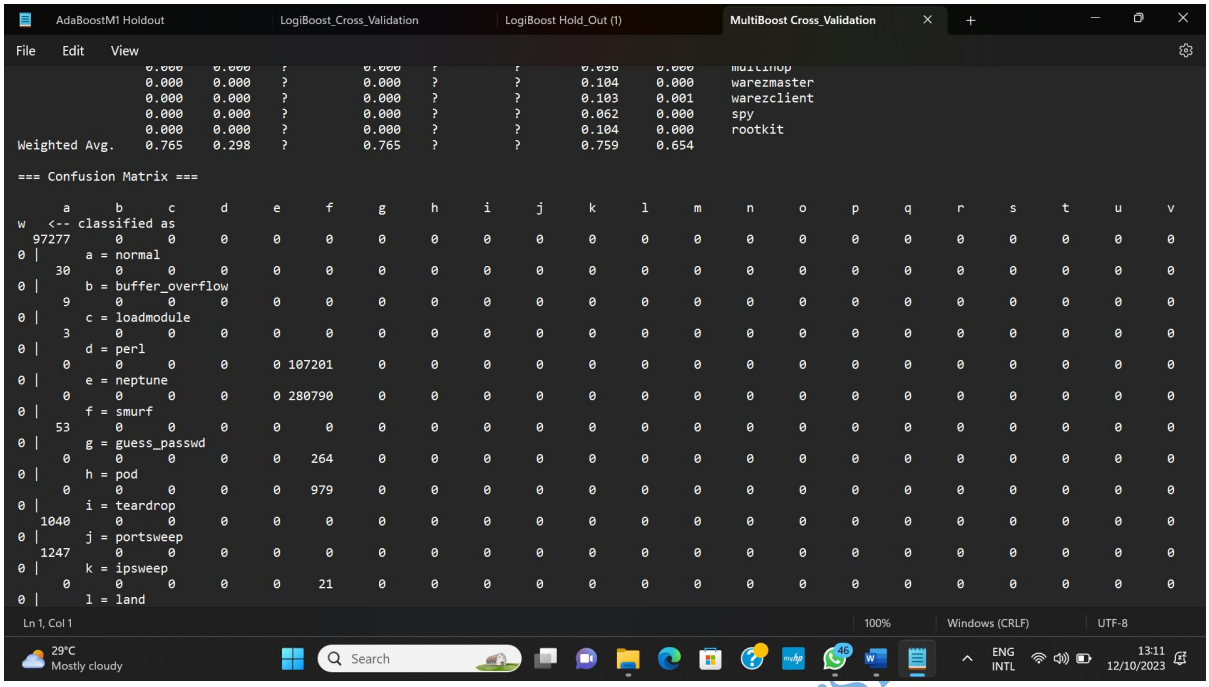
Confusion Matrix for AdaboostHoldOut



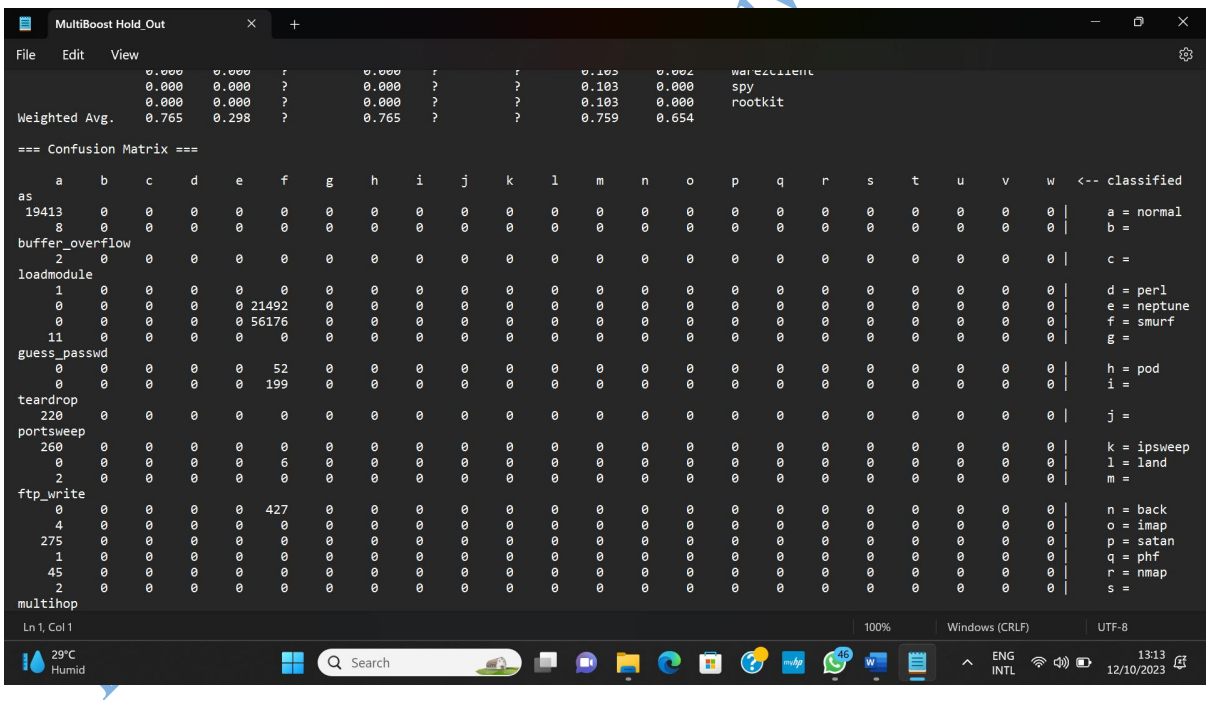
Confusion Matrix for LogiBoost Cross Validation



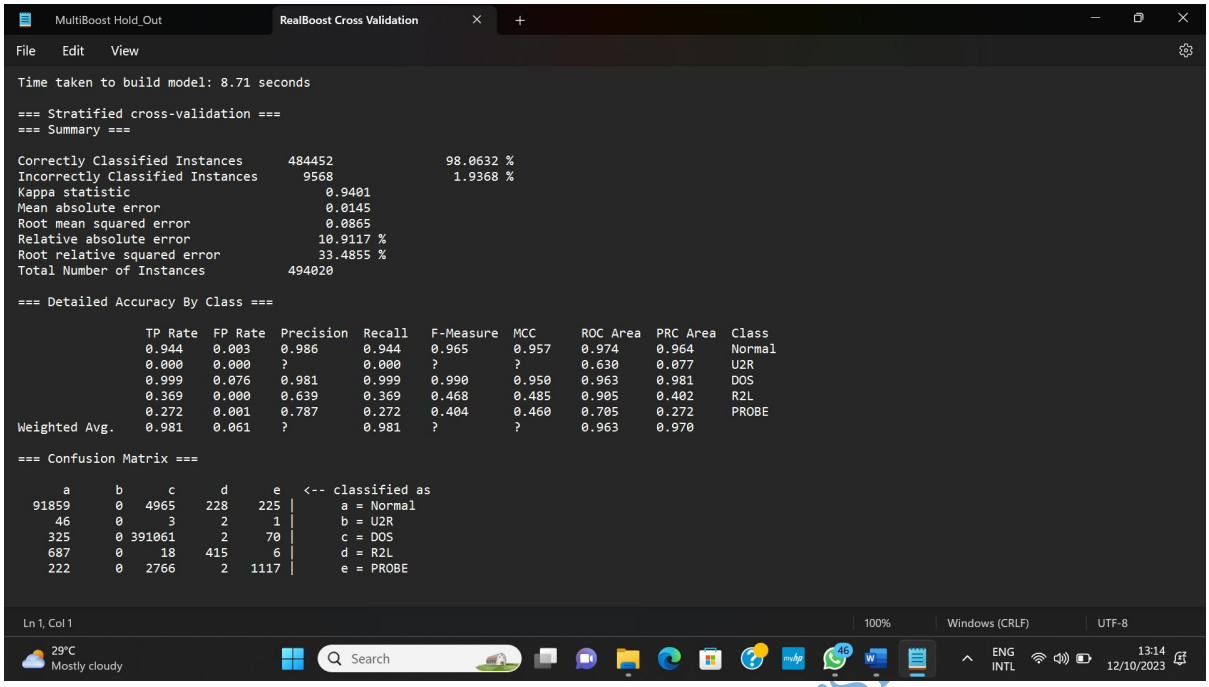
Confusion Matrix for LogiBoostHoldOut



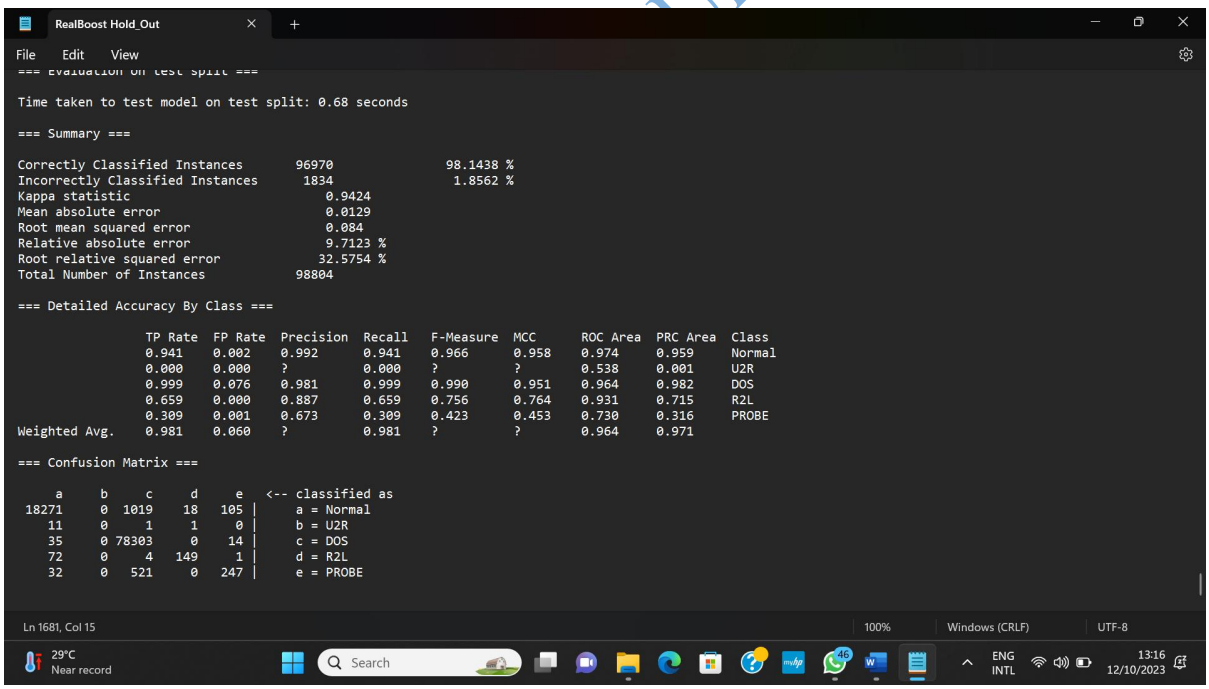
Confusion Matrix for Multi Boost Cross Validation



Confusion Matrix for MultiBoost Holdout



Confusion Matrix for RealBoost Cross Validation



Confusion matrix for RealBoostHoldOut

Biodata

Name: **Jiboku Folahan**

Address: 7, El-Shaddai House, Judah parish str, FPI expressway, Ilaro, Ogun.

E-Mail: jibokufola@gmail.com

Phone Number: 09062050721

Date of Birth: 29th October, 1994

Nationality: Nigerian

Name of Next of Kin: Dr Oladele Jiboku

Address of Next of Kin: Directorate of Linkages and Affiliation, The Federal Polytechnic, Ilaro.

Institutions attended with dates

Babcock University, Ilishan Remo, Ogun State. 2017

Optimum Success College Ilaro, Ogun state. 2011

Grait international College, Ota, Ogun state. 2010

Gospel faith mission comprehensive high school, Ibadan, Oyo State. 2005

Federal Polytechnic Staff Nursery and Primary School,Ilaro, Ogun state. 2004

Academic Qualification with dates

BSc. (Hons) Computer Information System 2017

West African Senior School Certificate (WASSC)2011

First School Leaving Certificate 2004

Working Experience:

Lecturer III April 2023 till Date

Assistant Lecturer (The Federal Polytechnic Ilaro) Nov 2019 till Apr 2023

Internship: Infomix Solutions (Federal Polytechnic ilaro)June 2017 - Oct 2017

PUBLICATIONS:

1. **Title** : BIG DATA TESTING - CHALLENGES AND BEST PRACTICES

   Year of publication: February 2021

   Volume: **Volume - 63, Issue - 06]**

   Journal: **Technology Reports of Kansai University (ISSN: 04532198)**

   Journal ID: TRKU-10-08-2021-11475

CONFERENCES / SEMINAR / WORKSHOP ATTENDED WITH DATES.

1. Ojuawo O.O, and **Jiboku F.J.** (2022). Implementation Of Face Authentication using Active Appearance Model, (A Paper Presented at the 6th National Academic Conference, ASUP Zone C, Federal Polytechnic Ile-Oluji, Ondo State, 19th – 22nd July 2022)

2. **Jiboku F.J,** and Adegboye, O.J. (2022). Data Deduplication Removal in Cloud Computing using File Checksum, (A Paper presented at the 3rd International Conference, The Federal Polytechnic, Ilaro, Ogun State, 16th -17th August, 2022)

3. Sodehinde, V.O, and **Jiboku, F.J.** (2021). The Role of Big Data on Agricultural Growth, (A Paper Presented at The 5th National Conference of The School of Pure and Applied Sciences, Federal Polytechnic, Ilaro, 29th -30th September 2021)

4. **Jiboku, F.J** and Ayodele, E. (2020). Social Media Marketing as a Tool For The Sustainability of Small and Medium Enterprises in Ogun State, (A Paper presented at the 2nd International Conference, The Federal Polytechnic, Ilaro, Ogun State, 10th–11th November 2020)

5. Ojuawo O.O, and **Jiboku F.J.** (2021). The Role of Big Data in Crisis Mitigation.(A Paper Presented at The 5th National Conference of The School of Pure and Applied Sciences, Federal Polytechnic, Ilaro, 29th -30th September 2021)

6. Ojuawo O.O, and **Jiboku F.J.** (2023).Overview Of Edge Computing and Its Significance In The Era Of Iot And Big Data. (A Paper Presented at The4thinternational Conference, The Federal Polytechnic Ilaro in collaboration with Takoradi Technical University, Takoradi, Ghana. 3rd -7thSeptember, 2023)

7. **Jiboku F.J,** and Obarayi Z.O. (2023). User Experience and Interaction Design in Augmented Reality(A Paper Presented at The6th National Conference of The School of Pure and Applied Sciences, Federal Polytechnic, Ilaro, 19th -23thOctober, 2023)

8. Ojuawo O.O, and **Jiboku F.J.** (2023).Design of an Efficient Vehicle Registration System Using Binary Search Algorithm.(A Paper Presented at The 6th National Conference of The School of Pure and Applied Sciences, Federal Polytechnic, Ilaro, 19th -23thOctober, 2023)

Date---------- Signature----------

University Compliance Certification

This is to certify that this thesis by Folahan JIBOKU with Matriculation Number LCU/PG/001817 in the department of Computer Science, Faculty of Engineering and Technology, Lead City University, Ibadan is in full compliance with the approved University's Format and Style.