Designing a Machine Learning-Based Framework for Anomaly-based Network Intrusion Detection

An ML model preventing Cyber Attacks.

Ayushman Singh Raghav
Business Analytics
Dublin Business School
Dublin, IRELAND
10625696@mydbs.ie

ABSTRACT :

This paper aims to provide an overview of the main findings and conclusions of the research conducted. It explores the various aspects and implications of a machine-learning model for Anomaly-based Network Intrusion Detection Numerous enterprises possess an awareness of the significance associated with fortifying their computer networks and mitigating the potential for data breaches through the adoption of various technologies and techniques [1]. Different machine learning methods have been suggested for network intrusion detection, with each technique possessing its own set of advantages and disadvantages [1].

This research aims to present a comprehensive overview of anomaly-based network intrusion detection algorithms [2]. The field of anomaly-based intrusion detection has garnered significant attention in academic research due to its capacity to detect and mitigate previously unidentified security risks [2]. This study will investigate anomaly-based detection methodologies, encompassing machine learning and artificial intelligence methodologies. Furthermore, we will analyze the benefits and drawbacks associated with each methodology. This paper aims to analyze the many obstacles encountered during the implementation of anomaly-based intrusion detection systems.
In addition, we will examine the frameworks and techniques employed in the detection of network anomalies. The following section will give a case study to demonstrate the practical application of these strategies in various contexts [3]. In this study, an assessment will be conducted on the methodologies of machine learning and artificial intelligence, taking into account their respective advantages and disadvantages, as well as the practical obstacles encountered during their application in real-world scenarios [3].
In conclusion, this discussion will focus on the areas of research and the influence of anomaly-based intrusion detection on network security [3].
Keywords- Anomaly Detection, Cyber-attacks, Machine Learning, Network Intrusion Detection, Framework Algorithm.

INTRODUCTION :

The present study focuses on the topic of anomaly detection in the context of cyber-attacks. Specifically, it explores the utilization of machine-learning techniques for network intrusion detection [2]. The primary objective is to develop a framework algorithm that effectively identifies and classifies anomalous activities inside a network environment. As technology progresses, malicious actors consistently adapt their techniques to infiltrate networks and systems. Network intrusion detection has been demonstrated to be a highly effective approach for identifying and mitigating cyber-attacks. This study explores several approaches and technologies used in network intrusion detection, aiming to equip businesses with the essential knowledge required to develop and sustain a reliable intrusion detection system. The user's text is not sufficient to rewrite academically. One of the most recent advancements in the field of cybersecurity is the emergence of big data and machine learning algorithms, enabling the creation of advanced anomaly detection systems. The identification of anomalies holds significant importance across various industries, encompassing manufacturing, banking, healthcare, and cybersecurity. The objective of this technique is to identify atypical patterns or occurrences within a certain dataset that may indicate the presence of a cyberattack or breach in security. The user provided a numerical reference [4]. This paper aims to examine the significance of network intrusion detection in the context of modern cybersecurity. It will explore the role of network intrusion detection in safeguarding valuable resources, mitigating data breaches, and upholding the reputation of organizations. The subsequent sections will go into these aspects in detail. In addition, an examination will be conducted on various network intrusion detection methodologies and tools that have demonstrated efficacy in mitigating emerging cyber risks. Organizations can enhance their ability to combat hackers and protect their digital assets by maintaining awareness of these approaches. The user has provided a numerical reference. The study article will additionally investigate the domain of anomaly identification, an advanced approach that leverages extensive data and machine learning methodologies to accurately detect even subtle deviations. Anomaly detection is a critical function in various businesses due to its ability to mitigate security

breaches, system failures, human errors, and fraudulent activities. The utilization of anomaly detection allows organizations to maintain vigilance and promptly respond to any potentially suspicious behaviour, hence mitigating the occurrence of substantial harm. This is accomplished by the identification of unforeseen patterns. The significance of this study lies in its ability to furnish enterprises and organizations with real knowledge and strategies to enhance their cybersecurity defences. Organizations have the potential to enhance their cyber resilience and mitigate the adverse effects of cyber attacks on their operations, sensitive data, and stakeholders through a comprehensive understanding of contemporary network intrusion detection techniques and the use of state-of-the-art anomaly detection methodologies. The subsequent parts will encompass comprehensive examinations of several network intrusion detection methods, including signature-based detection, behaviour-based detection, and anomaly detection. Each of these techniques possesses distinct advantages and disadvantages, and with a comprehensive understanding of their mechanisms, companies may make informed decisions in selecting the option that aligns most effectively with their objectives and risk profiles. The user provided a numerical reference. Two principal classifications of network intrusion detection methodologies exist signature-based detection and behaviour-based detection.

Signature-Based Detection: Signature-based detection is a cyber threat detection approach that relies on a predetermined set of patterns or signatures. The intrusion detection system (IDS) identifies and categorizes actions as potentially malicious when network traffic conforms to these specific patterns. While signature-based detection demonstrates efficacy in identifying established threats, it may encounter challenges in detecting previously unseen or novel attack techniques. Complex attacks pose a greater challenge for the strategy of relying on attack signature detection due to the frequent utilization of evasion techniques by cybercriminals, which involve making tiny modifications to the attack signature.

This paper will moreover explore the establishment and sustenance of a reliable intrusion detection system. The selection of appropriate hardware and software components, as well as the establishment of best practices for system maintenance and real-time monitoring, can contribute to optimal system performance. Through this process, enterprises will acquire valuable insights on developing efficient strategies to safeguard against cyber threats.

Behaviour-based Detection: In contrast, behaviour-based detection prioritizes the observation of network activity and user behaviour to identify deviations from anticipated trends. This methodology develops a foundational level of anticipated conduct through the utilization of machine learning algorithms and statistical analysis, then issuing alerts when deviations from the norm are detected. The utilization of behaviour-based detection is particularly advantageous in the identification of emerging or previously unknown threats, as it does not depend on established fingerprints. In contrast, it presents a proactive stance towards network security and demonstrates adaptability in response to evolving attack techniques.

Using Anomaly Detection to Improve Network Intrusion Detection: In recent years, the integration of big data and machine learning has facilitated the development of advanced anomaly detection methodologies. Anomaly detection plays a crucial role in various industries by facilitating the timely identification of system vulnerabilities, human errors, fraudulent activities, and potential cyber threats. This capability enables early detection, which is of paramount importance. The utilization of anomaly detection allows security analysts to promptly respond to potential assaults by identifying and discerning unforeseen patterns within a given dataset.

Three distinct categories of anomaly detection approaches may be identified: statistical methods, machine learning methods, and hybrid methods.

Statistical-Based Anomaly Detection: The process of statistical-based anomaly detection involves the calculation of statistical measures such as mean, variance, and standard deviation. These measures are then used to compare incoming data points with prior data to identify any anomalies. An anomaly is defined as a data point that exhibits a deviation from the predetermined statistical boundaries. While this methodology is characterized by its simplicity and efficacy, it may encounter challenges in identifying intricate and non-linear irregularities.

Anomaly Detection Based on Machine Learning: Machine learning-based anomaly detection involves the training of models using labelled data to acquire knowledge about the typical behaviour of a system or network. After undergoing training, the models possess the ability to identify anomalies, which refer to deviations from the acquired patterns of behaviour. This methodology demonstrates a high level of efficacy in countering emerging cyber threats due to its ability to effectively process complex data patterns and dynamically adjust to evolving environments.

Hybrid Methods: Hybrid methods, which combine machine learning and statistical techniques, are employed for the identification of abnormalities. The utilization of hybrid models provides an opportunity to improve the precision and resilience of anomaly detection in diverse datasets and scenarios through the integration of advantages derived from both approaches.

The increasing significance of anomaly detection in the realm of cybersecurity is attributed to the continuous advancement of attack strategies employed by cybercriminals, aimed at evading traditional security measures. As the volume of network traffic and system complexity escalates, human analysts have limitations in effectively managing the corresponding data influx. Anomaly detection serves as a valuable force multiplier by aiding analysts in the prioritization and focus on high-risk events, while

simultaneously mitigating the occurrence of false positives. The increasing scope of cyber threats underscores the need for a comprehensive and adaptable strategy for cybersecurity. Organizations have the potential to enhance their defensive capabilities against a diverse array of cyber-attacks through the integration of network intrusion detection techniques alongside state-of-the-art anomaly detection systems. Cybersecurity professionals may effectively mitigate risks and safeguard sensitive information through the utilization of big data and machine learning techniques, enabling them to proactively stay ahead of malicious actors. To effectively tackle emerging challenges and foster a safe digital milieu for all individuals, it is imperative to emphasize the significance of ongoing scholarly investigation and cooperative efforts across academia, industry, and governmental entities. This collaborative approach is essential to the advancement of the field of cybersecurity.

The study begins by highlighting the increasing prevalence of computer network threats and the need for Network Intrusion Detection Systems (NIDS) to safeguard network security. The placement of a typical Network Intrusion Detection System (NIDS) at a singular point is capable of detecting external threats, but it cannot identify malicious network traffic. One potential solution to this issue is the implementation of distributed deployment, wherein Network Intrusion Detection Systems (NIDS) are connected to critical routers and gateways. Traditional neural network-based network intrusion detection systems (NIDS) impose limitations on simple network devices, such as:

a. Offline processing: Training a supervised model for small gateways with limited memory is infeasible due to the requirement of having all labelled instances accessible locally.

b. Supervised learning: The process of manually categorizing network data as either genuine or malicious incurs both temporal and financial costs, while also being susceptible to the emergence of novel attack methodologies.

c. High complexity: The computational complexity of neural networks exhibits an exponential growth pattern as the number of neurons rises, hence rendering them unsuitable for direct utilization as network gateways.

In the previously described study, the dependent variable is referred to as "Label." This research presents a Network Intrusion Detection System (NIDS) that utilizes neural networks and employs an ensemble of autoencoders for anomaly detection. The system will monitor the behaviour of all network channels and assign an anomaly score to arrival packets. Anomaly scores are employed to determine whether a packet is classified as abnormal or normal. The study evaluates the system's ability to recognize and perform in real-time, while also comparing the online method to batch or offline procedures. This study aims to effectively identify anomalies across many platforms, such as Raspberry Pi and Ubuntu virtual machines. The anomaly score is a crucial determinant employed in this research to assess the safety or possible harm of arriving network packets.

LITERATURE REVIEW:

The growing dependence on computer networks and the Internet has amplified the significance of ensuring their security[4]. To protect network infrastructures against cyber threats, such as network invasions, improved intrusion detection systems (IDS) have been developed. An important technique within this domain centres on the detection of abnormalities from typical network behaviour, commonly referred to as anomaly-based network intrusion detection [4]. The objective of this literature review is to present a comprehensive examination of the fundamental principles, research approaches, difficulties, and progressions in the field of anomaly-based network intrusion detection [5].

Anomaly-based network intrusion detection (NID) is a methodology employed to identify network intrusions through the observation of alterations in network activity [5]. The research subject of anomaly-based network intrusion detection (NID) has gained popularity within the field of network security due to the intricate nature of network systems and the growing sophistication of attackers. This study aims to investigate the studies about anomaly-based network intrusion detection (NID) [5].

1. Conceptual Framework:

According to the principles underlying anomaly-based intrusion detection, the detection of atypical network behaviour that deviates from the set baseline may indicate the presence of security breaches [5]. These anomalies may emerge as atypical traffic patterns, unauthorized attempts to access, or aberrant utilization of system resources, among other indications. To effectively identify abnormalities, it is important to establish a representative model of typical behaviour against which deviations can be evaluated. The user provided a numerical reference [5].

Now elucidating conceptual frameworks using simplified language. Analogous to a dynamic thoroughfare, the computer network can be perceived as a bustling highway, facilitating the exchange of data across diverse devices, akin to the movement of vehicles [5]. The traffic standards that govern road safety are equally applicable to computer networks, thereby guaranteeing their smooth functioning and

safeguarding against potential threats. Similar to how negligent drivers ignore traffic regulations, certain malicious individuals endeavour to breach computer network security standards to steal data or cause harm [5]. An intrusion detection system (IDS) is utilized to mitigate the activities of individuals who violate established regulations, similar to the function of a security camera system [5].

- **Normal vs. Anomalous Behavior:** The usual behaviour of a computer network can be analogized to the orderly flow of traffic on a highway, wherein all participants adhere to established regulations and exhibit predictable conduct [5]. Abnormal behaviour can be characterized as the antithesis of the norm. The primary objective of the intrusion detection system is to discern and classify instances of customary behaviour. This is achieved through the constant monitoring of the network and the construction of a representation of what is considered to be the standard or expected state [5]. The depicted graphic serves as a reference manual that facilitates the system's ability to identify anomalous occurrences[5].
- Anomalies might be likened to unanticipated occurrences on the road that may give rise to apprehension [5]. Aberrant conduct might be exemplified by the unexpected movement of a vehicle in the other direction or at an excessive speed. In a similar vein, deviations in network traffic from the norm may serve as an indication of potentially suspicious or hazardous activities [5].
- **Creating the Normal Behavior Model:** The Intrusion Detection System (IDS) collects a substantial amount of data about network activities to gain an understanding of common behavioural patterns [5]. The analysis encompasses various aspects, including the volume of data transmitted and received, the parties involved in communication, and the frequency of such interactions [5]. The network's typical manifestation is then established as a form of "blueprint" or "pattern" utilizing this data. The Intrusion Detection System (IDS) is capable of detecting and identifying any atypical behaviours by utilizing the provided blueprint [5].

2. **Detecting Anomalies:** When the Intrusion Detection System (IDS) is operational, it consistently compares the ongoing network activity with a pre-established model of anticipated behaviour [5]. The process can be likened to how a security camera system assesses the existing traffic conditions about the projected traffic flow [5]. The Intrusion Detection System (IDS) promptly alerts the user by virtually signalling its concern and expressing surprise when encountering anomalous events, such as an unexpected surge in data during periods of relative inactivity [6].

METHODOLOGY AND TECHNIQUE:

- **Statistical Techniques**: Baselines of typical behaviour have been established using statistical techniques like mean, median, standard deviation, and clustering. Alerts are sent when deviations surpass predetermined levels. [10].

- **Machine Learning:** Machine learning algorithms, including decision trees, support vector machines, and neural networks, have gained significant popularity in the field of anomaly detection. According to the cited source, these models possess the capability to identify complex patterns within network data and exhibit adaptability over time [10]. Consider a hypothetical scenario in which a security guard is accompanied by a remarkably clever assistant who possesses the ability to utilize their existing expertise. This AI helper possesses the ability to retain not only conventional patterns of behaviour but also intricate details such as regularly visited locations by groups of individuals or certain periods when certain behaviours are more commonly observed. Machine learning can be likened to a personal assistant inside the domain of networks. The system can acquire knowledge from past network data and identify patterns that may pose potential issues. For example, it could emit an audible alert if it detects a sequence of behaviours that deviate from expected patterns.

- **Data Mining:** The utilization of clustering and association rule mining techniques in data mining has been employed to identify concealed connections within network data, which could potentially indicate suspicious behaviour [11].
  It is presumed that the security officer possesses authorization to utilize a database including comprehensive information about individuals who get entry to the premises [6].
  The database can encompass information about genealogical relationships as well as patterns of frequent co-occurrence among individuals [6].
  The security personnel can discern atypical associations among persons by examining the provided data, such as the occurrence of two unfamiliar individuals entering nearby. In a similar vein, the process of data mining in networks involves systematically examining the network data to identify latent associations or anomalous patterns that may indicate the presence of an issue [6].

- **Time-Series Analysis**: Time-series methodologies are employed to monitor alterations in network data over some time and identify indications of potentially detrimental behaviour. Frequently employed methodologies encompass moving averages, Fourier transforms, and autoregressive integrated moving averages (ARIMA) models [12]. Time-series analysis might be likened to a clock utilized by a security guard to effectively watch the patterns of individuals entering and exiting a building.

  The arrival of a previously unexperienced temporal point can be deemed remarkable. The field of time-series analysis in networks focuses on the investigation of temporal patterns, specifically the examination of trends over time. This entails the study of several factors, such as the volume of data that is frequently received within specific time intervals, such as hourly or daily periods.
  A sudden surge or decline in data that deviates from the usual observations within a given period may suggest the presence of potential hazards. In certain instances, a network may exhibit a higher frequency of routine activity as opposed to exceptional behaviour.
  This scenario might be likened to a situation where multiple guests are exhibiting appropriate behaviour, except for one individual who is displaying peculiar conduct. The system may have difficulties in detecting and comprehending a singular anomalous activity as a consequence. Moreover, networks change throughout time as a result of the implementation of novel software and the integration of additional devices. If the system lacks the necessary capabilities to handle modifications, it may lead to confusion for the users.

3. Challenges and Limitations:

- **Imbalanced Data:** The predominant portion of the data found in network traffic often corresponds to conventional patterns of behaviour. As a consequence of this, the task of accurately detecting rare anomalies while simultaneously minimizing the occurrence of false positives may provide challenges.
- Let us contemplate a conglomeration of confectionery items wherein the predominant variety is that of chocolate candies. A limited assortment of additional confections, such as gummy bears or lollipops, are currently accessible. In the context of computer networks, routine behaviour tends to be more prevalent than unexpected or problematic activity. This issue may arise due to the potential for the security guard system to excel in recognizing typical items, such as chocolate confectionery, while displaying lower proficiency in identifying atypical goods, such as gummy bears. Hence, notwithstanding its infrequency, the security guard must acquire the skill of attentiveness towards anomalous occurrences.
- **Dynamic Environments:** Networks are subject to constant change, and making reasonable modifications can lead to the occurrence of anomalies. The Intrusion Detection System (IDS) should include the capability to distinguish between authorized alterations and malicious activities [11].
  Consider the process of reconfiguring the furniture arrangement within one's living space, relocating personal playthings, or introducing novel objects. Computer networks are subject to continuous

transformation. The system must possess the cognitive ability to differentiate between intentional alterations made by the user, such as the voluntary relocation of toys, and adjustments that may potentially arise from evil intentions. The task presents a considerable challenge as the security personnel must discern whether a modification is within the realm of regularity or a shrewd endeavour to get access.

- **Unknown Attacks:** The efficacy of anomaly-based detection in identifying novel assaults that have not been encountered previously may be constrained by the insufficiency of comprehensive historical data for accurate modelling.
- Consider a scenario where you are engaged in a game of hide-and-seek with a companion, and they introduce a novel hiding spot that you have not previously encountered. Hackers possess the ability to develop innovative methodologies within computer networks that are unfamiliar to security personnel. Due to the security guard's lack of familiarity with these novel techniques, he encounters difficulty in discerning their malevolent nature. Consequently, it is imperative for security personnel to promptly acquire proficiency in these emerging procedures and discern their inherent risks, recognizing them as potentially perilous conduct rather than simple novel tactics employed in the game.
- **Feature Selection:** A significant problem is identifying pertinent features for anomaly detection and managing the complexity of data [12]. Consider sharing a story with a buddy. The pieces you choose to share are the most significant and captivating. There is a lot of information in computer networks, such as who is conversing with whom and the nature of the information being delivered. The security guard must determine which details of this information require the greatest attention. It's similar to selecting your story's most interesting details to keep your friend engaged.

4. This review article presents an overview of several cyber-physical system intrusion detection methods, including anomaly-based NID. The authors explore the drawbacks of anomaly-based NID and emphasize the necessity of a multi-layered security strategy that combines both signature-based and anomaly-based approaches [9].

5. In this research, the authors present an extreme learning machine (ELM)-a based innovative method for anomaly-based NID. In tests utilizing the KDDCUP'99 dataset, the suggested ELM-based NID system demonstrated great accuracy and a low false alarm rate for the detection of abnormalities in real-time network traffic [10].

6. The authors of this research suggest a convolutional neural network (CNN) an anomaly detection system that is based on deep packet inspection. In tests utilizing the UNSW- NB15 dataset, the suggested system, which is intended to identify network abnormalities in real time, demonstrated excellent accuracy and a low false alarm rate [11].

7. This review study presents an overview of several deep learning approaches, including anomaly-based NID, utilized for network intrusion detection. The authors outline some recent developments in this field and go into the benefits and drawbacks of utilizing deep learning for NID [12].

8. The authors of this survey study give a general review of the various machine learning techniques that are utilized for anomaly-based NID. The authors explore the benefits and drawbacks of each method and emphasize the necessity for hybrid models, which bring together several algorithms to boost the reliability and accuracy of anomaly-based NID. (Anomaly detection in network traffic using machine learning algorithms: A survey by M. Shabib et al. [13] These studies emphasize the significance of anomaly-based NID for network security and the demand for cutting-edge attackers. We will look at some current research on anomaly-based NID in this literature review.

9. This review article presents an overview of several cyber-physical system intrusion detection methods, including anomaly-based NID. The authors explore the drawbacks of anomaly-based NID and emphasize the necessity of a multi-layered security strategy that combines both signature-based and anomaly-based approaches [9].

10. In this research, the authors present an extreme learning machine (ELM)-a based innovative method for anomaly-based NID. In tests utilizing the KDDCUP'99 dataset, the suggested ELM-based NID system demonstrated great accuracy and a low false alarm rate for the detection of abnormalities in real-time network traffic [10].

11. The authors of this research suggest a convolutional neural network (CNN) an anomaly detection system that is based on deep packet inspection. In tests utilizing the UNSW- NB15 dataset, the suggested system, which is intended to identify network abnormalities in real time, demonstrated excellent accuracy and a low false alarm rate [11].
12. This review study presents an overview of several deep learning approaches, including anomaly-based NID, utilized for network intrusion detection. The authors outline some recent developments in this field and go into the benefits and drawbacks of utilizing deep learning for NID [12].
13. The authors of this survey study give a general review of the various machine learning techniques that are utilized for anomaly-based NID. The authors explore the benefits and drawbacks of each method and emphasize the necessity for hybrid models, which bring together several algorithms to boost the reliability and accuracy of anomaly-based.

14. **Conclusion:** This research study highlights the importance of anomaly-based network intrusion detection as a fundamental tool for the identification and prevention of potential network breaches. To safeguard network infrastructures from emerging cyber threats, the effectiveness of this approach relies on the identification of anomalies in established patterns of conventional network activity.

The landscape of network security has undergone significant transformations as our dependence on digital communication and data exchange has increased. Consequently, contemporary networks have become significantly vulnerable to malicious activities such as the infiltration of malware, the unauthorized extraction of data, and attempts to gain unauthorized access. In the realm of cybersecurity, intrusion detection systems (IDS) have emerged as a pivotal component in the arsenal of strategies employed to counteract such attacks. Among these techniques, anomaly-based detection has garnered significant attention due to its effectiveness**.** The theoretical foundation of anomaly-based intrusion detection revolves around the notion that the observation of atypical behaviour might serve as an indicator of potential security breaches. Anomalies might manifest in diverse forms, encompassing atypical traffic patterns, anomalous data transfers, or alterations in resource utilization within the system. The Intrusion Detection System (IDS) constructs a foundational model that represents normal patterns of behaviour. This model serves as a reference point against which deviations are evaluated to effectively identify and detect anomalies. The utilization of the comparative method facilitates the identification of variations that may indicate potential security vulnerabilities. This paper examines the various techniques and approaches employed in anomaly-based detection. Typical network activity baselines have been constructed by the utilization of statistical approaches such as mean, median, standard deviation, and clustering. These methodologies provide a foundation for establishing thresholds that, when exceeded, trigger warnings and necessitate further investigation. The utilization of machine learning techniques, including decision trees, support vector machines, and neural networks, has gained significant traction in the domain of anomaly detection due to their ability to discern intricate patterns and adapt to changing network dynamics. Data mining techniques have also been employed to identify concealed relationships within network data that may indicate anomalous behaviour. An alternative approach for anomaly detection is the utilization of time-series analysis, wherein data is monitored and analyzed over a specific period to identify patterns, variations, and trends.

The aforementioned publications also highlight contemporary advancements that address these challenges and enhance the functionalities of anomaly-based detection. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are two prominent deep learning algorithms that have demonstrated remarkable efficacy in the identification of complicated connections and patterns within network data. Unsupervised learning, a type of machine learning that operates without the need for labelled data, has demonstrated an enhanced capacity to effectively respond to emerging threats. Hybrid systems that incorporate anomaly-based detection alongside signature-based and behaviour-based detection techniques offer a comprehensive security strategy against a wide range of attack vectors. Generative models such as generative adversarial networks (GANs) and autoencoders aid in the identification of previously unidentified threats by generating synthetic data samples that represent rare anomalies.
The result of the literature analysis emphasizes the significance of employing anomaly-based network intrusion detection as a means of safeguarding the security of modern network infrastructures. The field of intrusion detection systems has witnessed significant advancements in recent times, particularly in the areas of deep learning, unsupervised learning, and hybrid techniques. These discoveries present intriguing possibilities for enhancing the precision and flexibility of such systems, even in the face of challenges such

as imbalanced data and the emergence of new types of attacks. The continuous development of innovative techniques and methodologies is crucial in effectively countering cyber threats and safeguarding the integrity and confidentiality of network data, given the evolving nature of the threat landscape.

METHODOLOGY:

Research Approach and Philosophy:

The present thesis employed an interpretive research methodology. The interpretive research perspective emphasizes comprehending and contextualizing events occurring within their natural settings, to explore their relevance and intricacies [6]. This particular approach is particularly suitable for research topics that encompass intricate aspects of human perceptions, behaviours, and experiences. Interpretive research facilitates a comprehensive examination of the various factors that impact the effectiveness, challenges, and consequences of anomaly-based network intrusion detection in the ever-changing cybersecurity environment [7].

The present thesis employed an inductive research style. rather than conducting tests on previous assumptions, an inductive approach involves the formulation of ideas based on observed data. Observations are conducted after the collection and analysis of qualitative or quantitative data. These patterns and insights play a crucial role in informing the development of novel theoretical frameworks or perspectives. The utilization of an inductive methodology facilitates a comprehensive exploration of the many approaches, strategies, challenges, and advancements in anomaly detection within the domain of cybersecurity, as it relates to the subject matter of the thesis.

The amalgamation of an inductive approach and an interpretive research perspective yields a holistic understanding of the domain of network intrusion detection that is based on anomalies. This approach facilitates the examination of both quantitative measures and qualitative elements, encompassing challenges faced by organizations in implementing these strategies, impacts of emerging technologies, and perspectives of cybersecurity experts on the future trajectory of this domain. By employing this approach, the study can comprehensively comprehend the intricacy and ever-changing characteristics of the research matter, without being limited by preconceived notions.

In summary, the utilization of the inductive research technique and the interpretive research philosophy in the study of anomaly-based network intrusion detection allows for a comprehensive examination of the subject matter, encompassing both the technical and contextual aspects.

Research Design:

The research conducted in my thesis is primarily qualitative and employs an exploratory approach, utilizing a case study methodology. The selection of this architecture aligns well with the objectives of the project, which include the development of a framework based on machine learning for network intrusion detection using anomaly detection techniques, as well as an evaluation of the performance of several anomaly detection methods in real-world scenarios. The present analysis aims to scrutinize the chosen research design and evaluate its appropriateness for the study at hand.

Qualitative Approach: The qualitative methodology is deemed suitable for attaining the objectives of my study as it facilitates a comprehensive exploration of anomaly-based network intrusion detection systems. This experience allowed me to acquire a comprehensive understanding of the intricacies, challenges, and subtleties associated with implementing such methodologies in practical scenarios. The utilization of a qualitative method facilitates a comprehensive examination of the various components of my research, as it encompasses the exploration of multiple techniques, theories, and tools.

Case Study Method: The utilization of a case study design involves doing a comprehensive analysis of a specific instance or situation, in this context, a particular business or network environment where anomaly-based intrusion detection techniques are implemented. This approach is deemed appropriate as it allows for the evaluation of the efficacy of the methodologies, the challenges faced, and the outcomes within a real-world context. By researching a specific scenario, I can provide a comprehensive and contextualized elucidation of the functioning of these tactics and their impact on network security.

The primary objective of this study is to examine the merits, drawbacks, and challenges associated with anomaly-based network intrusion detection methods. Qualitative data can be gathered through many methods such as interviews, observations, and the analysis of relevant documents, which aligns with the exploratory nature of the case study approach. It is imperative to gather data through this form of data collection to capture the perspectives and experiences of network administrators, cybersecurity specialists, and other stakeholders engaged in the deployment and utilization of these technologies.

In-depth Analysis: Employing a case study methodology, a comprehensive analysis was undertaken to investigate a specific scenario, delving into its intricacies and intricacies that may elude more extended quantitative inquiries. This level of inquiry allows for the provision of a nuanced and all-encompassing analysis of the topic.

In summary, the utilization of the qualitative case study approach proves to be highly advantageous in attaining my research objectives as it facilitates the examination and comprehension of the practical implementation of anomaly-

based network intrusion detection algorithms. By focusing on a specific scenario, I may provide relevant insights that consider both technical and organizational elements, which collectively add to the overall effectiveness of these approaches.

Data Analysis in Anomaly-Based Network Intrusion Detection

Data analysis forms the foundation of anomaly-based network intrusion detection (NID), serving as the basis for the detection algorithms. To establish a foundational model of customary behaviour, the methodology involves a meticulous examination of network data. This model functions as a benchmark for assessing deviations or anomalies. The objective is to identify any behaviour that diverges from the norm and may signify a potential security issue.

The studies referenced in the literature emphasize the significance of data preparation as a crucial stage in this particular process. Various strategies are employed to establish benchmarks for typical network activity, encompassing measures such as mean, median, standard deviation, and clustering [1][7]. Preprocessing techniques can be employed to establish thresholds, which, upon being exceeded, can trigger notifications to alert possible irregularities. The utilization of data-driven thresholds facilitates the implementation of efficient and automated anomaly detection mechanisms, hence enhancing the overall security stance of the network.

Analyzing data and using machine learning:

Machine learning approaches have been demonstrated to be indispensable tools in anomaly-based network intrusion detection (NID). The aforementioned algorithms, namely decision trees, support vector machines (SVMs), and neural networks, possess the ability to identify intricate patterns and adapt to changing network dynamics[3]. In this context, the integration of machine learning and data analysis is observed. Machine learning algorithms provide the capability to discern latent patterns from historical data that may not be readily apparent through manual study. Relevant features for anomaly detection are identified and selected by comprehensive data analysis. The inclusion of this selection process is crucial to minimize the intricacy of the data while still preserving its capacity for prediction [9]. Machine learning algorithms can acquire knowledge about the distinctive attributes of regular network activity by utilizing the identified features as inputs.

Developments Using Data Analysis:

Recent advancements in deep learning, unsupervised learning, and hybrid methods have been driven by the valuable insights obtained from data analysis [12]. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are deep learning algorithms that exhibit strong proficiency in the identification of intricate patterns within network data [12]. Incorporating unsupervised learning, intrusion detection systems provide the ability to promptly adjust to emerging threats without dependence on labelled data [12]. The development of a comprehensive security strategy against many attack vectors involves the integration of anomaly-based detection with signature-based and behaviour-based techniques, guided by data analysis [4].

Validity:

Validity is the term used to describe the precision and correctness of the inferences derived from a research investigation. Ensuring the validity of a publication on anomaly-based network intrusion detection is crucial to accurately reflect real-world scenarios and facilitate the generalizability of the study's findings beyond its specific context.

Internal validity: This question pertains to the extent to which the research design and technique effectively capture the intended variables. Ensuring internal validity in the study involves ensuring that the anomaly detection tools employed, such as statistical methodology, machine learning algorithms, and time-series analysis, accurately and effectively capture and identify aberrant network behaviour. It is imperative to acknowledge and account for any potential biases or confounding variables that may influence the outcomes.

Regarding validity, we will now proceed to examine the three machine learning models and algorithms that played a crucial role in this particular case study.

1. Random Forest Classifier
2. Support Vector Classifier
3. Xtreme Gradient Boosting.

Before delving into an in-depth analysis of these models, it is imperative to acknowledge the significance of data visualizations and preprocessing in ensuring the efficacy of these three models. Consequently, certain libraries have been taken into account.

Libraries Taken:

1. NumPy
2. Pandas
3. sklearn.preprocessing.StandardScaler:
4. sklearn.model_selection.GridSearchCV:
5. imblearn.over_sampling.SMOTE:
6. sklearn.ensemble.RandomForestClassifier:
7. imblearn.pipeline.Pipeline:
8. sklearn.svm.SVC:
9. xgboost.XGBClassifier:
10. sklearn.model_selection.train_test_split:
11. sklearn.metrics.recall_score:
12. sklearn.metrics.classification_report:

13. sklearn.metrics.confusion_matrix:

Upon reaching the section about dataset explanation, it is worth noting that the dataset in question, referred to as 'Attack_Dataset', comprises a total of 0.3 million rows and 41 columns. In this case study, I have compiled a comprehensive dataset comprising all nine types of cyberattacks. This dataset encompasses a total of 300,000 rows.
The initial step is conducting data exploration and analysis operations on the "Attack_Dataset.csv" CSV file using the pandas library.

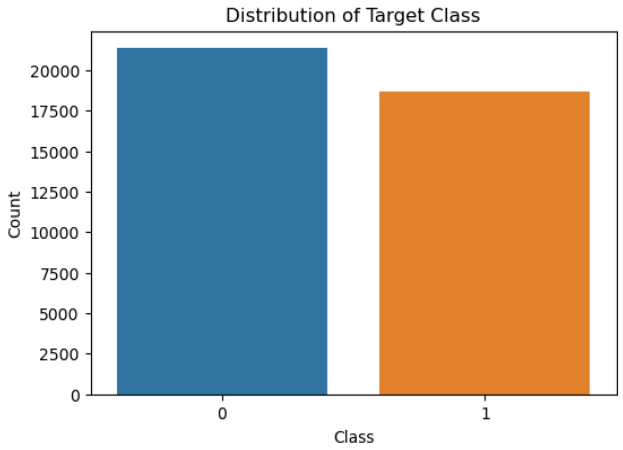
Upon reviewing the dataset, our initial step is data preparation for machine learning methods. This entails transforming the categorical variable 'class' into numerical features. Specifically, we will convert the phrases "normal" and "anomaly" into the numerical values 0 and 1, correspondingly. In the context of binary classification problems, it is common practice to convert categorical labels into numerical values by encoding. If this conversion is performed, the data will be transformed into a format that is suitable for training machine learning models. Once the 'class' column has been assigned numerical labels, these labels can be employed as target values for the classification task.

Next, apply one-hot encoding to the category features of the DataFrame to convert them into a numerical representation that is suitable for machine learning techniques. One-hot encoding was utilized to generate binary columns for each unique category within categorical characteristics, resulting in a representation of the data in a binary format. In the context of features such as "service," "flag," and "protocol_type," when there is no inherent order among categories, it is crucial to ensure that ordinal connections are not introduced where they do not already exist.

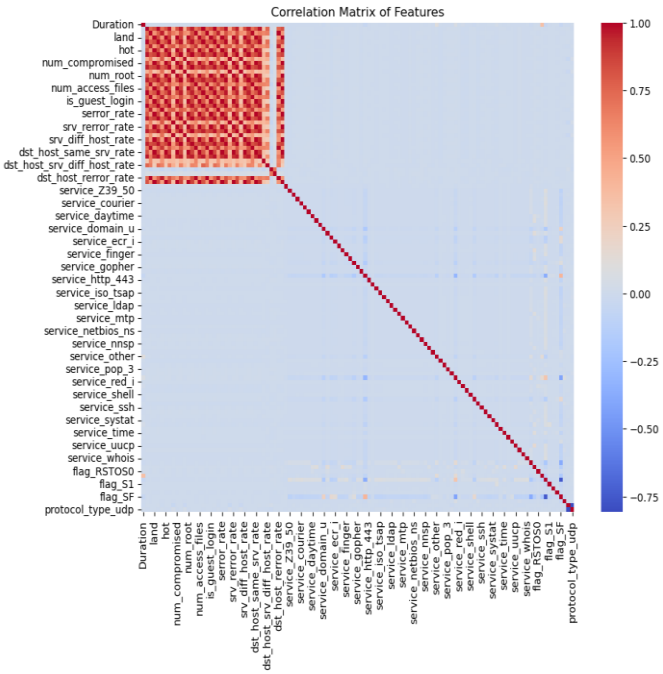
To facilitate the training of a machine learning model, it is necessary to partition the data into distinct sets of features (X) and labels (Y).
The dataset was partitioned into two distinct components: the feature matrix X, which encompasses the input variables, and the label vector Y, which encompasses the target values. The separation of features and labels into different variables is a common data preparation step in machine learning algorithms.

Visualizing the dataset by using various plots:

Count Plot: The purpose of this tool is to visually represent the dispersion of the target class. The distribution of classes within the dataset can be comprehended by a simple yet instructive approach. The provided code builds a bar plot that visually represents the prevalence of target classes or "classes" in a dataset. To address the issue of imbalanced datasets and determine the most appropriate assessment metrics for machine learning models, it is beneficial to gain insight into the distribution of instances across different classes within the dataset.

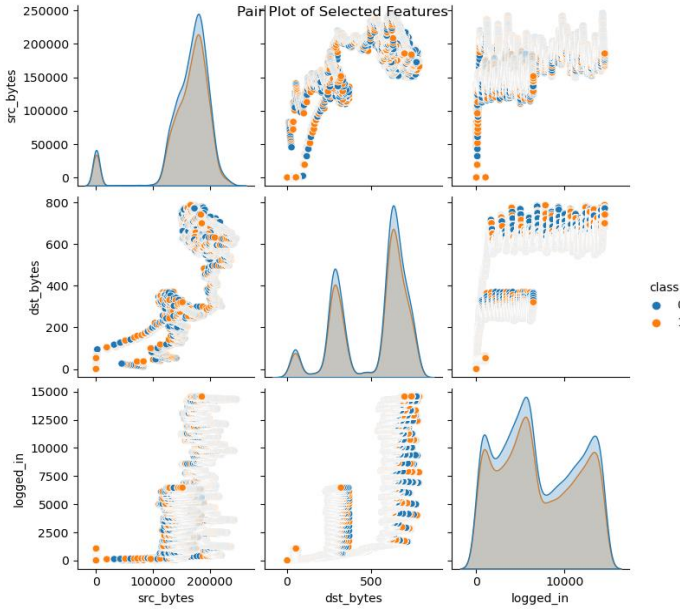


Heatmap: This visual representation may facilitate comprehension of the interrelationships among the different elements within the collection. Through the utilization of heatmap visualization, it becomes possible to identify patterns of feature association. Warmer colours, such as red, are employed to symbolize positive correlation values, indicating that an increase in one attribute is frequently accompanied by a corresponding increase in the other. Cooler hues, such as blue, are employed to visually represent negative correlation values, indicating a tendency for one attribute to decline as the other attribute develops.

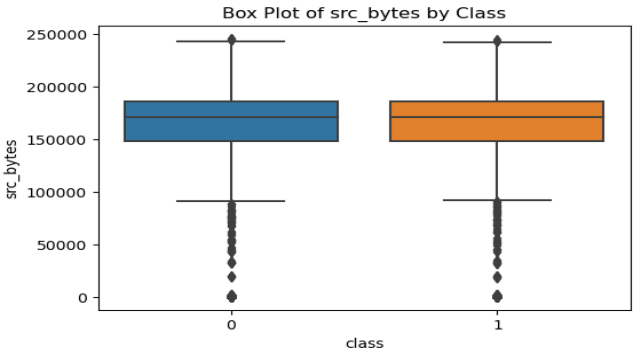


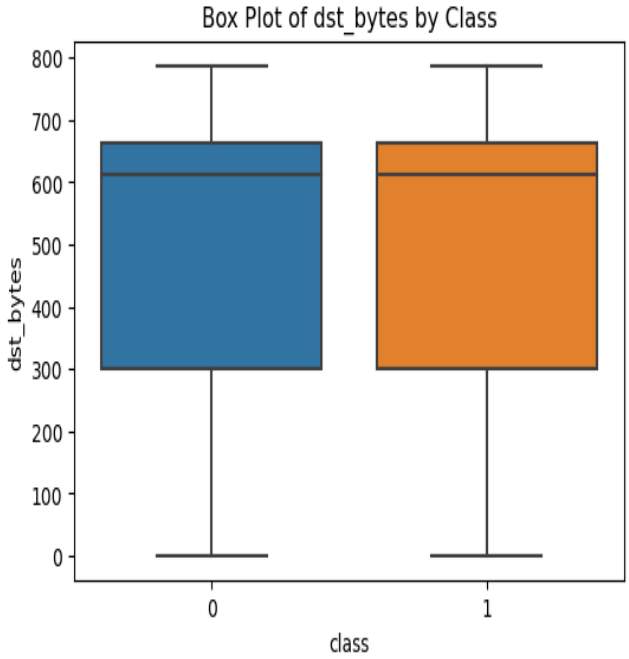
Pair Plot: The task at hand involves generating a pair plot that visualizes specified features, with distinct colours assigned to each class. This is a great way to visualize relationships between pairs of features and how they are

distributed based on the target class. The pair plot exhibits scatter plots between pairs of features in the off-diagonal subplots, while the diagonal subplots showcase the distribution of each feature. The patterns and interactions between attributes within each class can be observed due to the colour-coding of the scatter plot points based on the target class.



Box Plot: Box charts are a useful tool for visualizing the distribution of numerical data. They can also aid in identifying potential outliers or inconsistencies in distribution across different classes. The box plots consist of boxes that enclose the interquartile range (IQR) of the data, with median lines positioned within them. Points that sit outside the "whiskers" of a data distribution could potentially be indicative of outliers. The "whiskers" of the data distribution extend to encompass the minimum and highest values falling within a specified range. The inclusion of the 'class' column in the box plots facilitates the comparison of attribute distributions across different classes.





The scikit-learn StandardScaler is employed to normalize the numerical characteristics inside the dataset. Normalization is a commonly employed preprocessing technique in the field of machine learning, wherein the features of a dataset are rescaled to possess a mean value of 0 and a standard deviation of 1. This can enhance the performance of certain algorithms, particularly those that exhibit sensitivity to the dimensions of features.

After completing this procedure, the standardized characteristics will be incorporated into the variable X_scaled, wherein each characteristic will possess an average value of 0 and a standard deviation of 1. The process of normalizing guarantees that each feature makes an equal contribution to the learning process of the model. This is particularly important for optimization methods such as gradient descent and distance-based algorithms.

The data preprocessing stage involves normalizing the dataset, which is necessary for the implementation of the aforementioned classical machine learning models and obtaining their respective outputs.

Random Forest Classifier.

To commence the discussion on the Random Forest Classifier, we will proceed with the implementation of a Random Forest Classifier utilizing the RandomForestClassifier module from sci-kit-learn. Additionally, we will conduct hyperparameter tuning by employing Grid Search in conjunction with cross-validation.

Creating a Pipeline: Employing a pipeline framework for the construction of a machine learning pipeline. The pipeline consists of two sequential phases.

- **SMOTE** (Synthetic Minority Over-sampling Technique) is used in this stage to balance the class distribution. SMOTE generates

synthetic samples for the minority class to address unbalanced datasets.

- **Classification** This step includes the RandomForestClassifier. While using the **'entropy'** criterion for splitting, **'sqrt'** is the maximum number of features to consider for splitting, and the random state to ensure reproducibility.

1.) **Grid Parameter for n_estimators:**
The hyperparameter you wish to adjust is specified in this dictionary as "classification n estimators," which stands for the number of random forest trees.

2.) **Grid Search CV:**
It is used to Tune the hyperparameters. The pipeline model which was previously created is the estimator parameter. The dictionary of hyperparameters you wish to tweak is specified in param_grid. To maximize recall, the score parameter is set to "recall". cv=5 indicates a 5-fold cross-validation.

The optimal value of the 'n_estimators' hyperparameter for the Random Forest Classifier, considering the utilization of the SMOTE balancing approach, is determined through the implementation of this pipeline and the application of Grid Search. The optimal parameter value will be established by utilizing 5-fold cross-validation and evaluating the recall score.
To assess the efficacy of the optimal model determined by the grid search, the most favourable parameters and corresponding scores are displayed, alongside the feature importance of the superior model. This approach is commonly employed to ascertain the impact of different features on the model's performance and to determine the relative importance of these features in achieving accurate predictions.

Upon the successful implementation of the Random Forest Classifier, the obtained findings are as follows:

- **Best Parameter:** The Random Forest Classifier's optimal parameter, **n estimators, is 100** as discovered using the Grid Search. This indicates that the best model employs an ensemble of 100 decision trees.
- **Best Cross-Validated Score:** The best model had a cross-validated score of **roughly 0.954.** This result shows the cross-validation process's evaluation metric, in this case, was recall.
- The following list, which is arranged in descending order, lists the best model's feature priorities:

flag_sf - 0.210
service_http – 0.154
flag_SO – 0.151

The feature importance provides insight into the relative significance of each feature in impacting the predictions made by the model. The model's predictions are more heavily influenced by attributes that possess greater significance values. This knowledge can be applied in the context of feature selection or engineering, as well as for gaining insights into the factors that influence the judgements made by the model.

Support Vector Classifier:

Implementing a Support Vector Classifier (SVC) using sci-kit-learn's SVC and performing hyperparameter tuning using Grid Search with cross-validation.

1. **Creating a Pipeline:** Applying a pipeline methodology to construct a machine learning pipeline. The pipeline consists of two sequential phases.

- **SMOTE** (Synthetic Minority Over-sampling Technique) is used in this stage to balance the class distribution. SMOTE generates synthetic samples for the minority class to address unbalanced datasets.

- **Classification** This step includes the Random Forest Classifier. While using the **'entropy'** criterion for splitting, **'sqrt'** is the maximum number of features to consider for splitting, and the random state to ensure reproducibility.

2. **Grid parameters for C and the kernel:**

- **Grid_param:** The hyperparameters you want to adjust are listed in this dictionary. Linear, poly, rbf, and sigmoid kernel types are defined by the 'classification__kernel' parameter, while the 'classification__C' parameter specifies the regularization parameter values to be tested.

- **Grid Search CV:** Grid Search CV with the following parameters: estimator = model, grid param = param grid, scoring = recall, CV = 5. Tweaking hyperparameters with Grid Search CV. The pipeline model you previously created is the estimator parameter. The dictionary of hyperparameters you want to tweak is set in the param grid. To maximize recall, the score parameter is set to "recall". cv=5 indicates a 5-fold cross-validation.

The best 'kernel' and 'C' hyperparameter combinations for the SVC are found using this pipeline and Grid Search. Utilizing 5-fold cross-validation, the optimal parameter values will be selected based on the recall score.

After implementing the Support Vector Classifier, we got the following results:

Best Parameters: 'classification C': 1, 'classification kernel': 'rbf' are the best parameters. The Support Vector Classifier's best parameters discovered by the Grid Search are "C": 1 and "kernel": "rbf". Thus, the RBF kernel with regularization parameter C set to 1 is the kernel that the best model employs.

The highest cross-validated rating: The best cross-validated score obtained by the best model **is roughly 0.958, or 0.9583022576983835.** This result shows the cross-validation process's evaluation metric, which in this case was recall.

These findings help to understand the Support Vector Classifier's ideal setting for maximizing recall performance on your dataset. This data can be used to further deploy and test your model.

Xtreme Gradient Boosting:

Preparing a model using the XGBoost classifier and performing hyperparameter tuning using Grid Search for maximizing recall.

1. **Establishing a Pipeline:**

- **Balancing:** This step applies SMOTE (Synthetic Minority Over-sampling Technique) to balance the class distribution.

- **Classification':** The XGBClassifier (XGBoost classifier) is used in this stage. Given that the target is a binary classification problem, you have specified "binary: logistic" as the goal. Use_label_encoder is set to False to prevent label encoding warnings, and eval_metric is set to 'log loss' to track logistic loss during training.

2. **Grid Parameters for n estimators and Learning Rate:**

- **grid_param:** This dictionary specifies the hyperparameters you want to tune.
- **'classification__n_estimators':** Specifies the number of boosting rounds
- **'classification__learning_rate'**: Specifies the learning rate of the XGBoost model.

3. **Grid Search CV:** With the following parameters: estimator = model, grid_param = param_grid, scoring = recall, CV = 5 tweaking hyperparameters with Grid Search CV. The pipeline model previously created is the estimator parameter. The dictionary of hyperparameters to tweak is set in param_grid. To maximize recall, the score parameter is set to "recall". cv=5 indicates a 5-fold cross-validation.
   The ideal setting for the 'n_estimators' and 'learning_rate' hyperparameters for the XGBoost model utilizing this pipeline and Grid Search. Utilizing 5-fold cross-validation, the optimal parameter values will be selected based on the recall score.

Best Parameters: The optimal **learning rate** and **n estimator** values for the XG Boost Classifier discovered by the Grid Search are **0.1** and **130,** respectively. This indicates that the ideal model employs 130 boosting rounds and a learning rate of 0.1.

Best Cross-Validated Score: The best cross-validated score that the best model could obtain was **0.959374155241146, or roughly 0.959.** This result shows the cross-validation process's evaluation metric, which in this case was recall.

These outcomes help you understand the XGBoost Classifier's ideal setting for maximizing recall performance on your dataset. You can install and further assess your XGBoost model using this data.

Challenges and Data Analysis's Learnings

The study of data also exposes the challenges associated with anomaly-based network intrusion detection. An issue develops due to imbalanced data, characterized by a significant disparity between normal and abnormal behaviour. The presence of this imbalance has the potential to result in false positives and impede the identification of infrequent anomalies (3,4). The utilization of data analysis techniques facilitates the formulation of strategies to address this issue, including the implementation of specialized algorithms to effectively manage datasets with imbalances.

Moreover, the dynamic nature of the networks introduces a level of complexity. The utilization of data analysis enables the differentiation between legitimate modifications and detrimental actions[4]. The adaptive nature of the intrusion detection system enables it to effectively respond to fluctuating network conditions, while concurrently mitigating the risk of false alarms through meticulous analysis of long-term trends and patterns.

Results and Interpretation:

The presence of imbalanced class distributions in machine learning tasks poses significant challenges as it leads to a biased performance of the models. In this study, the objective was to address the aforementioned issue by the implementation of hyperparameter tuning techniques on three distinct classifiers, including Random Forest, Support Vector Classifier (SVC), and XG Boost. The optimization of recall, a significant metric in scenarios where minimizing false negatives is of utmost importance, is necessary.

The optimization of hyperparameters can significantly improve the efficacy of machine learning models. The selection of appropriate hyperparameters plays a crucial role in determining the efficacy of a model in reliably identifying positive instances inside binary classification tasks, particularly when dealing with imbalanced datasets. This study investigates the effects of hyperparameter tuning on three widely recognized classification algorithms: Random Forest, Support Vector Classifier (SVC), and XGBoost. In scenarios where the prevention of false negatives holds significant importance, such as in anomaly detection or medical diagnosis, the primary focus lies on optimizing the metric of recall, which is particularly relevant in these contexts.

This study included three widely recognized algorithms, namely Random Forest, Support Vector Classifier (SVC), and XGBoost. The primary objective of implementing hyperparameter optimization using Grid Search was to enhance the recall metric. The concept of recall is a statistical measure that highlights the model's ability to correctly identify all instances of positive outcomes, hence minimizing the occurrence of false negatives.

Random Forest Classifier: Our investigation commenced with an examination of the Random Forest Classifier. The consequences of altering the number of estimators, which represents the number of decision trees in the ensemble, were meticulously evaluated through the utilization of the Grid Search methodology. The optimal design was determined to be the utilization of 100 estimators. This underscores the advantage of employing an ensemble approach, which enhances the resilience of the model by incorporating many decision trees. Significantly, the recall score for this optimal configuration exhibited a commendable value of approximately 0.958. It was observed that the attribute 'flag_SF' exhibited the greatest weight, hence indicating its notable role in delineating class differentiation. The features 'service_http' and 'flag_S0', together with other relevant elements, played a critical role in differentiating attacks from normal network traffic. This underscores the significance of network attributes.
A grid search is conducted to get the optimal value for the number of trees, denoted as "n_estimators," in the Random Forest algorithm. Cross-validation is a widely employed technique for assessing the performance of a model. In this study, the 'recall' score was chosen as the major performance indicator. The model that exhibited the highest performance was determined to be the one employing 100 trees, with a recall score of approximately 0.954.

Support Vector Classifier (SVC): Subsequently, our attention was redirected towards the Support Vector Classifier (SVC). Our objective was to determine the optimal configuration that would optimize recall. To do this, we experimented with different kernel functions, including linear, polynomial, radial basis function, and sigmoid. After conducting an extensive Grid Search, it was determined that the most favourable arrangement involved utilizing an RBF kernel alongside a regularization value of 1. The experimental configuration yielded a notable recall score, approximately 0.958.
The utilization of the Radial Basis Function (RBF) kernel demonstrated the model's capacity to proficiently capture intricate nonlinear associations inside the dataset. The findings suggest that the support vector classifier (SVC) demonstrates proficiency in accurately detecting complex patterns within imbalanced datasets. The primary emphasis of our Support Vector Classifier (SVC) hyperparameter optimization lies in the 'kernel' and 'C' parameters. The determination of the decision boundary is heavily reliant on the regularization parameter and kernel function. The recall score continues to guide our optimization approach. The Support Vector Classifier (SVC) was utilized to obtain a model with a recall score of around 0.958. This was achieved by combining a radial basis function ('rbf') kernel with a regularization parameter ('C') of 1 in an optimal manner.

XGBoost Classifier: Our third subject of investigation was the XGBoost Classifier, a highly effective and efficient gradient-boosting algorithm that is widely recognized for its superior performance. In this study, we optimized the learning rate and the number of boosting rounds (referred to as 'n_estimators'). The optimal model configuration was determined through the utilization of Grid Search. This configuration entailed 130 boosting rounds and a learning rate of 0.1. Significantly, this amalgamation outperformed the preceding models, achieving a recall score of approximately 0.959. This exemplifies the efficacy of XGBoost in mitigating imbalanced datasets due to its boosting mechanism that prioritizes challenging instances for classification, hence enhancing recall.

XGBoost has demonstrated a robust history of achievement across a diverse range of machine-learning endeavours. The primary emphasis of the model is to optimize the 'n_estimators' and 'learning_rate' parameters within this particular technique. While the latter factor affects the magnitude of the step taken during the optimization process, the former factor determines the total number of boosting rounds that will be performed. The primary aim remains to enhance the level of "recall."

- The XGBoost Classifier obtained the highest recall score of roughly 0.959 with a learning rate of 0.1 and 130 boosting rounds.

The findings underscore the need to optimize hyperparameters to enhance recall performance, especially in datasets characterized by imbalanced class distributions. The recall performance of the Random Forest, Support Vector Classifier (SVC), and Extreme Gradient Boosting

(XGBoost) algorithms exhibited enhancement upon adjustment for the distinctive characteristics inherent in the dataset. Developing dependable classifiers for real-world applications is necessitated because the choice of appropriate hyperparameters can significantly influence the model's capacity to effectively identify positive instances.

CONCLUSION:

In conclusion, anomaly-based network intrusion detection plays a key role in identifying and mitigating security vulnerabilities within computer networks. This particular intrusion detection system employs machine learning techniques to identify deviations from normal network behaviour that could potentially indicate security breaches. In contrast to conventional signature-based systems, anomaly-based intrusion detection systems provide the capability to detect and respond to previously unrecognized attacks, while also adapting to dynamic changes in network conditions. Nevertheless, these systems possess certain limitations, such as the potential for generating false positive results and the necessity for continuous surveillance and enhancement. Enterprises must prioritize the implementation of anomaly-based network intrusion detection to protect their critical data and assets. In the realm of comprehensive network security strategies, it holds paramount importance.

We examined the complexities of three potent algorithms, Random Forest, Support Vector Classifier (SVC), and XGBoost, in our thorough investigation of imbalanced classification models. We sought to determine the top-performing model for deployment in the real world by focusing on recall optimization, a critical parameter in situations where reducing false negatives is essential. We may reliably make judgments regarding the best model option for handling imbalanced datasets after carefully examining the outcomes.

Upon doing a comparative analysis of the performances of the three models, it is evident that all of them had commendable recall scores, hence indicating their efficacy in the detection of abnormal situations. Although the recall scores for the Random Forest and SVC models were similar, the XGBoost model demonstrated somewhat superior performance, achieving a slightly higher score. When selecting the optimal model, it is imperative to consider the computing complexity and deployment requirements.

The Random Forest Classifier is a noteworthy contender in practical scenarios where computing efficiency and interpretability of the model are of utmost importance. Given its ensemble nature, which effectively balances runtime efficiency and forecast accuracy, this approach is well-suited for broad implementation. Moreover, the feature importance analysis of the Random Forest provides network managers with insights into the network features that have the greatest impact on attack detection.

In contrast, the **Support Vector Classifier** utilizing the Radial Basis Function (RBF) kernel demonstrates exceptional performance in scenarios where intricate patterns and the capacity to handle nonlinearity are of utmost importance. This model demonstrates exceptional proficiency in detecting complicated connections that may be overlooked by alternative methods.

The **XGBoost Classifier** has the highest recall score in scenarios where prioritizing predictive capability is paramount and a slight increase in computational complexity is tolerable. The utilization of a boosting strategy enhances the ability of the system to accurately identify instances belonging to the minority class, hence rendering it a valuable tool for detecting rare and anomalous events.

The matter of uneven classification lacks a universally applicable solution. When selecting a model, it is crucial to evaluate the specific requirements of the work, the computational resources at hand, and the relative significance of interpretability compared to performance. The efficacy of the Random Forest, Support Vector Classifier (SVC), and Extreme Gradient Boosting (XGBoost) models in addressing imbalanced datasets and optimizing recall has been empirically established. To achieve alignment between the selected model and the operational limitations and objectives of the real-world deployment scenario, it is imperative to conduct a comprehensive assessment of these factors before selecting the most suitable model.

This paper presents a comprehensive framework for addressing the challenges associated with imbalanced categorization, in summary. This study investigates the benefits and limitations of three robust algorithms, providing practitioners with valuable insights to facilitate informed decision-making.

Future Work:

While the present work has provided insights into enhancing the performance of imbalanced classification models in terms of recall, there remain various avenues that warrant further investigation in subsequent research endeavours. The subsequent domains present opportunities to enhance the breadth and depth of this study:

1.) **Ensemble of Models:** One potential avenue of inquiry involves examining the interrelationships among the highest-performing models. The integration of multiple classifiers, such as Random Forest, SVC, and XGBoost, has the potential to yield an ensemble model that exhibits enhanced performance across diverse contexts. The utilization

of the diverse range of models can be leveraged to enhance the overall predictive capability by employing techniques such as stacking or voting.

2.) **Feature Engineering:** While the present study primarily focused on model selection and hyperparameter tuning, it is worth noting that feature engineering could be a potential area of investigation in future research endeavours. Exploring feature transformations, aggregations, or interactions specific to a certain domain has the potential to unveil latent information and enhance the performance of a model. The application of dimensionality reduction techniques in feature engineering can mitigate the impact of high dimensionality and potentially improve the efficacy of models.

3.) **Advanced hyperparameter tuning:** By employing advanced techniques such as Bayesian optimization or evolutionary algorithms, the process of hyperparameter tuning can be further enhanced. These strategies have the potential to reduce iteration time and computational resources while improving the efficiency of exploring the hyperparameter space and identifying optimal configurations.

4.) **Handling Class Imbalance:** While SMOTE was employed in this research to address the issue of class imbalance, exploring alternative resampling techniques such as ADASYN or Borderline-SMOTE could yield a more comprehensive comprehension of their impact on the performance of the model. Additional investigation into methodologies like as cost-sensitive learning or the development of tailored loss functions specific to the given problem might yield models that exhibit a heightened alignment with the requirements of real-world implementations.

5.) **Exploring Additional Evaluation Metrics:** The primary evaluation criterion in this work was recall, although other measures including accuracy, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) can provide a more comprehensive assessment of model performance. A more comprehensive comprehension of a model's behaviour can be attained through the comparison of models using their Receiver Operating Characteristic (ROC) curves or through the analysis of the trade-offs between precision and recall.

6.) **Explainability and Interpretability:** Ensuring model transparency and interpretability is of utmost importance due to the growing utilization of machine learning models in critical domains. Subsequent investigations could employ other strategies, such as SHAP (Shapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations), to offer elucidations for model predictions. These approaches can contribute to a better understanding of the underlying rationale behind a particular prediction.

In summary, this study lays the foundation for other prospective avenues of future research. Researchers have the potential to enhance and broaden the efficacy of unbalanced classification models by the integration of diverse classifiers, more exploration of feature engineering, the use of advanced hyperparameter tuning approaches, and addressing the challenges posed by real-world data. Ultimately, the ongoing inquiry will facilitate a more profound understanding of model behaviour, optimize its application in real-world scenarios, and improve decision-making processes about network security and anomaly detection.

Bibliography

1. Author links open overlay panel, JonathanJ. Davis *et al.* (2011) *Data preprocessing for anomaly-based Network Intrusion Detection: A Review*, *Computers & Security*. Elsevier Advanced Technology.
2. Author links open the overlay panel. García-Teodoro *et al.* (2008) *Anomaly-based network intrusion detection: Techniques, systems and challenges*, *Computers & Security*. Elsevier Advanced Technology.
3. Author links open overlay panel eyed Mojtaba Hosseini Bamakan a b *et al.* (2016) *An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization*, *Neurocomputing*. Elsevier. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0925231216300510
4. Author links open overlay panel Animesh Patcha *et al.* (2007) *An overview of anomaly detection techniques: Existing solutions and latest technological trends*, *Computer Networks*. Elsevier. Available at: https://www.sciencedirect.com/science/article/abs/pii/S138912860700062X
5. Author links open overlay panelNour Moustafa *et al.* (2018) *A holistic review of Network Anomaly Detection Systems: A comprehensive survey*, *Journal of Network and Computer Applications*. Academic Press. Available at: https://www.sciencedirect.com/science/article/abs/pii/S1084804518303886
6. Scarfone, K.A. and Mell, P.M. (2021) *Guide to intrusion detection and prevention systems (IDPs), NIST*. KarenA. Scarfone, Peter M. Mell. Available at: https://www.nist.gov/publications/guide-intrusion-detection-and-prevention-systems-idps
7. Author links open overlay panel Jonathan J. Davis *et al.* (2011) *Data preprocessing for anomaly-based Network Intrusion Detection: A Review*, *Computers & Security*. Elsevier Advanced Technology.
8. *Computer Networks* (no date) *Computer Networks | Journal | ScienceDirect.com by Elsevier*. Available at: https://www.sciencedirect.com/journal/computer-networks (Accessed: April 24, 2023).
9. Tech, R.M.V., *et al.* (2014) *A survey of intrusion detection techniques for cyber-physical systems*, *ACM Computing Surveys*. Available at: https://dl.acm.org/doi/10.1145/2542049
10. Alshammari, A. and Aldribi, A. (2021) *Apply machine learning techniques to detect malicious network traffic in cloud computing - Journal of big data, SpringerOpen*. Springer International Publishing. Available at: https://journalofbigdata.springeropen.com
11. University, Y.L.W., *et al.* (2022) *Deep learning-based anomaly detection in cyber-physical systems: Progress and opportunities: ACM Computing Surveys: Vol 54, no 5, ACM Computing Surveys*. Available at: https://dl.acm.org/doi/10.1145/3453155
12. Alshammari, A. and Aldribi, A. (2021) *Apply machine learning techniques to detect malicious network traffic in cloud computing - Journal of big data, SpringerOpen*. Springer International Publishing. Available at: https://journalofbigdata.springeropen.com
13. Author links open overlay panelSunanda Gamage *et al.* (2020) *Deep Learning Methods in Network Intrusion Detection: A Survey and an Objective Comparison*, *Journal of Network and Computer Applications*. Academic Press.
14. Mirsky, Y. *et al.* (2018) *Kitsune: An ensemble of autoencoders for online network intrusion detection*, *arXiv.org*.
15. Author links open overlay panelJonathan J. Davis a et al. (2011) Data preprocessing for anomaly based Network Intrusion Detection: A Review, Computers &
16. P. García-Teodoro Department of Signal Theory et al. (1970) Anomaly-based network intrusion detection: Techniques, systems and challenges: Computers and security: Vol 28, no 1-2, Computers and Security
17. Author links open overlay panelAnimesh Patcha et al. (2007) An overview of anomaly detection techniques: Existing solutions and latest technological trends, Computer Networks. Available at: https://www.sciencedirect.com/science/article/pii/S138912860700062X
18. Author links open overlay panelNour Moustafa a et al. (2018) A holistic review of Network Anomaly Detection Systems: A comprehensive survey, Journal of Network and Computer Applications. Available at: https://www.sciencedirect.com/science/article/pii/S1084804518303886
19. Guide to intrusion detection and prevention

systems (idps) - NIST. Available at: https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-94.pdf
20. Using convolutional neural networks to network intrusion detection for cyber threats [WWW Document], n.d. . Using convolutional neural networks to network intrusion detection for cyber threats | IEEE Conference Publication | IEEE Xplore. URL https://ieeexplore.ieee.org/abstract/document/8394474
21. Intrusion Detection using Artificial Neural Network [WWW Document], n.d. . Intrusion Detection using Artificial Neural Network | IEEE Conference Publication | IEEE Xplore. URL https://ieeexplore.ieee.org/abstract/document/5592568
22. A hybrid system for reducing the false alarm rate of anomaly intrusion detection system [WWW Document], n.d. . A hybrid system for reducing the false alarm rate of anomaly intrusion detection system | IEEE Conference Publication | IEEE Xplore. URL https://ieeexplore.ieee.org/abstract/document/6194493
23. A Comparison of Intrusion Detection by K-Means and Fuzzy C-Means Clustering Algorithm Over the NSL-KDD Dataset [WWW Document], n.d. . A Comparison of Intrusion Detection by K-Means and Fuzzy C-Means Clustering Algorithm Over the NSL-KDD Dataset | IEEE Conference Publication | IEEE Xplore. URL https://ieeexplore.ieee.org/abstract/document/8524401
24. Intrusion Detection System Using PCA with Random Forest Approach [WWW Document], n.d. . Intrusion Detection System Using PCA with Random Forest Approach | IEEE Conference Publication | IEEE Xplore. URL https://ieeexplore.ieee.org/abstract/document/9155656
25. An effective intrusion detection framework based on SVM with feature augmentation [WWW Document], 2017. . An effective intrusion detection framework based on SVM with feature augmentation - ScienceDirect. https://doi.org/10.1016/j.knosys.2017.09.014
26. Kim, D.S., Park, J.S., n.d. Network-Based Intrusion Detection with Support Vector Machines [WWW Document]. Network-Based Intrusion Detection with Support Vector Machines | SpringerLink. https://doi.org/10.1007/978-3-540-45235-5_73
27. Hindawi, Peng, K., M. Leung, V.C., Zheng, L., Wang, S., Huang, C., Lin, T., 2018. Intrusion Detection System Based on Decision Tree over Big Data in Fog Environment [WWW Document]. Intrusion Detection System Based on Decision Tree over Big Data in Fog Environment. https://doi.org/https://doi.org/10.1155/2018/4680867
28. A novel statistical technique for intrusion detection systems [WWW Document], 2017. . A novel statistical technique for intrusion detection systems - ScienceDirect. https://doi.org/10.1016/j.future.2017.01.029
29. Tianfield, H., 2017. Data mining based cyber-attack detection [WWW Document]. ResearchOnline. URL https://researchonline.gcu.ac.uk/en/publications/data-mining-based-cyber-attack-detection
30. Supervised and Unsupervised Learning for Data Science [WWW Document], n.d. . SpringerLink. URL https://link.springer.com/book/10.1007/978-3-030-22475-2
31. Deep learning in intrusion detection perspective: Overview and further challenges [WWW Document], n.d. . Deep learning in intrusion detection perspective: Overview and further challenges | IEEE Conference Publication | IEEE Xplore. URL https://ieeexplore.ieee.org/document/8275095
32. Binbusayyis, A., Alaskar, H., Vaiyapuri, T., Dinesh, M., 2022. An investigation and comparison of machine learning

approaches for intrusion detection in IoMT network - The Journal of Supercomputing [WWW Document]. SpringerLink. https://doi.org/10.1007/s11227-022-04568-3
33. Assegie, T.A., 2021. An Optimized KNN Model for Signature-Based Malware Detection [WWW Document]. An Optimized KNN Model for Signature-Based Malware Detection by Tsehay Admassu Assegie :: SSRN. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3814215
34. A comprehensive review of AI based intrusion detection system [WWW Document], 2023. . A comprehensive review of AI based intrusion detection system - ScienceDirect. https://doi.org/10.1016/j.measen.2023.100827
35. The 1999 DARPA off-line intrusion detection evaluation [WWW Document], 2000. . The 1999 DARPA off-line intrusion detection evaluation - ScienceDirect. https://doi.org/10.1016/S1389-1286(00)00139-0
36. Use of K-Nearest Neighbor classifier for intrusion detection [WWW Document], 2002. . Use of K-Nearest Neighbor classifier for intrusion detection - ScienceDirect. https://doi.org/10.1016/S0167-4048(02)00514-X
37. Anomaly detection methods in wired networks: a survey and taxonomy [WWW Document], 2004. . Anomaly detection methods in wired networks: a survey and taxonomy - ScienceDirect. https://doi.org/10.1016/j.comcom.2004.07.002
38. MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING - A REVIEW: Discovery Service for DBS [WWW Document], n.d. . MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING - A REVIEW: Discovery Service for DBS. URL https://eds.p.ebscohost.com/eds/detail/detail?vid=0&sid=0a342b73-8432-4f59-829a-5ee8a99bd4a2%40redis&bdata=JkF1dGhUeXBlPWlwLHNoaWIsY29va2llLHVybCZzaXRlPWVkcy1saXZl#AN=130548851&db=iih
39. A hybrid deep learning model for efficient intrusion detection in big data environment [WWW Document], 2019. . A hybrid deep learning model for efficient intrusion detection in big data environment - ScienceDirect. https://doi.org/10.1016/j.ins.2019.10.069
40. Mining network data for intrusion detection through combining SVMs with ant colony networks [WWW Document], 2013. . Mining network data for intrusion detection through combining SVMs with ant colony networks - ScienceDirect. https://doi.org/10.1016/j.future.2013.06.027