Tampere University

Wajeh Ahmad

ARTIFICIAL INTELLIGENCE IN MODERN FIREWALLS

Role of Artificial Intelligence in Network Security

Masters' Thesis
Faculty of Information Technology
and Communication Sciences
Juha Vihervaara
Matti Monnonen
October 2024

ABSTRACT

Wajeh Ahmad: Artificial Intelligence in Modern Firewalls, Role of Artificial Intelligence in Network Security
Master's Thesis
Tampere University
Masters' of AI and Robotics
August 2024

This thesis explores the role and impact of Artificial Intelligence on the design of modern firewalls due to the advancement of Artificial Intelligence (AI), particularly how AI-based methods transcend the barriers of the traditional rule-based approaches and enhance overall network security. The research traces the evolution of firewalls over time and highlights the necessity for more advanced security measures. It also outlines the progress of AI, setting the stage for a deep dive into the role of AI in security systems.

The research critically assesses the limitation of conventional firewalls, such as their inability to detect zero-day attacks and their vulnerability to inspect encrypted traffic. It is investigated that how machine learning techniques, such as shallow neural networks or decision trees, can complement network security by aiding the firewall's decision making and automatically classifying network packets. Furthermore, the thesis demonstrates the application of intrusion detection systems in cyberspace by employing machine learning, normal distribution, clustering, and t-plotting that can facilitate prevention and/or detection of such malicious attacks.

In addition, the thesis investigates the suitability of Explainable AI (XAI) to explain and make understandable AI based firewall decisions. It also highlights the difficulties and constraints of AI-centric firewalls, like the fact that they require a large set of data, the need of removing biases from such data, as well as taking privacy into account when using these capabilities, pushing further research on development of robust training methodologies and privacy-preserving approaches. The thesis features two prominent case studies: Google Chronicle Security Operations Platform, and Darktrace Enterprise Immune System. These case-studies highlight AI can be put to practical use for improving the abilities of threat detection, investigation and response.

The thesis concludes by highlighting the significant progress made in integrating AI into contemporary firewalls, such that they can start learning proactively against continuous adapting cyber threat landscape. The report concludes by suggesting for more research to be done in the refining of AI algorithms, the investigation into reinforcement learning, and adding user and entity behavior analytics (UEBA) onto models as well as continuing the process of creating explainable AI models. The findings from this work will contribute to a greater understanding of AI-powered network security and offer direction for those looking at how they might develop more resilient firewalls in the future both in practice as well as in research.

Keywords: Artificial Intelligence, Machine Learning, Neural Networks, Deep Learning, Explainable AI, Clustering, t-plotting, Intrusion Detection System.

USE OF AI IN THESIS

I have utilised AI tools in my thesis:

☐ No
✔ Yes

The AI tools utilised in my thesis and their purposes are described below:

Names and versions of AI tools:

- OpenAI ChatGPT (GPT 4.0): Used for generating ideas, providing information on topics related to network security.
- Google Gemini (1.0): Use to get holistic picture about few topics

Purpose of using AI tools:

- Assisted in brainstorming ideas and structuring sections of the thesis
- Enhance the readability and professionalism of the thesis

Sections where AI tools were used:

- 2.5.5. Challenges and Limitations of AI-powered Firewalls: A Critical Examination
- 4. AI Ethics in Security
- 6. Conclusion
- Few other minor places

I acknowledge that I am fully responsible for the entire content of my thesis, including the parts generated by AI, and accept accountability for any violations of ethical standards in publications.

PREFACE

This thesis would not have been possible without the usual guidance and support of numerous individuals. First and foremost, I extend my gratitude to my esteemed professor, Juha Vihervaara, whose support and encouragement have been invaluable throughout this endeavor. Your insightful feedback and subject knowledge have been unwavering and played a key role in pushing me to complete my research.

I would like to thank my beloved mother, Samina Khan, I owe an immeasurable debt of gratitude. Your unwavering love, patience, sacrifices, and constant prayers have been my guiding light, motivating me to strive for excellence not just in my academic journey but in all aspects of life. Your steadfast support has always been the foundation of all my accomplishments and achievements.

To my dear wife, Arooba, thank you for your love, patience, and understanding during the countless hours I dedicated to this thesis. Your unwavering support and encouragement have been my constant source of strength and inspiration.

I would also like to express my appreciation to my friends and colleagues at Tampere University for their camaraderie and support. Your stimulating discussions and collaborative spirit have enriched my academic experience and fostered an environment conducive to growth and learning.

Finally, I extend my gratitude to all those who have contributed to my academic and personal development, directly or indirectly. Your collective influence has shaped me into the person I am today, and I am deeply grateful for your presence in my life.

Thank you all from the bottom of my heart.

Wajeh Ahmad

Tampere, Finland

October 2024

Table of Contents

1. Introduction 10
   1.1 Background and Motivation for Thesis 10
   1.2 Research Objective and Scope 11
2. Literature Review 13
   2.1 Network Components and Management 13
   2.2 Need for Network Security and Firewalls 17
   2.3 Traditional Firewall Technologies and Evolution 18
   2.4 Evolution of Artificial Intelligence 20
   2.5 Role of Artificial Intelligence and Machine Learning in Firewall 24
      2.5.1 Limitations of Traditional Firewall 24
      2.5.2 Use of Machine Learning Techniques to Identify Firewall Decisions 26
      2.5.3 Machine Learning Techniques for Intrusion Detection Systems 28
      2.5.4 Explainable AI (XAI) and the Future of Cloud Security 32
      2.5.5 Challenges and Limitations of AI-powered Firewalls: A Critical Examination 33
   2.6 AI in Network Segmentation and Microsegmentation 35
3. Case Studies and Real-World Examples 38
   3.1 Google Chronicle's Security Operations Platform 38
   3.2 Darktrace's Enterprise Immune System 41
4. AI Ethics in Security 44
5. Regulatory and Compliance Challenges in AI-Powered Firewalls 48
   5.1 Compliance with Data Protection Laws 48
   5.2 Compliance with Industry-Specific Standards 49
   5.3 Future Regulatory Trends 50
6. Conclusion 52
   6.1 Future Directions for AI in Firewalls 53
   6.2 Recommendations 54
7. References 55

LIST OF FIGURES

Figure 1 - TCP/IP with OSI model 13

Figure 2 - Software Defined Network (SDN) Controller 16

Figure 3 - Application firewall 19

Figure 4 - Subset of AI 21

Figure 5 - Demonstration of Supervised learning 22

Figure 6 - Example of Supervised learning.....23

Figure 7 - Example of Reinforcement learning 24

Figure 8 - Packet filtering firewall preventing spoofing attack 25

Figure 9 - System Evaluation: Overall evaluation metrics and Confusion matrix (Al-Haijaa & Ishtaiwia, 2021) (Al-Haijaa & Ishtaiwia, 2021) 27

Figure 10 - Google Security Operations platform overview 38

LIST OF TABLES

Table 1 - Comparison of Machine Learning Techniques for IDS 29

Table 2 - Comparison of Deep Learning Techniques for IDS 31

Table 3 - Comparative summary of the main XAI techniques for cybersecurity.....33

LIST OF SYMBOLS AND ABBREVIATIONS

ACC	Classification Accuracy
AI	Artificial Intelligence
ANN	Artificial Neural Network
APT	Advanced Persistent Threats
ASIC	Application-Specific Integrated Circuit
CNN	Convolutional Neural Network
CEP	Classification Error Percent
DBN	Deep Belief Network
DL	Deep Learning
DNN	Deep Neural Network
DNS	Domain Name System
DoS	Denial-of-Service
DT	Decision Tree
FDR	False Discovery Rate
FN	False Negatives
FNR	False Negative Rate
FP	False Positives
GAN	Generative Adversarial Network
ICMP	Internet Control Message Protocol
IDS	Intrusion Detection System
IDP	Intrusion Detection and Prevention
IFW	Internet Firewall
IP	Internet Protocol
IoT	Internet of Things
LIME	Local Interpretable Model-Agnostic Explanations
LSTM	Long Short-Term Memory
ML	Machine Learning
MTTD	Mean Time to Detect

NGFW	Next-Generation Firewall
NLP	Natural Language Processing
NV	Network Virtualization
ODT	Optimizable Decision Tree
OSI	Open Systems Interconnection
PCA	Principle Component Analysis
PPML	Privacy-Preserving Machine Learning
PPV	Positive Predictive Value
RBM	Restricted Boltzmann Machine
RNN	Recurrent Neural Network
SDN	Software-Defined Networking
SHAP	SHapley Additive exPlanations
SNA	Software Network Analysis
SNMP	Simple Network Management Protocol
SNN	Shallow Neural Network
SVD	Singular Value Decomposition
SVR	Support Vector Regression
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TDIR	Threat Detection, Investigation, and Response
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
UDP	User Datagram Protocol
VNF	Virtualized Network Function
VPN	Virtual Private Network
XAI	Explainable AI
YARA-L	Yet Another Recursive Acronym for Logs

1. Introduction

1.1 Background and Motivation for Thesis

The sophistication of cyber threats is growing faster than what conventional firewalls can handle. There is a new technology that could be the answer — Artificial Intelligence (AI) has been making strides to aid in the efficiency of security systems and lighten the workload for human professionals. Using ML, DL or NLP algorithms, AI-powered firewalls can also analyze network traffic in real-time and identify any anomalous behavior then takes proactive actions to deal with emerging threats.

Machine Learning (ML) is a subset of AI that makes computer systems capable of performing tasks without being explicitly programmed. Deep Learning (DL) is another advanced form of AI that uses neural networks (NN) to learn from large datasets. One more popular AI method is natural language processing (NLP) that allows computer systems to analyze and processes human language, containing text and speech.

AI-embedded firewalls let organizations leverage the power of AI and real-time analysis by providing advantages such as high threat detection, response accuracy, rapid productivity, virtual capability scaling that allows better network security while decreasing the load on security professionals. However, AI firewalls rely on large number of samples to predict cyber threats, risk of false positives and violate user privacy.

This thesis aims to have a comprehensive study on the benefits and challenges faced by AI implemented in modern-day firewalls. We will discuss various AI methods for firewalls and analyze those different approaches to see their capabilities as well as shortcomings. We will examine some use cases about AI firewalls; try to evaluate how effective they are for preventing cyber threats at large. Finally, we will provide future directions and recommendations for AI-powered firewalls based on our analysis and findings.

1.2 Research Objective and Scope

This research aims to find out how AI is/can be integrated into contemporary firewalls for securing network traffic. The research studies the various AI approaches, such as ML, DL and NLP to determine which approach is more efficient for real-time cyber threat detection and to avoid cyber threats in real-time. The research will discuss the effectiveness and limitations of using AI-powered firewalls, including their capability to assess extremely large data volumes without delay or inaccuracy, how we can reduce our dependency on human security professionals, being able to preemptively respond against imminent threats. Additionally, the research will also examine the difficulties in practical implementation of these modern firewalls such as the need for large, labeled data sets, how to reduce potential for false positives, and privacy concerns. The study will also analyze real-world examples of AI-powered firewalls and assess their effectiveness in protecting against different types of cyber threats. This will involve evaluating their performance through different parameters, such as enhanced detection, accelerated investigation and proactiveness against different types of attacks. In conclusion, the research's goal is to include learning better AI models and methods for incorporating AI into contemporary firewalls favoring network security, as well as providing recommendations and future directions for AI-driven firewalls.

To achieve this objective, the study will follow a comprehensive review of existing literature on network elements, firewalls technologies, AI methodologies, cybersecurity landscape coupled with empirical analysis of some real-world case studies and performance of AI-based security components. This will help us develop a thorough understanding of theoretical concepts and practical applications of AI in network security. The findings will be presented in a clear and concise manner, utilizing visual aids and statistical analysis where appropriate, to facilitate comprehension and knowledge dissemination.

This thesis will focus on the following topics:

- A comprehensive study of the challenges faced by conventional firewalls in protecting networks against modern cyber threats and how AI-based firewalls can provide a solution.
- The evolution of AI, how it is reshaping the technological sphere especially in terms of cybersecurity

    - An overview of AI techniques that are being used in modern firewalls: ML, DL, NLP and how much these technologies contribute to traffic security
- A detailed analysis of the challenges and limitations of existing methods for creating AI-based firewalls including the need for labeled data, false positives, as well as privacy concerns (for the end user).

This thesis will not cover the following topics:

- Implementation details of specific AI algorithms or techniques.
- Detailed technical specifications of specific firewall products or vendors.
- Legal or regulatory issues related to the use of AI-powered firewalls.

2. Literature Review

2.1 Network Components and Management

A computer network facilitates the connection of multiple computers to enable them to communicate, serve data and share resources. A network architecture consists of a series of building blocks that are required to build communication. This largely depends on the kind of network one is looking to implement, and some elements are optional or non-existent in certain cases such as use of cables for a wireless network.

In computer networks, protocols are a way of communication, which provides a set of rules to be followed for defining network communication. The Open Systems Interconnection (OSI) model was introduced in 1983 and is a way of thinking about the layered architecture of networking protocols, and how they relate to each other. There are seven different layers in the OSI model, and each one does something a little bit different from the other, all the way down to layer 1 where your data gets transmitted physically to through to application-specific interactions. This provides the framework on which common protocols like transmission control protocol (TCP), user datagram protocol (UDP) and internet protocol (IP) work to make communication over networks seamless.

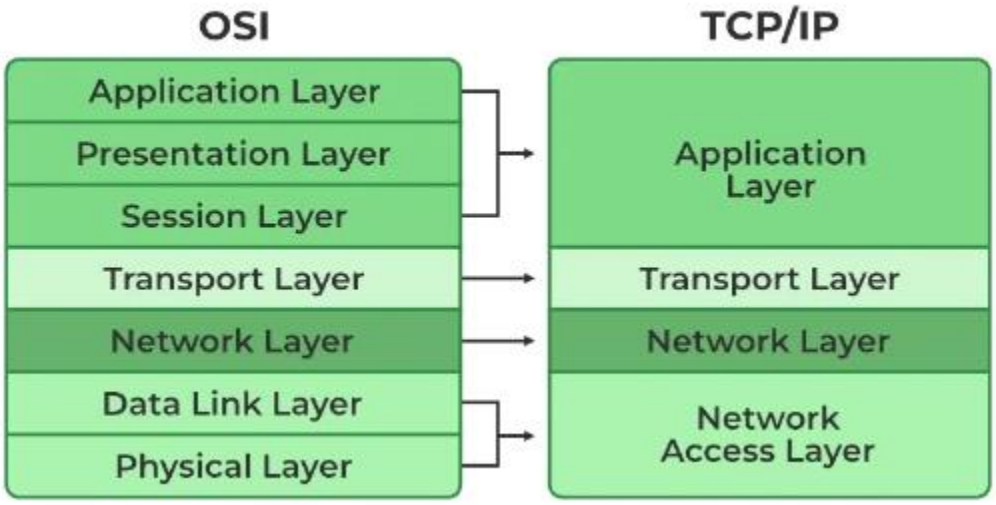


Figure 1 - TCP/IP with OSI model

Figure 1 shows TCP/IP protocol stack, a practical implementation recognized in the OSI model comprises four layers Link Layer Network Layer Transport Layer Application Layer Every layer uses its own protocols tailored to its function. For instance, the Application Layer uses protocols such as HTTP and HTTPS to stipulate how data should be represented in a web browser. TCP and UDP, who are responsible for delivering data reliably or connectionless-ly at the Transport Layer. On the other hand, IP protocol of Network Layer deals with addressing and routing of data packets across networks, while Link Layer works on physical transmission of data over the underlying network medium.

The OSI model depicts a comprehensive view of network communications, and each layer plays a vital role in facilitating communication between devices. Such design architecture contributes to modular, flexible and interoperable building blocks that encourage multiple networking technologies implementation and deployment.

Traditional Network Components

Traditional networking works off a foundation of specialized hardware - such as switches and routers - and each component plays an integral part in the overall network functionality. These devices, which often come in the form of dedicated Application-Specific Integrated Circuits (ASICs), are designed to perform certain tasks incredibly fast and efficient, making them well-suited for accelerating various areas of network performance (Today, 2021). However, with the expansion of organizations the complexity of network traffic increases, the challenges of managing a sprawling infrastructure become apparent.

Network topology discovery, a key aspect of network management, traditionally involves manual configuration and limited visibility. The absence of robust discovery tools can impede even the most experienced administrators and network veterans in gaining an in depth and comprehensive understanding of the network's structure, impeding tasks such as future resource planning, root cause analysis (RCA) and traffic bottleneck identification. Automated discovery methods like Simple Network Management Protocol (SNMP), Domain Name System (DNS) and Internet Control Message Protocol (ICMP) provide some relaxation, but they have their own drawbacks and complexities (Pandey, 2010).

Modern Network Components

Networking equipment and methodology need to keep up with the unrelenting march of technology. This transition is driven by the migration from wired to wireless networks and

virtualization. Virtualizing network functions can provide the ability to create Virtual Machines, which operate on their own platform, allowing the chaining of these functions through its decoupling off physical hardware. This is accelerated even more by the wide spread of Internet of Things (IoT) devices and aggressive uptake of cloud-based services.

This is why network virtualization (NV) and more recently, virtualized network functions (VNFs), are on the rise with organizations driven by factors like environmental concerns. Widely recognized tech giants like VMware, Microsoft, Google and AWS stand at the helm of this revolutionary march aiming to drastically reduce the reliance on hardware and hence electricity consumption by subjecting us instead to a wireless transmission-centric network ecosystem. 4G and 5G networks have accelerated the proliferation of IoT devices, which in turn require scalable, flexible network architecture.

Network Management and Automation

With increasing complexity, network management plays a crucial role in ensuring smooth operation along with the security of foundations. Today, network management systems use a combination of software and hardware to track, analyze, manage and secure data traffic. This approach, called network automation, is one of the key elements of modern network management, that allows us setup and configuration or deploy on operating a service in minimum human intervention.

Simple Network Management Protocol (SNMP) an application layer management protocol that facilitates the exchange of information between network devices, regardless of hardware or software configuration (Oros, 2016). However, older versions of SNMP have issues ranging from lack of security to scalability, it is clear that these limitations drive the ever-changing protocol management.

Software-Defined Networking (SDN)

SDN is a new approach in the network architecture, providing flexibility and agility using centralized software control. SDN decouples traffic control from forwarding (data) and allows administrators to change flow through networks no longer with physical hands on a box. In an SDN scenario, we have a central controller that controls the behavior of network switches and routers: this tells every device how to route/process packets and in which manner. That centralized control allows for automation, streamline network management, and can speed-up be changing to business needs.

This move from old to new in networking, combined with those early-days SDN technologies and trial-by-fire lessons learned through using AI-powered network management tools marks an historic shift in how networks are built for scale. Going forward, it's all about moving at the speed of software-defined automation. In response to these demands — in an era of increasingly complex and interconnected digital ecosystems — the advent of this technology represents a roadmap towards a more agile, efficient, and secure future for organizations. The Figure 2 shows the architecture of a Software Defined Network (SDN) controller, with three key components: the Infrastructure Layer (containing network devices), the Control Layer (where SDN control software manages network services), and the Application Layer (which hosts business applications using APIs to interact with the control software).

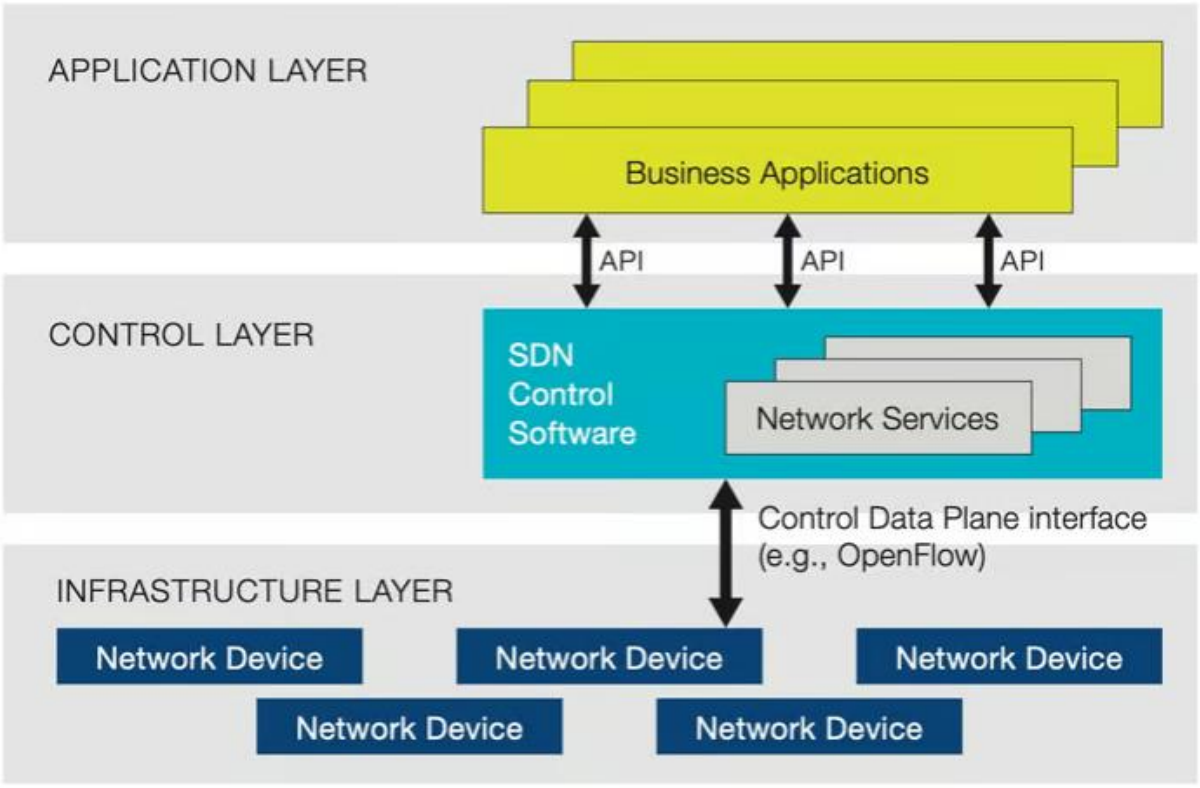


Figure 2 - Software Defined Network (SDN) Controller

2.2 Need for Network Security and Firewalls

The emergence of firewalls and network security can be credited to multiple crucial factors in the world of computer networks and information technology. The most important factor was the rapid growth of computer networks and the internet, this contributed to the increased risk in unauthorized access to valuable information or system resources. In response, network security devices and firewalls were invented to prevent unwanted personals or malicious entities who tried to infiltrate and access network systems, data, services. On the other hand, data confidentiality depends on network security. With the influx of organizations and individuals transmitting massive amounts of sensitive information across networks, steps needed to be taken to prevent this data from being viewed or accessed by unauthorized parties. Security solutions on networks made encryption, access controls and secure communication protocols to maintain data confidentiality.

Additionally, the rising number of data breaches necessitated the implementation of network security as well as firewalls. Data breaches were frequent as unauthorized parties gained access to or stole sensitive data. Network security measures, including firewalls, to detect and block unauthorized access attempts, leading to a lower risk of data breaches, as well as minimizing financial/ legal/reputational costs related to data leaks. In addition, the utilization of network security and firewalls saved the networks from disturbances as well as downtime. Any network service could be targeted by cyberattacks, malware, and network intrusions which could lead to downtimes, system failures, or productivity delays. Firewalls and network security solutions were introduced in order to identify and prevent malicious traffic, preventing denial-of-service (DoS) attacks on-line ensuring the availability as well as reliability of network resources. The network security requirement also arose in compliance with regulatory requirements. Most industries, geographies and jurisdictions imposed regulations or compliance standards regarding data security and privacy (for example General Data Protection Regulation (GDPR) in Europe). In order to comply with these requirements, so as not to be penalized legally, and to protect their reputation, organizations started employing network security measures like firewalls. In addition, the changing face of cyber threats made it clear that network security and firewalls were needed. Countless attack vectors, vulnerabilities, and sophisticated techniques began appearing at a regular pace. Such network security measures including those that protect against the spread of malware, ransomware, phishing attacks, and advanced persistent threats (APTs) form a part of an overall cybersecurity strategy.

Finally, network security and firewalls were designed to safeguard data as well as the infrastructure which the data traverses. Unauthorized access to network devices, uncontrolled network traffic and insecure connections between networks pose a significant risk on network elements. Firewalls were also strategically placed to block unauthorized access, shape network traffic, and protect network infrastructure machines by preserving their integrity and availability. Ultimately, the security of network and firewall services was necessitated by increasing risks related to unauthorized access, interception for data breaches, intentional network interruption done through protocol conformance violation attacks, evasion of compliance requirements (e.g., sensitive data leaks) as well as the inherent characteristics of cyber threats that are in constant evolution. Strong network security methods were adopted to safeguard confidential data, ensure business continuity and limit the potential fallout of cyber threats.

2.3 Traditional Firewall Technologies and Evolution

Firewalls have evolved significantly since their inception in the late 1980s. Conventional firewalls primarily engage in the tasks of packet forwarding and blocking based solely on header information. However, their effectiveness is limited in the face of application-layer attacks due to their inability to thoroughly inspect individual packets (Junyan & Yoohwan, 2022). In the mid-1990s, stateful inspection firewalls were introduced, which examined traffic at the transport layer and maintained a state table of active connections. This allowed firewalls to make more intelligent decisions about allowing or blocking traffic based on the context of the connection. To enhance protection against these sophisticated threats, organizations increasingly adopt security measures that leverage deeper packet inspection and application-layer gateways to provide comprehensive defense at the network's highest layer.

A proxy server or an application firewall serves as an intermediary between a client application and a server, positioned to facilitate communication between the two hosts. This intercepts original client requests that come from the internal network and then sends them to the destination servers, which prevents direct connection between them. Functioning as a server to the client and a client to the server, it ensures that no direct connection is established between them. Operating on the application layer, proxy servers inspect data in the application protocol headers or payloads and decide based on that information whether to allow traffic to pass or not. While standard firewalls are primarily header-based, proxy servers have the ability to filter application specific traffic. These characteristics help them to provide an advanced security

feature with flexibility, empowering them to incorporate additional security features. Basically, proxy server or application firewall is a gateway that providing heightened security to the client server connection by adding an extra layer of control and inspection client and server (Mohd Zafran Abdul Aziz, Muhammad Yusoff Ibrahim, Abdullatif M. Omar, & Ruhani Ab Rahman, 2012). These proxies are capable of inspecting the contents of the data stream to identify and block suspected nefarious activities, thereby only permitting legitimate and authorized traffic. They can operate on the application layer which also provides them with better insight to communicate, giving more granular control over traffic and strengthening network security. Figure 3 demonstrates a network architecture where a firewall and proxy server act as intermediaries between the Internet and an internal network. The firewall provides security by filtering external traffic, while the proxy forwards requests to the internal network.

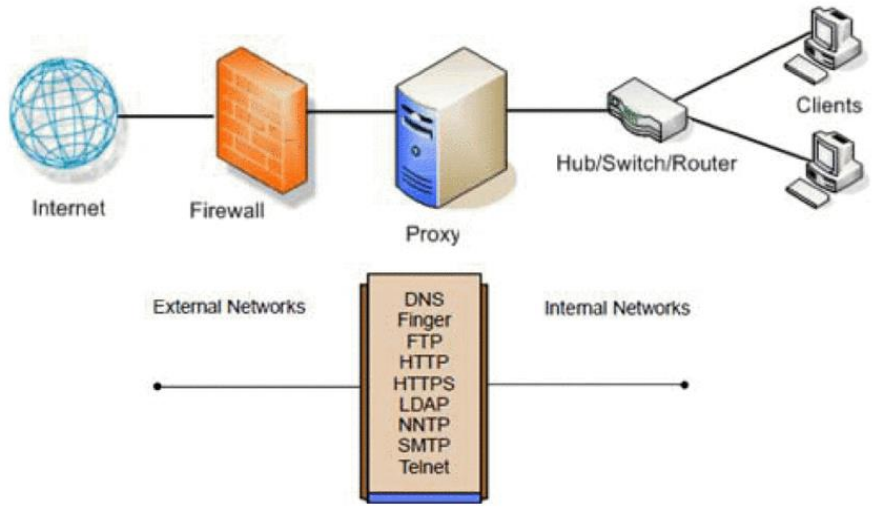


Figure 3 - Application firewall

In recent period the internet and its applications are gone through a rapid growth and development, leading to countless millions of users. At the same time, massive changes have occurred to user access and how they utilize the Internet (Kende, 2016). New threats with their sophistication make it mandatory to put up advanced protection systems for the safeguard of internet users. Despite being of great use, traditional firewalls are not effective enough to meet these evolving challenges. To address this, Next-Generation Firewall (NGFW) managed to

respond to this evolving threat landscape by offering an increased level of network traffic visibility, cross-cutting OSI layers operational abilities as well as new security capabilities to shield networking infrastructure from emerging risks (Kende, 2016) which integrates the classic firewall functionality with intrusion detection and prevention (IDP), virtual private network (VPN), and other security features. NGFWs offered more advanced security features and could conduct packet inspection much deeper to identify and prevent active threats. Like WAF, NGFW provides the capability of application awareness by decrypting protocols, inspecting payloads and using heuristic analysis. The same policy also helps in user identification thus offering user-based packet filtering at its best. Such features allow NGFW to offer strong, multi-level protection for the network and perform highly effective incident responses (Junyan & Yoohwan, 2022).

2.4 Evolution of Artificial Intelligence

Artificial Intelligence (AI) is no new term. It has been evolving for several decades and has been through several stages of development. The term "artificial intelligence" was first used by the Dartmouth University scholars in 1956 during a conference. With this, a major milestone in the field was set that engendered a new area of research discussing how machines can simulate human intelligence processes. In the beginning, AI systems were rule-based that meant they followed hard-coded rules trying to process the human line of reasoning. These systems were of a fix rule-based design, which both made it difficult to cope with complex tasks and were hard to maintain. Later, a new terminology was introduced, focusing on capturing and representing human knowledge in specific domains. Based on simple programmatic rules, capable to mimic expert human reasoning, they allowed solving problems within explicit domains. They served as a crucial foundation for comprehending and understanding human expertise. But that was just the beginning. Heuristic decision-making based on rule-based approach was its primary focus often using 'if-then' weighted rules to deal with uncertainty and ambiguity. These systems paved the way for many AI breakthroughs and innovations that followed. (Daniel E & O'Leary, 2021). Figure 4 represents how deep learning and machine learning are subsets of AI in general.

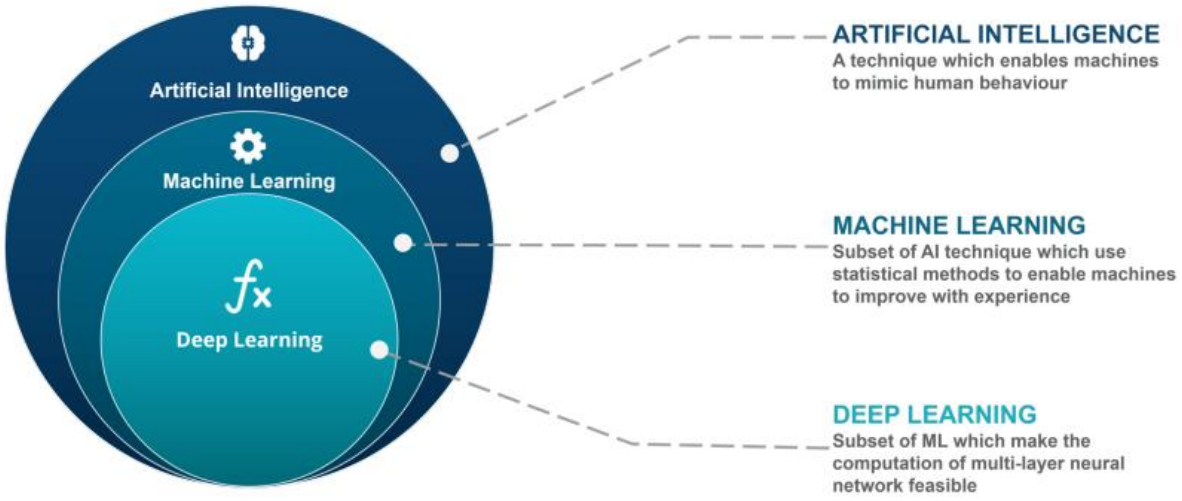


Figure 4 - Subset of AI

At the end of the decade progress in research related to AI slowed, leading to a period known as the AI Winter. Limited computational power and unrealistic expectations contributed to a decline in funding and interest in the field. In the 1990s, machine learning (ML) emerged as a new approach to AI, in which systems could learn from data and improve their performance over time. Machine learning, a subfield of artificial intelligence, involves the automated acquisition of patterns and relationships from data, eliminating the need for explicit programming (M. Mijwil & E. Salem, 2020). It has demonstrated considerable efficacy across various domains, including business, research, and optimization (S. N. Abd, Alsajri, & R. Ibraheem, 2020). Machine learning algorithms autonomously generate knowledge, train models, discern relationships, and identify previously unknown patterns. These extracted patterns and relationships are subsequently utilized to predict outcomes and optimize processes for novel and unobserved datasets. Machine learning algorithms can be categorized into three main divisions: the decision process, where predictions or classifications are made based on labeled or unlabeled data; the error function, which evaluates the accuracy of model predictions and assesses overall performance; and the model optimization process, which iteratively adjusts weights to minimize discrepancies between the proposed and estimated models, continuously improving accuracy (Aggarwal, Mijwil, Sonia, & Al-Mistarehi, 2022). Diverging from traditional software development, machine learning focuses on autonomous learning from data, enabling algorithms to develop their own approach without explicit programming. Such algorithms adapt to specific contexts and dynamically refine their capabilities based on changing circumstances.

Machine learning can be classified into three main categories: supervised learning, unsupervised learning, and reinforcement learning. As figure 5 illustrates, in supervised learning, the algorithm learns from labeled training data, where each data point is associated with a corresponding target or output value. By generalizing from labeled examples, the algorithm can learn how to map input features to the correct output. It can then use this mapping to make predictions on new, unseen data. Some of the popular supervised learning algorithms used are decision trees, linear regression, support vector machines (SVM), and neural networks (NN) to name among others. A very common example is Email Spam Classification, in which the algorithm works on a dataset that's made up of spam & non-spam emails. By utilizing the patterns within the existing data, the algorithm is able to classify new incoming emails as either spam or non-spam, effectively filtering out scam or unwanted messages. The Stock Price Prediction (Regression) where it is trained with historical market data from a certain business domain. The algorithm uses regression analysis to find patterns and relations within the data, thus it can then make predictions for future stock prices. This prediction capability assists investors and traders in making informed decisions based on projected price trends.

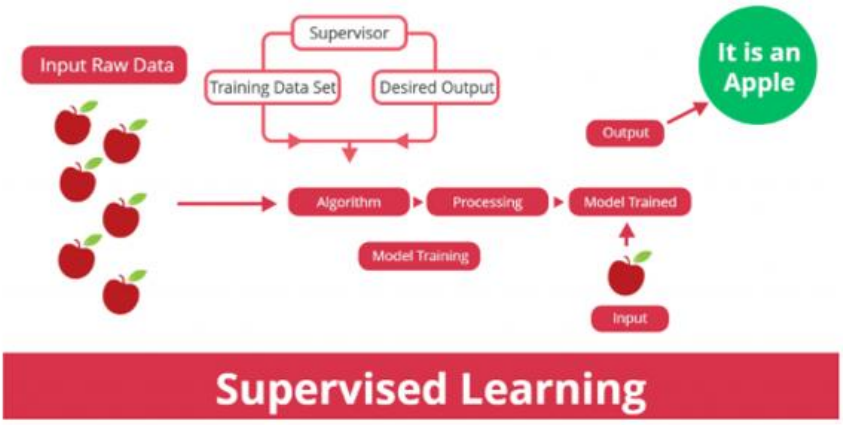


Figure 5 - Demonstration of Supervised learning

On the other hand, as shown in figure 6, unsupervised learning involves no labeled data and thus discovering arbitrary patterns or structures by optimizing the algorithms. When the target values are not directly given, unsupervised learning algorithms search for clusters, anomalies, or patterns etc. Types of unsupervised learning algorithms include clustering algorithms (e.g., k-means, hierarchical clustering), dimensionality reduction (principal component analysis, singular value decomposition), and generative models (Gaussian mixture model, hidden Markov model).

For example, Clustering is an unsupervised learning where similar data points are grouped together based on rules or similar patterns not defined by us. K-means Clustering is one of the most popular algorithms. For example, take a data with client properties like age, income and spending habits. Unsupervised learning methods, for instance k-means clustering can do a very similar thing in which you do not need to specify the number of customer segments upfront and instead clusters are formed based on similarity of attributes like the one mentioned above. Businesses use this method to find deeper patterns, structure or customer segments for creating targeted marketing strategies or personalized recommendations. Unsupervised works well for such not labelled data and provide some benefited insights regarding decision making as customer behavior.

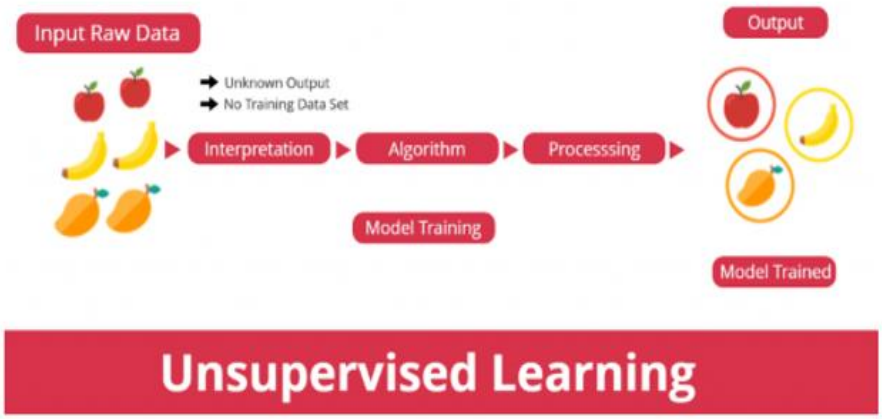


Figure 6 - Example of Supervised learning

Reinforcement learning is a subset of machine learning in which an agent learns how to make sequential decisions in an environment to maximize cumulative rewards. The agent interacts with the environment by taking actions based on its current state and receives feedback in the form of rewards or penalties. The agent goes through a trial and error process of trying out different actions, noticing their outcomes then updating its policy/strategy accordingly in order to optimize decision making. These techniques involve value iteration or policy gradients. For example, reinforcement learning can be used to train an autonomous robot to move through a maze. The robot attempts a possible path, positive reward when reaching the goal and negative in case of hitting an obstacle. The robot learns to move effectively with repeated iterations and discern the shortest path to the destination. Reinforcement learning is generally considered for the case when it has to operate in high dimensional uncertain and complex environments, an

example of these tasks are robotics, game playing and autonomous systems. Figure 7 shows reinforcement learning in which how machine keeps on learning based on reward.

Figure 7 - Example of Reinforcement learning

2.5 Role of Artificial Intelligence and Machine Learning in Firewall

2.5.1 Limitations of Traditional Firewall

While traditional firewalls form an essential pillar for network security, their functional boundaries confine them within the dynamic and evolving threat landscape. These limitations are because of their dependence on the unchanging hard-coded rule sets used in them, which often proves ineffective in preventing novel and complicated attack vectors (Mirsky, 2018). Consequently, these rule sets can be circumvented by tailored attack techniques, necessitating frequent manual updates and thus systems vulnerable to zero-day exploits (Institute., 2019). Furthermore, the most firewalls are located on network and transport layers of Open Systems Interconnection (OSI) model and deal with filtering packets by the IP addresses or ports (Scarfone, 2007). This approach fails to solve threats designed for application-layer protocols, which have been leveraged by attackers to avoid detection (Antonakakis, 2017). Compounded even further by the fact that we can't see inside of encrypted traffic, as malicious payloads can be concealed within seemingly benign communication channels (Cisco, 2021). Additionally, traditional firewalls often have a hard time with scalability and performance in high-traffic environments. Firewall that use deep packet inspection and complex rule matching can create

overheads that may lead to network bottlenecks and latency problems, ultimately degrading network performance (Firewall.cx., 2023). This limitation becomes even more evident within today´s networks filled with variousity of traffic and increasing bandwidth demands. Figure 8 shows how a packet filtering firewall prevents IP spoofing by checking incoming packets' source IP addresses against established rules.

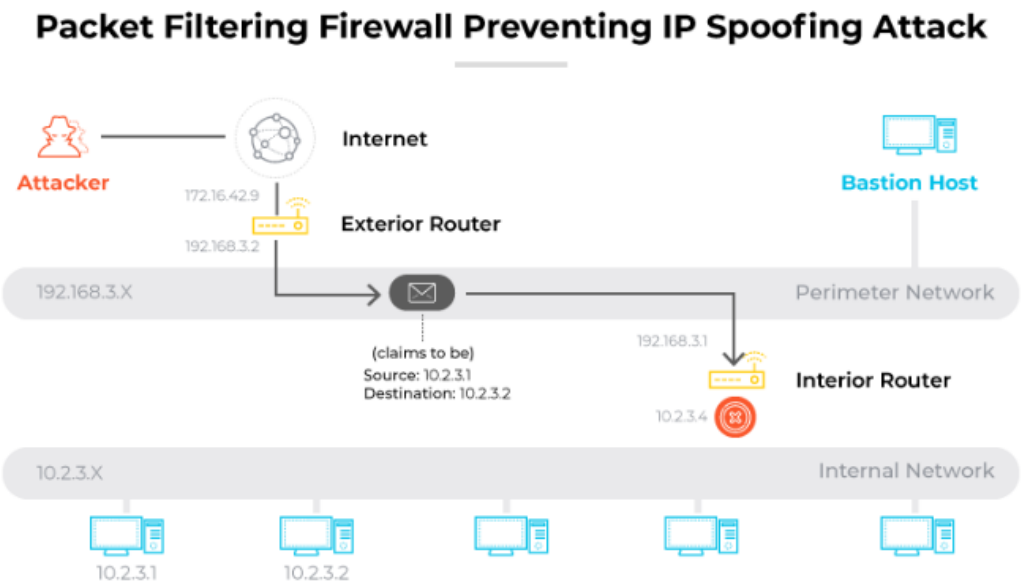


Figure 8 - Packet filtering firewall preventing spoofing attack

To address these challenges, the integration of AI-driven solutions with firewall architecture has emerged as a promising solution. Machine learning and deep learning algorithms have the potential to process large-scale network traffic data at high speeds, detecting suspicious anomalies that can be used as malicious activity (Al-Haijaa & Ishtaiwia, 2021). Since the threats are always changing, these AI-powered firewalls continuously learn and improve their defenses as well to provide better protection against zero-day attacks or other unknown emerging sophisticated threats. For instance, as explored by (Mirsky, 2018), the implementation of negative security models within firewalls can significantly reduce computational overhead while maintaining robust security through blacklisting attributes. This technique is in line with the growing trend of anomaly-based intrusion detection, in which deviations from established baseline behaviors are marked as potential threats. Moreover, AI can help in automating the rule

set generation and updating process so that manual intervention is kept to a minimum as this will bring down the possibility of any human errors. Using Natural Language Processing (NLP) AI can interpret threat intelligence feeds and security advisories, updating firewall policies to reflect the latest threats automatically (Cisco, 2021).

While traditional firewalls remain a foundation of network security, their limitations mandate the adoption of more adaptive and intelligent solutions. By increasing the inclusion of AI techniques into firewall architectures, it opens a way for improving security posture, providing proactive defense against novel emerging threats, and automating critical security tasks. More research in this domain is carried out to focus on exploring the synergy between firewalls and other emerging technologies like blockchain and zero-trust networking to further fortify network defenses.

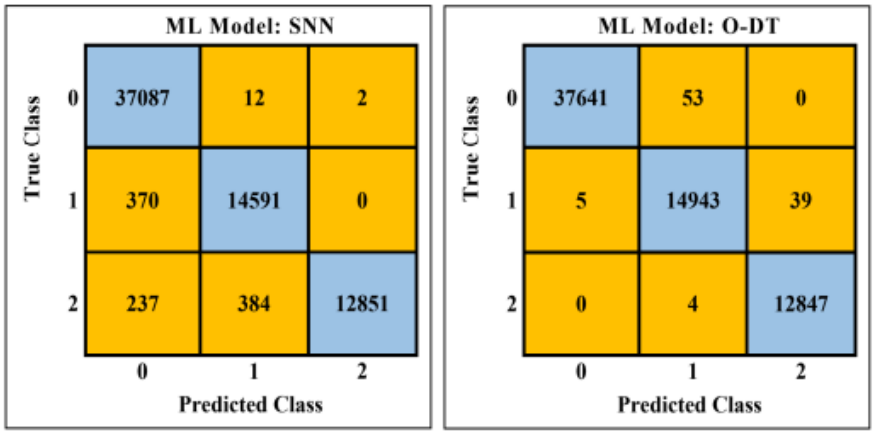
2.5.2 Use of Machine Learning Techniques to Identify Firewall Decisions

A firewall helps to regulate traffic flow for packets in and out of communication networks. Its primary objective is to make intelligent and informed decisions to enhance cyber defenses and deal with malicious packets. This is done through a filtering process in which traffic packets are evaluated against set rules that should prevent cyber threats from attacking the network. The firewall system therefore decides whether to "allow," "deny, or "drop/reset" each incoming packet according to these rules.

A study was conducted to develop an intelligent classification model that can be combined with firewall systems to effectively handle and respond to every network traffic packet. This model utilizes two machine learning approaches: Shallow Neural Network (SNN) ANN architecture having an input vector of size 11 features connected to 150 neurons in a hidden layer. For each 3-output neuron, a "allow", a "deny" and a "drop/reset" class. For computer vision, all networks use trainable weight vectors for each layer. The input elements are multiplied by the corresponding weights and passed through a summation operation, followed by the addition of bias to obtain the neuron's output value and optimizable decision tree (ODT) which is a decision tree-based classification and prediction tool. Our ODT configuration includes 11 predictors and one response variable. The tree is built using a split criterion of maximum deviance reduction, with a maximum number of 30 splits and 30 iterations using 5-fold cross-validation, enabling analysis of packet attributes. For classification module, SoftMax activation function was utilized, which normalizes a vector of real numbers into a probability distribution. The final neuron outputs from the previous layer are activated using the sigmoid function. The numerical

probabilities for each class are calculated using SoftMax, ensuring that the probabilities sum up to 1. The class with the highest probability is selected as the output label. The multi-class internet firewall (IFW-2019) dataset with 70% of the data used as a training dataset, 15% for validation purpose, and 15% to be used as testing dataset, consisting of packets classified into three classes ("allow," "deny," and "drop/reset"), was utilized to train and evaluate the proposed models.

The developed classification model was evaluated using various performance metrics (Q. A. Al-Haija, M. Smadi, & S. Zein-Sabatto, 2020). These metrics include the multi-class confusion matrix, which consists of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Additionally, the evaluation metrics include classification error percent (CEP), classification accuracy (ACC), true positive rate (TPR), positive predictive value (PPV), false-negative rate (FNR), and false discovery rate (FDR). The confusion matrix results for the aforementioned classifier are shown in Figure 9 below



	ACC%	CEP%	TPR%	FNR%	PPV%	FDR%
SNN	98.50	1.50	98.47	1.53	98.51	1.49
ODT	99.8	0.20	100.0	0.00	99.7	0.30

Figure 9 - System Evaluation: Overall evaluation metrics and Confusion matrix (Al-Haijaa & Ishtaiwia, 2021) (Al-Haijaa & Ishtaiwia, 2021)

The experimental results demonstrated the superiority of the classification model, achieving an overall accuracy of 98.5% for SNN and 99.8% for ODT. Furthermore, additional machine learning metrics, including positive predictive value, true positive rate, and others, were assessed to provide deeper insights into the system's performance. Ultimately, through a comparison with the current state-of-the-art models in the field, the achieved results surpassed

existing automated classification models for firewall actions. Ultimately, the proposed system outperformed many existing contemporary firewall systems, indicating its significant advancements and superiority in terms of accuracy and performance. (Al-Haijaa & Ishtaiwia, 2021).

2.5.3 Machine Learning Techniques for Intrusion Detection Systems

Supervised algorithms possess the characteristics of training with fully labeled data and establishing the relationship between input and target units. They can be categorized into classification and regression. Well-known classification algorithms include Support Vector Machine (SVM), Naïve Bayes, Nearest Neighbor, Neural Network , Discriminant Analysis, and Logistic Regression . Regression algorithms commonly employed in intrusion detection analysis encompass Linear Regression, Support Vector Regression (SVR), Ensemble methods, Decision Tree (DT) , and Random Forest . On the other hand, unsupervised learning techniques uncover hidden structures in unlabeled data without explicit training. Reduction and clustering serve as major techniques for grouping data for comparison and compression with unique identification. Popular clustering algorithms include K-Means and C-Means. Feature reduction techniques such as Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) are also widely used.

Table 1 presents a comprehensive overview of the properties, advantages, and issues associated with different machine learning approaches in Intrusion Detection Systems (IDS). This table underscores the necessity and significance of each technique in the detection process, enabling future IDS developers to make informed choices regarding the appropriate approach to adopt (P. L. S. Jayalaxmi, Rahul, Kumar, Conti, & Kim, 2022).

Table 1 - Comparison of Machine Learning Techniques for IDS

Approaches	Properties of the Method	Advantage	Issues
Signature (knowledge-based Detection)	Processing with defined signatures. Pattern matching technique.	Low False positive rate with fast and accurate detection Effective in detecting known pat- terns, Flexible, and Robust. Suitable for all network levels Source	Difficulty in analyzing information/state new attacks are ignored. Complex in updation.
Anomaly-based Detection	Create a default profile for identi- fying the normal state Deviation from baseline detected as an anomaly. Profile updates based on new behavior and critical events	Identification based on labels and behavior (normal/abnormal). Specialized in identifying Zero- day attacks Multiple profile management dif- ficult for the hacker hide identifi- cation.	More time consuming for training phase. Alteration of the threshold is dif- ficult for managing false rates. Low efficiency and high computational cost.
Hybrid Detection	Combined detection process. Both Predefined pattern behavior Detection	Suitable for both Supervised and unsupervised methods Integration of methods for best performance.	Integrated product of Anomaly and signature detection system with added advantages the implementation cost is very high comparatively.
Probabilistic packs marking- based attack source detection	Encodes information into a packet header. Identification field are used to mark and reconstruct attack path.	Manage with the regular traffic for communication Reconstruction of attack path without ISP influence.	High false-positive rates Requirement of large number of packets.
Deterministic Packet Marking Based Attack Source detection	Marks packets which are near to the source of the attack. 16-bit identification fields are used.	Reduces storage and computational overhead flexible for small packet.	High false-positive rates. Trace only nearby source. identi- fication of origin is delayed.
Support Vector Machine	Hyper plane setup for traffic area, suitable for classification and re- gression. Effective in parameter identifi- cation, implemented for discreet valued kernels.	High accurate Handle complex nonlinear deci- sion boundaries Less over fitting problems.	Complex in implementation and extensive memory requirement Choice of Kernel is difficult Slow in training and testing.
K-Nearest Neigh- bour	Classification and decision based on behavior patterns /classes. Multiple parameters and transfor- mation with K nearest value.	Analytically tractable, Simple in implementation. Use local information and yield highly adaptive behavior. Suitable for parallel processing.	Huge storage space requirement Highly susceptible to the curse of dimensions Slow in the classification of test tuples.
Bayesian Method	Follow Joint probabilities rules. Eliminate condition with relative frequencies form training sets,	Simplifies the computations. High speed and accurate for large database	Decision based on assumptions Lack of available probability data (less updated).

Decision Tree	Based on binary classification nodes correspond to Variable/attributes Branches for positive and negative instance.	Construction does not require any domain knowledge. Capable of handling high dimensional data. Suitable for numerical and cate- gorical data.	Output attribute must be categori- cal. Unstable result patterns. Complex as created with numeri- cal datasets
Artificial Neural Network models	An adaptive system with chang- ing structure based on informa- tion flow in the network Depended on training element find the distance of comparison	Require less formal statistical training. Detect complex nonlinear rela- tionships between variables High tolerance to noisy data, with multiple training algorithms	Black box nature (based on speci- fications). Greater computational burden. Requires long training time

Also, Deep learning techniques, known for their exceptional self-learning capabilities, offer significant advantages in the development of intrusion detection models. These models exhibit lower false rates and higher accuracy when compared to traditional machine learning methods. The standard architecture for deep learning in this context is the multi-layer perceptron, constructed using a linear stack classifier. Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) have emerged as the most widely used methods for detecting malware activities, employing powerful self-learning techniques. Natural Language Processing (NLP) exhibits strength in identifying spam and social engineering attacks, leveraging new forms of communication and language patterns. Artificial Neural Networks (ANN) excel in monitoring network traffic and detecting imminent attacks. ANN, DNN, and CNN are examples of supervised instance learning utilizing feed-forward neural networks, while RNN and Long Short-Term Memory (LSTM) fall under the category of supervised sequence learning methods. Restricted Boltzmann Machine (RBM) and Deep Belief Network (DBN) belong to the semi-supervised instance learning approach. Deep Learning (DL) also supports transfer learning methods, allowing the generation of generic problem statements that can be reused with other models. CNN and RNN are widely employed for conditional/discriminate models, while AE, DBN, RBM, and GAN are considered generative DL techniques. Combining both approaches forms an ensemble technique. In Table 2, we discuss the importance of various DL techniques for IDS and highlight associated issues (P. L. S. Jayalaxmi, Rahul, Kumar, Conti, & Kim, 2022).

Table 2 - Comparison of Deep Learning Techniques for IDS

Approaches	Properties of the Method	Advantages and Applications	Issues
Convolutional Neural Network	The convolutional layer utilizes a kernel function, while the pooling layer employs max pooling. These layers are widely used in image processing applications	Automatic feature learning methods are known for their high accuracy and performance. They are commonly applied in tasks such as face recognition, image classification, action recognition, and human pose estimation	Complex in updation and high computational cost. Also limited to some applications
Recurrent Neural Network	Store previous input state and pre- serves the relationship Self loop structure Good for time series prediction	Feed forward method with back- ward connection points. Long Short-Term Memory for lengthy- time period dependencies. Anomaly Detection, Stock Price Forecasting, Sentiment Analysis etc.	Existence of vanishing gradient problem. Fixed Model Size Compatibility issues with Tanh or Relu activation feature.
Auto-Encoder (AE)	Preferred for dimensionality re- duction. Equal Input and output layers. compress and decompress the data	low-dimensional abstraction and training with back propagation. Sparse E, Denoising AE ,Con- tractive Data Compression, Image Denoising,Dimensionality Reduction,Image Generation etc.	Additional computation time. Deterministic bias resulting with over fitting problem
Restricted Boltzmann machine	Bidirectional data flow Transform high dimensional data.	Restrictions connections faster performance. Motion-capturing video sequenc- ing, image procession etc.	Unsupervised training lack of general application Procession on unstructured in- put, but explicit structure is not considered.
Deep belief net- work	Integrated component with RBM and sigmoid Generates deep hierarchical rep- resentation.	Sequential learning strategies. Unsupervised learning and avoiding over-fitting and under-fitting problems.	Increased run time complexity Low processing rate for clamped inputs.
Generative adversarial network	Combination of generative and distributive model High potential rate self training to mimic distribution of data.	Easy training compared to RBM and DBN. Domains used are music, image, speech, prose.	Unstable training. Complex for Text representation.

ML and DL techniques reduce human intervention and automate detection in a short time. DL models are not compatible with large datasets and complex structures as compared to the ML techniques. ML techniques are mostly used for signature intrusion detection that act according to the stored patterns. On the other hand, DL has a capability for self-learning; hence, it is more

compatible for anomaly detection. Analyzing and detecting the attacks based on behavior helps in handling zero-day vulnerabilities. Though the ML techniques require less computational power, the DL techniques are faster than the ML techniques. In short, the use of ML and DL methods enhances the performance of IDS systems.

2.5.4 Explainable AI (XAI) and the Future of Cloud Security

Explainable AI (XAI) emerges as a powerful tool to address long grappled with limitations like high false positive/negative rates and opaque decision-making processes by providing interpretable explanations for AI-driven security decisions and improving security analysts' ability to understand alerts, leading to efficient resource allocation and quick responses. XAI techniques like LIME and SHAP offer a glimpse into the "black box" of machine learning models used in IDS. These explanations empower analysts to differentiate between genuine threats and benign activities, leading to more informed security decisions. The benefits of XAI extend beyond immediate threat detection. By analyzing XAI explanations, security teams can identify weaknesses within the IDS model and use this knowledge for targeted retraining and adaptation, ensuring the IDS remains effective against evolving threats. Additionally, XAI facilitates collaboration by providing a shared understanding of the rationale behind alerts, fostering a more cohesive security posture. (Biggio B. D., Fairness in Machine Learning: A Survey of the State of the Art, 2022, August)

The future of XAI and IDS integration presents exciting possibilities. Researchers are exploring ways to refine XAI techniques for cloud security and develop hybrid models that combine machine learning accuracy with XAI interpretability. Furthermore, XAI's integration with threat intelligence feeds holds promise for proactive threat identification and mitigation strategies. However, implementing XAI in cybersecurity necessitates careful consideration of ethical implications like accountability, fairness, and potential biases within explanations. Collaborative efforts are crucial to establish robust ethical frameworks that govern the development and deployment of XAI-enhanced IDS. Widespread adoption requires collaboration across stakeholders – researchers, security practitioners, and cloud service providers – to develop and integrate XAI solutions. Educational and training initiatives are crucial to equip security analysts with the necessary skills to leverage XAI-generated insights, ultimately strengthening the overall security posture of cloud environments. In conclusion, XAI presents a transformative paradigm shift in cloud intrusion detection. By fostering transparency, interpretability, and continuous

learning, XAI empowers security analysts, optimizes decision-making, and paves the way for a more robust and adaptive cloud security landscape. (Utsav Upadhyay, 2023)

Table 3 - Comparative summary of the main XAI techniques for cybersecurity

Approach	XAI Method Level	Cybersecurity Class	Technique
Harder et al. [84]	Local explanation	Privacy	locally linear maps (LLM)
Löbner et al. [85]	post-hoc explanation, Model-agnostic explainer	Privacy	Information Gain (IG) and extended Iterative Dichotomiser 3 (ID3) classification tree
Rjoub et al. [86]	Local explanation, Post-hoc explanation	Trust	Double Deep Q Learning (DDQN)
Rjoub et al. [90]	Local explanation, Post-hoc explanation	Trust	SHAP
Machlev et al. [88]	Model-agnostic explainer, Post-hoc	Trust	Grad-CAM, LIME
Kuppa et al. [89]	post-hoc explanation, Model-agnostic explainer	Trust	Input*Gradient(I*G), Layer-Wise Relevance Propagation(LRP), Guided Back Propagation(GBP), Smooth- Grad(SG), Gradient(GRAD), and Integrated Gradients(IG)
Mankodiya et al. [93]	Model-specific explainers	Trust	Decision Tree-based algorithms
Elayan et al. [94]	post-hoc explanation	Trust	IoB-XAI
R. Karn et al. [96]	Local explanation, post-hoc explanation	ID	SHAP & LIME
M. Wang et al. [97]	Local & Global explanation	ID	SHAP
G. Baryannis et al. [98]	Model-specific	ID	Decision Tree & SVM
L. Aguilar. [99]	Local explanation, post-hoc explanation	ID	Decision Tree
G. Iadarola et al. [100]	Global explanation, Model-specific	Intrusion prevention	Gradient-weighted Class Activation Mapping (Grad-CAM)
L. Marino et al. [101]	Local explanation, Model-specific, post-hoc explanation	Intrusion prevention	Adversarial machine learning
Y. Al Hammadi et al. [102]	Local explanation, pre-hoc explanation, Model-specific	Access Control	SHAP
C. Seibold et al. [103]	Global explanation, Post-hoc explanation	Access Control	Focused Layer-wise Relevance Propagation (FLRP)
W. Garcia et al. [104]	Model-agnostic explainer, Post-hoc explanation, Local explanation	Authentication	LIME
R. Rocha et al. [105]	Global explanation, Model-agnostic explainer, Post-hoc explanation,	Authentication	Gedeon method

The above table summarizes how Explainable Artificial Intelligence (XAI) techniques are applied in cybersecurity. It compares the approaches used in various studies, the level of explanation provided by the XAI method (local vs. global, pre-hoc vs. post-hoc, or model-specific), and the cybersecurity category it addresses (privacy, trust, intrusion detection/prevention, access control, or authentication) (Yin, 2020).

2.5.5 Challenges and Limitations of AI-powered Firewalls: A Critical Examination

The Achilles' heel of AI-powered firewalls lies in their inherent data dependence. The quality and quantity of training data significantly influences the model's performance. Insufficient data creates a vulnerability known as underfitting, where the model performs well on familiar patterns but struggles to recognize novel attack tactics or network behaviors. This is particularly concerning the ever-shifting landscape of cyber threats, where attackers constantly devise new methods (Biggio B. e., 2022). An even more insidious issue arises from biased data sets. Imagine a model trained primarily on data dominated by a specific type of cyberattack. It might become adept at detecting those attacks but miss entirely novel attack vectors that differ from the training data. Additionally, biased data can lead to unfair flagging of legitimate traffic, causing unnecessary disruptions and wasting valuable security resources.

Addressing these limitations requires a multi-faceted approach. One promising avenue involves generating synthetic training data that mimics real-world network traffic patterns. Techniques like generative adversarial networks (GANs) can create realistic and diverse data sets, improving

model generalizability and reducing reliance on scarce real-world attack data. Active learning offers another solution. This approach involves iteratively training the model by prioritizing the most informative data points. The model actively queries security analysts or utilizes uncertainty sampling techniques to identify network traffic that is most challenging to classify. This focused learning process can significantly improve model performance with minimal training data. Researchers are also actively developing methods to identify and mitigate bias in training data sets. Techniques like data augmentation (adding variations to existing data points) and adversarial debiasing (training a secondary model to identify and remove bias from the primary model) hold promise for creating fairer and more robust AI-powered firewalls. By leveraging these advanced approaches, researchers can create AI-powered firewalls that are less susceptible to data dependence and bias (Yin, 2020). This will pave the way for robust and generalizable network security solutions capable of adapting to the ever-evolving threat landscape.

Security analysts often lack clear insights into the rationale behind the firewall's decisions due to the inherent "black box" nature of AI systems. This lack of explainability and transparency hinders trust and the ability to fine-tune security strategies. Explainable AI techniques offer solutions, but developing effective XAI methods for complex security applications remains an ongoing challenge. The efficacy of AI-driven firewalls hinges on their ability to ingest and meticulously analyze enormous volumes of network traffic data. This, however, presents a substantial challenge in terms of user privacy. The collected data can potentially encompass sensitive details concerning user behavior and network activity. Striking a critical equilibrium between robust security and user privacy is paramount. Fortunately, advancements in privacy-preserving machine learning (PPML) offer promising solutions (Chakraborty, 2023).

Training and running complex AI models can be computationally expensive, requiring significant processing power and memory resources. This can be a challenge for smaller organizations with limited IT infrastructure. Additionally, high computational demands can introduce latency into network traffic processing, potentially impacting network performance. Then AI models themselves can be susceptible to security vulnerabilities. These vulnerabilities can be exploited by attackers to gain unauthorized access to the firewall or manipulate its decision-making process. Regular security audits and vulnerability assessments are essential for mitigating these risks. The regulatory landscape surrounding AI-powered security solutions is still evolving. As AI adoption increases, clear guidelines and regulations are necessary to ensure responsible development and deployment of these technologies, balancing security needs with ethical

considerations. While AI can automate many security tasks, human expertise remains essential. Security analysts are crucial for interpreting XAI explanations, validating model performance, and making critical security decisions that cannot be fully automated.

Overcoming these challenges requires a multi-pronged approach. Researchers are actively developing robust AI models resistant to adversarial attacks and biases. Advancements in XAI techniques are crucial for improving model transparency and fostering trust among security professionals. Additionally, privacy-preserving machine learning methodologies hold promises for balancing security with user privacy. Despite their potential, AI-powered firewalls have limitations that must be addressed for widespread adoption. The success of these firewalls' hinges on the development of robust training methodologies, effective XAI techniques, and privacy-preserving approaches. Collaboration between researchers, security professionals, and policymakers is essential to ensure the responsible development and deployment of AI-powered firewalls, creating a future of robust defense against ever-evolving cyber threats while maintaining security, transparency, and ethical considerations.

2.6 AI in Network Segmentation and Microsegmentation

Network segmentation and particularly Microsegmentation have emerged as a key technique in modern cybersecurity architectures to mitigate against cyber risks and lateral movement of threats while also reducing the attack surface. Traditionally the segmentation would involve dividing networks into sub-networks (or zones) to separate various parts of an organization's digital infrastructure. However, these methods have worked to an extent, the increasing complexity of modern networks, particularly in cloud environments, mandates more dynamic and automated approaches. AI revolutionizes the implementation of network segmentation, especially in Microsegmantation where it provides dynamic and granular level of controls with ability to detect & response to real-time threat. This section delves into how AI is revolutionizing network segmentation and Microsegmentation, and why this development is critical for the next generation of firewalls.

Previously, network segmentation methods use static rules, allow or block rules, which often fall short in dynamic, distributed environments such as cloud networks, hybrid infrastructures, and software-defined networks (SDNs). On the other hand, AI-based segmentation provides a more dynamic framework that continuously monitors network traffic patterns and detects aberrations indicative of potential security threats. This is particularly helpful when it comes to the

identification of anomalies with abnormal behavior that could be indicative of a threat attempting lateral movement across your network. Machine learning (ML) algorithms power AI-driven segmentation systems, enabling them to recognize deviations by building models of typical traffic behavior over time. As an example, deep learning algorithms can classify, and cluster network activities based on massive datasets enabling the implementation of microsegmentation policies that dynamically adapt to traffic flows. The ability of AI to learn from real-time data allows for proactive segmentation that responds to security issues before they escalate into critical vulnerabilities (Wang Y. Z., 2022).

AI-driven microsegmentation provides finer control by segmenting the network down to the level of individual workloads or applications, rather than merely at the sub-network level. This level of granular segmentation is necessary in environments where you have a variety of applications, services and devices interacting. Systems, with the aid of AI can also dynamically assign policies on each application or device based on this behavior such that they will only be allowed to communicate when and where it is necessary. For instance, In a healthcare setting, AI-powered microsegmentation can prevent medical devices from communicating with only authorized systems and thereby limiting the threat of lateral movement targeting critical patient data repositories. AI systems autonomously enforce security policies that restrict access between different zones or microsegments, preventing attackers from exploiting vulnerabilities in one part of the network to compromise other areas. AI's can assist in microsegmentation by dynamically adjusting policies based on risk factors like anomalous traffic behavior or malware detection (Feng, 2020).

One of the most important benefits that AI adds to network segmentation and microsegmentation is its capability for segmentation self-optimization without human intervention. Traditionally, security experts were asked to create and manage ACLs for each segment manually or create firewall rules which were prone to error and hard to scale. AI-powered systems automate this process by using algorithms that continuously refine segmentation rules based on real-time data and threat intelligence. They can recommend improvements in security policies, identify rules that provide no value or are overly permissive (leading to excessive risk) and even force automatic enforcement of new, more secure network policy. With more and more organizations adopting multi-cloud and hybrid cloud environments, this automation gets even more important. Research show AI algorithms tune firewall rule sets to better enforce the microsegmentation policies by reducing overhead related to manual configuration (Gupta, 2021).

A bottleneck associated with traditional microsegmentation resides in the risk of performance degradation since granular security policies become more difficult to handle. AI offers a solution by optimizing how a microsegmentation system will perform, whilst still enforcing security controls without causing performance hits on network. AI-based systems guarantee critical traffic flows are not unnecessarily impacted by security enforcement mechanisms, latency-sensitive applications being a verifying example. For instance, AI has the potential to strike a balance between security needs and performance requirements without compromising on high-speed data driven applications like those used in financial trading systems or large scale Big Data platforms. AI also automatically fine-tunes segmentation policies based on security requirements and operational realities, so that it ensures that security does not come at the cost of operational efficiency. (Liu, 2020) has shown model-driven segmentation with the help of AI reduces policy enforcement overhead at SDN level leading to better security and network performance.

Lastly, AI in network segmentation and microsegmentation can help to build a more holistic zero-trust security model. Zero trust is predicated on the notion that no entity, whether inside or outside of a network, should be trusted by default and all communications must be verified first. AI-driven microsegmentation elevates the zero-trust model by validating, in real-time via behavioral analysis and threat intelligence of every communication between segments. The monitoring of every interaction at the most granular level possible allows AI systems to identify and prevent suspect Link activity, blocking its access before that traffic can spread across a network. To sum it all up, AI is revolutionizing the way organizations handle network segmentation and microsegmentation, with an entirely new set of security models characterized as dynamic adaptive mechanisms at scale. Advanced AI-driven systems, using machine learning and real-time analytics, can continually improve their segmentation policies while also rapidly detecting and responding to threats across the network and optimize security enforcement mechanisms performance. As networks get more complicated and spread out, microsegmentation's reliance on AI would only increase giving organizations the opportunity to implement a strong security measure. The future of network security resides at the intersection between next-gen AI and microsegmentation, enabling organizations to accurately defend themselves with agility and precision in a rapidly changing threat landscape.

3. Case Studies and Real-World Examples

3.1 Google Chronicle's Security Operations Platform

Google Chronicle is a case study that marks a significant step forward in utilizing AI for cloud security. It represents a paradigm shift, reshaping how organizations manages threat detection, investigation, and how to shape response in the ever-evolving cloud environment. Built upon the foundation of Google's vast experience in data analytics and machine learning, Chronicle's Security Operations platform empowers security teams to transcend the limitations of traditional, rule-based approaches (Google, https://cloud.google.com/chronicle/docs/overview, 2022).

By harnessing the power of AI, the Chronicle enables proactive identification and neutralization of sophisticated threats lurking within the vast expanse of cloud data, often before they can inflict substantial harm. This is achieved through a harmonious blend of cutting-edge technologies, including a unified data lake, advanced detection engines, and seamless threat intelligence integration. Figure 10 shows an overview of google security operations platform.

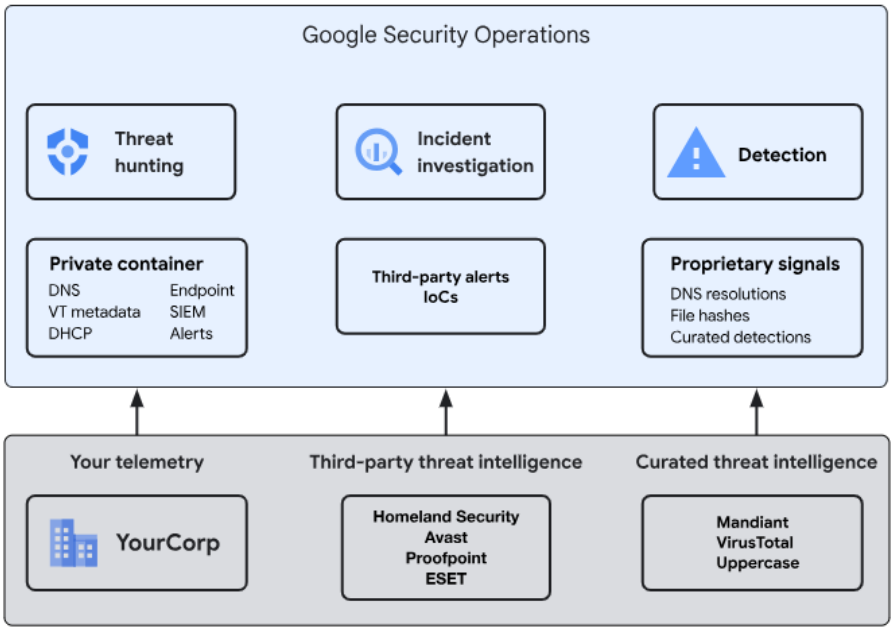


Figure 10 - Google Security Operations platform overview

As cloud adoption continues to accelerate, the complexity and volume of security data pose an ever-growing challenge. Chronicle's AI-driven capabilities equip organizations with the tools to

navigate this complexity, extracting actionable insights from the noise and enabling rapid, informed responses to emerging threat (Google, https://cloud.google.com/security/products/security-operations?hl=en, 2022). In essence, Google Chronicle embodies the convergence of cloud security and artificial intelligence, offering a potent solution for organizations seeking to safeguard their digital assets in an increasingly complex and interconnected world.

Focus: Chronicle's primary focus lies in enhancing threat detection, investigation, and response (TDIR) within cloud environments. It aims to do this by providing security teams with a powerful, unified platform to ingest, analyze, and act upon massive volumes of security telemetry data. These features include:

Unified Data Lake:

- The Chronicle excels at ingesting and normalizing security data from a multitude of sources, creating a comprehensive and readily searchable repository. This centralized data lake enables analysts to effortlessly move between wide data sets, gaining a holistic and contextualized view of the threat landscape. The power of this unified approach is exemplified in case studies like "Solve cloud security mysteries: Crack cases with Google Chronicle and Mandiant," which highlights how Chronicle's data lake facilitates rapid investigation and response to complex threats. (https://sada.com, 2023)

Natural Language Processing (NLP) for YARA-L Rules:

- (YARA for Logs) is a specialized language used to create rules for searching, filtering, and detecting patterns or anomalies within large volumes of log data. It is an extension of the original YARA language, widely used in malware analysis for identifying and classifying malicious files based on their characteristics. At the core of Chronicle's detection capabilities is the YARA-L rule engine, a powerful tool that enables advanced pattern matching and anomaly detection across the unified data lake. By utilizing YARA-L rules, security teams can detect even the most subtle signs of compromise, greatly enhancing their ability to detect emerging threats. NLP techniques aid in parsing and understanding the intent of YARA-L rules, potentially allowing for more natural language-like expressions in defining patterns

Deep Learning for Threat Intelligence Analysis:

- Chronicle seamlessly integrates with Google's extensive deep learning techniques for threat intelligence feeds, bolstering its detection capabilities and providing invaluable context for potential threats. This DL integration ensures that security teams are armed with the latest information on emerging threats, enabling them to make informed decisions and respond proactively.

Case Management and Collaboration:

- To streamline investigations and foster collaboration, the Chronicle offers intuitive case management tools. These tools facilitate efficient communication and coordination among security analysts, enabling them to work together seamlessly to address threats.

AI's Impact on Enhancing Google Chronicle:

1. **Enhanced Detection:**

- AI-powered anomaly detection and threat intelligence integration lead to a significant reduction in the time it takes to detect threats. A study by Google found that Chronicle customers experienced a **60% reduction in mean time to detect (MTTD)** for threats compared to traditional security tools. The Chronicle's ability to detect threats early in the attack chain, often before they cause significant damage, is further evidenced by a **92% reduction in dwell time** (the time an attacker remains undetected in a network) observed in some customer environments. (Google, https://cloud.google.com/chronicle/docs/overview, 2022)

2. **Accelerated Investigations:**

- The unified data lake and powerful search capabilities offered by the Chronicle can significantly accelerate investigations. According to Google, analysts using Chronicle can complete investigations **up to 10 times faster** compared to manual processes. The automated incident triage feature helps security teams focus on the most critical threats first. This prioritization leads to a **reduction in alert fatigue** and enables faster response times.

3. **Proactive Threat Hunting:**

- While concrete statistics on the impact of AI-driven proactive threat hunting in Chronicle are limited, industry studies indicate that organizations using AI for threat hunting

experience a **25% reduction in the overall cost of breaches**. (IBM, 2024). Behavioral analytics and predictive capabilities enabled by AI can also contribute to a **decrease in the number of successful attacks** by identifying and mitigating threats before they cause significant harm.

Overall Impact:

AI's integration into Google Chronicle has resulted in measurable improvements in threat detection, investigation, and response. The platform's AI-powered capabilities translate into faster detection times, accelerated investigations, and proactive threat hunting, ultimately helping organizations strengthen their security posture and reduce the risk of breaches.

Limitations in Research:

- Lack of Explicit Information: The publicly available information about Chronicle doesn't delve into the specifics of its AI algorithms. This limits our ability to pinpoint the exact technologies used.
- Proprietary Nature: Some aspects of Chronicle's AI implementation might be proprietary or confidential, making it difficult to obtain detailed technical information.

3.2 Darktrace's Enterprise Immune System

Darktrace, a cybersecurity company, employs an AI-powered solution called the Enterprise Immune System that acts as an intelligent layer on top of existing firewall infrastructure.

Behavioral Analytics

At the core of Darktrace's approach is the concept of behavioral analytics. The Enterprise Immune System employs unsupervised machine learning algorithms to continuously monitor and analyze vast amounts of network traffic, user behavior, and device activity. This data is then used to build a comprehensive "pattern of life" for the entire organization, creating a dynamic baseline of normalcy for every user, device, and application within the network (Darktrace, 2017)

This self-learning AI model employs a variety of techniques, including:

- **Clustering:** Identifying groups of entities with similar behavior patterns to understand normal interactions and communication pathways.

- **Time Series Analysis:** Tracking changes in behavior over time to detect deviations from established norms.
- **Bayesian Inference:** Quantifying the probability of an event being malicious based on observed evidence and prior knowledge.

Anomaly Detection

Once the model establishes the baseline behavior, the AI model excels at detecting any deviations from this standard, irrespective of whether they correspond to any recognized threat signatures. The shortcomings of traditional signature-based security solutions against insider threats, novel attacks can be delt with such anomaly detection system, which can also pick subtle signs of compromise.

By continuously comparing established baseline with the current activities the system can detect even the slightest divergence, such as unusual behavior of a user, unexpected network traffic patterns or attempting to access sensitive information at unusual hours. The understanding of organization's unique "pattern of life" enables the system to perform such actions with maximum accuracy.

Autonomous Response

In some scenarios, the Enterprise Immune System is able to respond automatically in order to halt threats and mitigate the potential damage of an attack. These actions span from quarantining compromised devices or blocking suspicious connections to alerting security analysts (Ingenia, 2019).

This autonomous response capability is especially beneficial in situations which require immediate action like zero-day attacks or insider threats. Taking immediate action can stop the malware from spreading, reduce data exfiltration, and drastically decrease remediation time.

Real-World Impact

Darktrace's AI-powered approach has shown substantial practical impact, making the organizations capable of detecting and responding to a large variety of cyber-attacks.

Detection of Novel Attacks

The Enterprise Immune System has effectively detected and prevented a range of attacks which commonly go undetected by traditional security solutions, including:

- Zero-day exploits that leverage previously unknown vulnerabilities.
- Ransomware attacks are designed to evade signature-based detection.
- Insider threats where employees exfiltrate sensitive data or engage in malicious activity.

Reduced False Positives

Through in-depth understanding of organization's normal behavior, the number of false alerts is significantly reduced by Darktrace, hence reducing alert fatigue on security analysts. This allows them to focus on genuine threats and channel their energy to manage security incidents more proficiently.

Faster Incident Response

The system provides early detection and its response that allows security teams to detect threats faster, hence reducing the damage an attack could inflict. This ensures business continuity and risk mitigation as it significantly reduces the time it takes to contain and remediate incidents.

Darktrace's case study gives compelling evidence of the system's effectiveness:

- 92% reduction in time to investigate threats.
- 63% decrease in the number of security alerts requiring investigation.
- Detection of an average of 44 previously unknown threats per customer each week.

These statistics highlight the transformative impact that AI-powered firewalls can have on an organization's security posture, enabling them to stay ahead of evolving threats and protect their critical assets (Ismail, 2019)

This comprehensive overview of Darktrace's Enterprise Immune System provides a technical deep dive into how AI is revolutionizing firewall technology, offering a glimpse into the future of network security. By harnessing the power of AI, organizations can build more resilient and adaptive defenses against an increasingly sophisticated threat landscape.

4. AI Ethics in Security

As the integration between AI and network security frameworks deepens, ethical considerations have become paramount. The deployment of AI technologies to enhance security measures raises important questions about bias, accountability, and the implications of automated decision-making processes. This section explores these ethical concerns, which are critical to ensuring that AI-driven solutions do not inadvertently exacerbate existing vulnerabilities or create new risks.

Bias in Machine Learning Algorithms: One of the foremost ethical issues in AI applications within cybersecurity is the presence of bias in machine learning algorithms. Bias can manifest in various forms, from the data used to train algorithms to the design of the algorithms themselves. Training datasets may reflect historical prejudices or over-represent certain groups, leading to discriminatory outcomes. As highlighted how biased algorithms can negatively affect outcomes, such as the delivery of online ads, raising concerns about fairness and equity in broader applications, including cybersecurity contexts (Sweeney, 2013). If AI systems are trained on biased datasets, they may misclassify threats or overlook vulnerabilities in specific demographic segments, leading to unequal protection levels. For example, AI-driven intrusion detection systems (IDS) may be less effective at identifying attacks targeting underrepresented groups, thereby leaving them exposed to heightened risks. The issue of bias is further compounded by the complexity of the algorithms used in AI systems. Many contemporary machine learning models, particularly deep learning networks, operate as "black boxes," making it challenging to understand how they arrive at specific decisions. This lack of transparency raises ethical concerns regarding fairness and accountability. If stakeholders cannot decipher the reasoning behind an AI system’s decisions, it becomes difficult to ascertain whether the system operates impartially or if it perpetuates existing biases. Research emphasizes the necessity for explainable AI (XAI) to address these challenges, providing insights into the model's decision-making process and allowing for a more equitable evaluation of its performance across different demographics (Lipton, 2018).

Accountability for AI Decisions: Accountability is another significant ethical concern surrounding AI in cybersecurity. As AI systems increasingly make autonomous decisions in real-time, determining who is responsible for these decisions becomes increasingly complex. The need for clear accountability frameworks that define the roles and responsibilities of

stakeholders involved in the deployment of AI in cybersecurity. This is particularly critical in scenarios where an AI system fails to detect a security breach or misidentifies an attack, potentially resulting in significant consequences. For example, if an AI-based firewall erroneously blocks legitimate traffic while failing to identify a true threat, the repercussions can be damaging to both the organization and its customers. Establishing accountability mechanisms not only promotes ethical use but also enhances trust in AI systems among users and organizations. This necessitates a collaborative approach between developers, users, and regulatory bodies to formulate guidelines that clarify accountability and encourage responsible AI deployment. Moreover, the ethical use of AI in cybersecurity extends to privacy considerations. The increasing reliance on AI for threat detection often involves processing vast amounts of sensitive data, raising concerns about individual privacy rights. Organizations must navigate the delicate balance between effective security measures and respecting users' privacy. The General Data Protection Regulation (GDPR) in Europe emphasizes the importance of data protection and privacy, compelling organizations to ensure that their AI-driven solutions comply with legal and ethical standards. Implementing privacy-preserving techniques, such as differential privacy or federated learning, can mitigate these concerns by enabling AI systems to learn from data without compromising individual privacy (Dwork, 2014).

Human Oversight and Ethical AI Implementation: Human oversight plays a crucial role in mitigating the risks associated with AI in cybersecurity. Sometimes it is imperative to introduce human oversight in AI decision-making processes. Ensuring humans stay in the loop allows organizations to better evaluate what AI-driven actions mean, especially in critical situations that demand nuanced understanding and ethical considerations. This hybrid approach can help balance the efficiency of AI with the moral responsibility that comes with security management. Human oversight also provides a mechanism for identifying and correcting biases that may emerge from AI systems, ensuring that ethical principles are upheld in security practices. To effectively integrate AI into cybersecurity frameworks while addressing these ethical concerns, organizations must foster a culture of ethical AI development. This involves training data scientists and cybersecurity professionals in ethical considerations, encouraging interdisciplinary collaboration, and establishing oversight committees to monitor AI deployments. By cultivating an ethical AI culture, organizations can create an environment where ethical principles are embedded in the design and implementation of AI technologies.

Privacy Concerns: The deployment of artificial intelligence (AI) in cybersecurity have garnered significant attention as the balance between effective security measures and the protection of

individual rights becomes increasingly complex. The extensive collection and processing of personal data in AI-driven security solutions raise critical questions about user privacy and data protection. One primary concern is the potential for intrusive surveillance enabled by AI systems, which often require vast amounts of data, including sensitive personal information. This can lead to situations where individuals are monitored without their explicit consent. The implications of surveillance capitalism, where personal data is commodified for various purposes, including security, creating an environment of pervasive surveillance that infringes on individual autonomy. Additionally, the risk of data breaches poses significant privacy challenges, as organizations that aggregate and analyze sensitive information become attractive targets for cybercriminals. According to the Identity Theft Resource Center, the number of data breaches continues to rise, with millions of records compromised each year, underscoring the importance of robust security measures to protect individuals' personal information from unauthorized access and misuse (Data Breach Report, 2021). Ethical implications also extend to the principles of data minimization and purpose limitation; organizations should strive to collect only the data necessary for security purposes, adhering to regulations like the General Data Protection Regulation (GDPR), which emphasize limiting data collection to relevant and necessary information. Furthermore, the transparency of AI algorithms poses another critical concern, as many systems operate as "black boxes," making it difficult for users to understand how their data is processed and what factors influence security decisions. This lack of transparency can lead to skepticism among users, who may be unaware of how extensively their personal information is utilized. Enhancing the explainability of AI models is crucial for promoting accountability and ensuring user trust in data governance. In conclusion, privacy concerns surrounding the integration of AI in cybersecurity require careful consideration, and organizations must implement measures to ensure user privacy through transparent data practices, adherence to data protection regulations, and responsible data handling. By prioritizing privacy alongside security, stakeholders can work toward a balanced approach that protects individuals' rights while leveraging AI technologies to enhance cybersecurity defenses.

In cybersecurity, the ethical implications of AI are multifaceted and warrant careful examination. The fast penetration of AI technologies in security framework can change the way organizations secure their networks, however, these developments also raise critical ethical considerations that must be addressed proactively. One of the main concerns is bias in machine learning algorithms. When the dataset with historical prejudices is taken as an input to train these AI models, the outcomes can be discriminatory. It not only breaches equality and fairness but can also lead to inadequate security measures for certain groups, exposing them to greater risks.

Therefore, AI systems implemented in organizations need to be tested and validated with rigor before we expect them to function in a way committed to equity. As AI continues to evolve and its integration with security practices it is crucial for stakeholders to continuously engage in ongoing dialogue and work on building ethical frameworks. It requires partnership among technologists, ethicists, policy makers and end users to define the norms for ethical conduct in AI use with respect to cyber security. By doing so, we can take advantage of the potential benefits AI offers to improving cybersecurity while keeping in check against some unintended consequences that may arise because of its misuse. Through responsible implementation of AI in cyber security practices, we can realize that benefit much sooner and also make our online environment all the more equitable for everyone. The responsible integration of AI into cybersecurity practices will not only help build a more equitable digital landscape for all users but also improve security outcomes.

5. Regulatory and Compliance Challenges in AI-Powered Firewalls

The increasing integration of Artificial Intelligence into firewall technologies presents new avenues for enhancing network security but also poses complex regulatory and compliance challenges. AI-powered firewalls, by their very nature, operate with a high degree of automation, often utilizing vast amounts of data to detect and mitigate threats. These processes, however, must be carried out within the constraints of existing data protection laws and industry-specific regulations, and organizations must prepare for future regulatory trends that will likely impose stricter standards on AI applications in cybersecurity. The following section elaborates few important regulatory aspects: compliance with data protection laws, adherence to industry level standards and the readiness for upcoming regulations which may have potential implications on AI enabled firewall implementations.

5.1 Compliance with Data Protection Laws

Regulatory laws like General Data Protection Regulation (GDPR) in Europe, the California Consumer Privacy Act (CCPA) are formulated to protect the individual's privacy by regulating the collection, storage and use of personal data. Modern firewalls most of the time rely on large training dataset as input to train their algorithms in order to improve effectiveness. As this dataset often contains personal information, this causes potential compliance issues if not handled carefully considering legal requirements and compliance. Hence these data protection laws pose a significant regulatory challenge for modern AI-based firewalls.

The GDPR mandates that all entities with access to the data of EU citizens abide by strict requirements regarding minimum necessary processing, transparency and individual's right to consent. AI-powered firewalls, particularly with deep learning capabilities, may facilitate storage of large amounts of traffic metadata, and possibly some of the collected data may inadvertently include personally identifiable information (PII) like IP addresses, geolocation information or behavioral patterns of pertinent users. GDPR enforces data minimization principle, i.e. the minimal amount of data which is necessary for a specific purpose only can be collected and the processing of this data should not violate any other legal requirements. Consequently, it creates a technical and legal challenge for AI applied firewalls as their sophisticated algorithm often

needs massive data to perform effectively. As a result, organizations must strike the right balance between the need for robust security through AI and maintaining legal compliance to preserve user privacy.

In addition, the GDPR also gives people a “right to explanation” for decisions made by automated systems; yet another compliance concern with AI-driven firewalls. The firewalls most of the time make decisions autonomously, for example to block or allow network traffic based on a model behavior they learned. Yet the decision-making process of AI models, and specifically deep learning algorithms, is often a black box that makes it hard to understand why decisions were taken. This lack of transparency could potentially result in a violation if users or regulators enquire about why the firewall acted as it did, as GDPR necessitates. Furthermore, businesses must be sure that AI-run firewalls do not overstep the data protection laws when handling or storing data for longer than needed, as per regulatory retention rules. A similar problem is presented under the CCPA, which allows consumers to opt-out from their data being collected and know what is being collected. AI models are complex and involve a high volume of data, which makes it challenging to comply with these regulatory demands especially as enterprises need smart dater governance strategies to control risk.

5.2 Compliance with Industry-Specific Standards

AI-powered firewalls also need to comply with industry specific cybersecurity regulations beyond general data protection regulations. These types of regulations are used in specific industries, such as sectors like healthcare, finance and critical infrastructure which have very unique needs and risks because the consequences of a security breach in those sectors are severe. For example, in the healthcare industry the Health Insurance Portability and Accountability Act (HIPAA) places strict regulations around preserving protected health information (PHI). In such areas, AI-powered firewalls must comply with these regulations and ensure that no PHI is compromised in utilizing network traffic data for monitoring and analyzing. This creates a critical challenge because AI systems must be trained on vast amounts of data to function effectively, and even anonymized datasets can sometimes become identifiable in the process. Implementing HIPAA-compliant AI firewalls calls for robust data encryption and intelligent access control mechanisms to avoid unauthorized entry into the patient's sensitive health records.

Also, for example, the financial sector follows the Payment Card Industry Data Security Standard (PCI-DSS) among other regulations to ensure payments made are not defrauded or compromised. To meet these requirements, AI powered firewalls used in these sectors must ensure no vulnerabilities or violation of PCI-DSS standards when processing financial data. One of the major concerns is that the use of machine learning algorithms in firewalls could potentially introduce security blind spots unless models are also trained to recognize new or novel types of network traffic. Further, PCI-DSS requires an annual penetration test and periodic audits for network security vulnerabilities, which means that AI firewalls must be regularly tested for compliance with these standards. The dynamism and adaptability of AI models can complicate this process further, as the system's behavior may evolve over time, which means regular re-evaluation is required to ensure compliance with standard.

For critical infrastructure sectors, like energy or transportation for example, network security solutions must abide by the rules and regulations of governing bodies such as the North American Electric Reliability Corporation (NERC) or International Civil Aviation Organization (ICAO). Due to their critical nature and strategic importance, these sectors are more sensitive to cyber attacks and hence their strict regulation emphasizes the need for fail safe and robust security measures. While AI-powered firewalls bring modern threat detection techniques, sometimes they introduce risks related to their unpredictability. Regulatory bodies in such sectors may require additional oversight and rigorous testing of AI firewalls to ensure that they do not inadvertently disrupt operations or introduce vulnerabilities.

5.3 Future Regulatory Trends

The integration of firewalls with Ai is still relatively new, hence the regulatory landscape is still evolving rapidly. Experts around the world are constantly focused on potential risks of using such security solutions, especially in highly sensitive environments. This will lead to one of the most significant developments in terms of AI-specific legislation. For example, in the European Union, a proposed AI Act has been introduced to identify types of AI systems based on risk levels and the high-risk systems shall be subjected to more stringent regulatory requirements. AI-powered firewalls, given their importance in protecting sensitive data and critical infrastructure, are likely to be classified as high-risk under such frameworks. This would require organizations deploying AI firewalls, to take strong risk management measures and make regular audits as well as ensure the transparency in operation of their models. In order to maintain consumer trust and to avoid regulatory penalties, the organization will need to explain

how and why certain security decisions were made, even though those decisions are a result of a complex set of algorithms. Finally, rise of AI auditing and certification standards is another important trend. In the future, regulators could begin to establish audit processes for AI systems, including firewalls, to ensure that they comply with ethical and legal standards. This could involve the creation of certification programs that verify the safety, reliability, and fairness of AI models used in cybersecurity. Organizations deploying AI-powered firewalls would have to invest heavily in technical infrastructure and compliance management efforts just be able to certify their systems.

The regulatory and compliance challenges associated with AI-powered firewalls are complex and multifaceted. Compliance with data protection laws such as GDPR and CCPA require organizations to carefully manage the data processed by AI models, while industry-specific regulations impose additional requirements on the security and handling of sensitive information in sectors such as healthcare and finance. As AI continues to evolve, organizations must also prepare for future regulatory trends, including AI-specific legislation and the growing demand for transparency and accountability in automated decision-making. Navigating these challenges requires a proactive approach, combining advanced technical solutions with rigorous compliance management to ensure that AI-powered firewalls not only enhance security but also operate within the bounds of legal and ethical standards.

6. Conclusion

The ever-evolving nature of cyber threat landscape necessitates a paradigm shift in terms of network security. Traditional monolithic firewalls with rule-based detection systems seem to struggle when faced with the contemporary sophisticated and dynamic modern threats. With Advanced AI, firewalls are empowered with the ability to learn, adapt, and proactively counter even zero-day threats. Machine learning, a subset of AI, lies in the heart of these modern AI-powered firewalls, which enables systems to learn from data without explicit programming. Firewalls leverage supervised learning techniques, trained on meticulously labeled datasets containing diverse network traffic patterns. This data consists of both malicious and benign traffic, allowing the firewall's ML model to discern the subtle nuances that differentiate them (Biggio B. e., 2022). Through a process of iterative learning, with time, the model continuously refines its understanding of the traffic, effectively adapting to the ever-evolving threat landscape (Li, 2020). One of the most common ML techniques used in firewalls is anomaly detection. The model can detect deviations that might indicate malicious activity, by comparing it with the established robust baseline for "normal" network traffic behavior. This method is particularly effective in identifying novel attacks, not yet in the database, that could go undetected in traditional signature firewalls (Biggio B. e., 2022). Deep learning, a subset of ML characterized by its complex, multi-layered artificial neural networks, is also gaining traction in firewalls. DL architectures excel at processing high-dimensional network traffic data, extracting intricate patterns that might escape simpler ML models (Wang W. e., 2020). DL, for example, can uncover hidden connections between seemingly benign events that may collectively represent an advanced orchestrated attack through its ability to analyze network traffic sequences on deeper level. (Bhavsar, 2022). However, the quantity and quality of dataset heavily influence the efficacy of said ML models, as biased or insufficient data can cause model vulnerabilities and inaccuracies. A significant challenge is faced when by adversarial learning techniques, attackers deliberately craft malicious traffic to confuse the model. Ongoing research in robust and privacy-preserving training methodologies is crucial to mitigate these concerns and ensure the long-term viability of AI-powered firewalls (Papernot, 2016).

AI-powered firewalls are not just limited to mere threat detection. By integrating threat intelligence feeds, they can maintain real-time awareness of emerging attack vectors and malicious actors. This equips the system to proactively analyze and block any suspicious traffic

before it infiltrates and sabotages the network (Shafahi, 2020). Furthermore, AI can make prognosis on the patterns of the networks' behaviors to identify potential security threats, enabling preventative measures to be taken before a full-blown attack unfolds (Meng, 2020). Moreover, AI can help automate the actions required in response to these incidents and cut the time taken to mitigate threats drastically. By applying techniques like behavior analysis and correlation, the system is able to pinpoint the source of an attack and automatically block malicious IP addresses or isolate infected devices in the network (Publication, Risk Management Framework for Information Systems and Organizations, 2013). Such a quick response helps to minimize the disruption and damage to the network operations. However, to prevent unintended consequences, automated response techniques require careful consideration of potential false positives and the need of a human oversight (Biggio B. e., 2022).

A major challenge with the AI-based systems is their "black box" nature, where security analysts often lack visibility and insight into the reasoning behind the firewall's decision. This lack of transparency can hinder trust and limit the ability to fine-tune security strategies. Fortunately, the Explainable AI (XAI) offers a solution by offering interpretable explanations for the system's outputs. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) can help reveal the features within the network traffic data that most heavily influenced the model's decision (Arreche, 2023). This increased transparency fosters trust between security analysts and the AI system, empowering them to validate model accuracy. Analysts can identify errors or potential biases within the model and make corrections, by understanding the rationale behind these decisions. Moreover, by offloading routine threat detection tasks to AI, analysts can concentrate their expertise to investigating complex cyberattacks that require human insight and tailor security strategies. XAI explanations also provide valuable insights that can guide the development of more effective security policies and network segmentation strategies.

6.1 Future Directions for AI in Firewalls

This research contributes to the rapidly growing field of AI in cybersecurity by providing a comprehensive analysis of AI's role in modern firewalls, particularly focusing on the technical advancements and their implications. The findings provide a valuable insight into the effectiveness of AI techniques in enhancing network security, hence assisting both researchers and practitioners in developing more resilient defense mechanisms. This thesis can serve as a guide for organizations looking to leverage AI to enhance their network security. While the

current research highlights the promising potential of AI in firewall technology, several avenues warrant further exploration. Future research could delve into the optimization of AI algorithms for specific network environments, considering factors like network size, traffic volume, and industry-specific threats. Additionally, exploring the synergy between AI and other emerging technologies like blockchain for enhanced security, as proposed by (Dorri, 2017), could yield novel and innovative solutions.

Another key area is the development of explainable AI models for firewalls. Ensuring transparency and interpretability in AI-driven decision-making is crucial for building trust and understanding among users. Research in this area could focus on developing techniques to explain the rationale behind AI-generated alerts and actions, thereby facilitating human intervention and validation.

6.2 Recommendations

Based on the research findings, several recommendations can be made for the effective integration of AI into modern firewalls:

- Embrace Deep Learning: Leverage the power of deep learning architectures for anomaly detection and intrusion prevention.
- Explore Reinforcement Learning: Investigate reinforcement learning algorithms for adaptive and self-learning firewall systems.
- Incorporate UEBA: Integrate user and entity behavior analytics into firewalls for enhanced insider threat detection.
- Focus on Explainability: Prioritize the development of explainable AI models to foster trust and transparency.

By embracing AI as a strategic asset, organizations can significantly enhance their network security and stay ahead of the ever-evolving threat landscape.

The successful integration of AI into firewalls necessitates a collaborative effort between security researchers, firewall developers, and security analysts. Researchers must continue to develop robust and secure AI

7. References

Aggarwal, K., Mijwil, M. M., Sonia, & Al-Mistarehi, A.-H. (2022). Has the Future Started? The Current Growth of Artificial. *Iraqi Journal for Computer Science and Mathematics*, 3.

Al-Haijaa, Q. A., & Ishtaiwia, A. (2021). *Machine Learning Based Model to Identify Firewall Decisions to.* Amman 1196, Jordan: Department of Data Science & Artificial Intelligence, University of Petra.

Antonakakis, M. A. (2017). Building a dynamic reputation system for DNS. In 26th {USENIX} Security Symposium ({USENIX} Security 17).

Arreche, C. e. (2023). *Explainable AI for Network Intrusion Detection Systems: A Survey and Open Research Questions.* arXiv preprint.

Bhavsar, V. e. (2022). Anomaly Detection Using Machine Learning for Network Intrusion Detection System. *In 2022 International Conference on Intelligent Systems and Information Management (ICISIM)* (pp. 0429-0434). IEEE.

Biggio, B. D. (2022, August). *Fairness in Machine Learning: A Survey of the State of the Art.* arXiv preprint arXiv:2208.00073.

Biggio, B. D. (2022, August). *Fairness in Machine Learning: A Survey of the State of the Art.* arXiv preprint arXiv:2208.00073.

Biggio, B. e. (2022). Security and Privacy in Machine Learning for Cyber Security. In *Security and Privacy in Machine Learning for Cyber Security* (pp. 3-28).

Biggio, B. e. (2022). Security and Privacy in Machine Learning for Cyber Security. Springer, Cham.

Chakraborty, S. e. (2023). Privacy-Preserving AI for Network Security: Challenges and Opportunities. *In 2023 IEEE International Conference on Communications (ICC)* (pp. 1-6). IEEE.

Cisco. (2021). *Encrypted Traffic Analytics.*

Daniel E, & O'Leary. (2021). *Expert Systems – History, Structure, Definitions, Characteristics, Life Cycle and Applications.* CA, USA: Marshall School of Business, University of Southern California.

Darktrace. (2017). *https://darktrace.com/*. Retrieved from https://darktrace.com/: https://darktrace.com/news/enterprise-immune-system-named-ai-project-of-the-year#:~:text=Our%20brilliant%20engineering%20team%20has,on%20live%2C%20complex%20networks%2C%E2%80%9D

(2021). *Data Breach Report.* Identity Theft Resource Center.

Dorri, A. K. (2017). Blockchain for IoT security and privacy: The case study of a smart home. *Proceedings of the 2nd ACM International Workshop on IoT Privacy, Trust, and Security (IoTPTS '17)*, (pp. 61-66).

Dwork, C. &. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, pp. 9(3–4), 211-407.

Feng, J. L. (2020). Microsegmentation with AI: A Novel Approach for Dynamic Policy Enforcement. *Journal of Cybersecurity and Privacy*, 3(4), 123-137.

Firewall.cx. (2023). *Firewall Performance Tuning.*

Google. (2022). *https://cloud.google.com/chronicle/docs/overview*. Retrieved from https://cloud.google.com: https://cloud.google.com/chronicle/docs/overview

Google. (2022). *https://cloud.google.com/security/products/security-operations?hl=en*. Retrieved from https://cloud.google.com: https://cloud.google.com/security/products/security-operations?hl=en

Gupta, R. &. (2021). AI and Firewall Rule Optimization: Towards Scalable Microsegmentation. *IEEE International Conference on AI and Security*, 12(5), 56-64.

https://sada.com. (2023, 10). Retrieved from https://sada.com: https://sada.com/blog/solve-cloud-security-mysteries-crack-cases-with-google-chronicle-and-mandiant/

IBM. (2024). *Cost of a Data Breach.* IBM.

Ingenia. (2019). An immune system for computers. *Ingenia*.

Institute., S. (2019). *Zero-Day Attacks.* SANS Institute. .

Ismail, N. (2019, September 4). *https://www.information-age.com/*. Retrieved from https://www.information-age.com/: https://www.information-age.com/darktrace-cyber-ai-analyst-faster-response-threats-14560/#:~:text=With%20transformational%20implications%20for%20the,in%20the%20time%20required%20to

Junyan, L., & Yoohwan, K. (2022). *Evolution of Firewalls: Toward Securer Network Using Next Generation Firewall.* Las Vegas, NV, USA: IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC).

Kende, M. (2016). *Global internet report 2016.* Internet Society.

Lashgari, A. e. (2022). *Interpretable Machine Learning: A Guide for Product Managers and Data Scientists.* arXiv preprint arXiv:2205.06078.

Li, S. e. (2020). *A Survey of Machine Learning Applications in Network Security.* arXiv preprint arXiv:2003.11131.

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Communications of the ACM*, 61(3), 36-43.

Liu, D. L. (2020). A reinforcement learning based adaptive intrusion detection framework for software defined networks. *In IEEE Transactions on Network and Service Management*, 2309-2322.

M. Mijwil, M., & E. Salem, I. (2020). Credit Caqird Fraud Detection in Payment Using Machine Learning Classifiers. *Asian Journal of Computer and Information Systems*, 50-53.

Meng, W. e. (2020). *A Survey of Machine Learning Techniques for Network Security Prediction.* IEEE Access.

Mirsky, Y. D. (2018). *Network and Distributed System Security Symposium (NDSS).*

Mohd Zafran Abdul Aziz, Muhammad Yusoff Ibrahim, Abdullatif M. Omar, & Ruhani Ab Rahman. (2012). *Performance analysis of application layer firewall.* Bandung, Indonesia: IEEE.

Oros, A. (2016). What is SNMP and how it works. *Network World*.

P. L. S. Jayalaxmi, Rahul, S., Kumar, G., Conti, M., & Kim, T.-H. (2022). *Machine and Deep Learning Solutions for Intrusion Detection and Prevention in IoTs: A Survey.* Los Angles: IEEE.

Pandey, S. R. (2010). Network Topology Discovery Techniques. *International Journal of Computer Applications*, 1-5.

Papernot, N. e. (2016). Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. *In Proceedings of the 2016 Symposium on Security and Privacy (SP)* (pp. 146-161). Symposium on Security and Privacy (SP).

Privacy-Preserving AI for Network Security: Challenges and Opportunities. (2023). In 2023 IEEE International Conference on Communications (ICC).

Publication, N. I. (2013). *Risk Management Framework for Information Systems and Organizations.* National Institute of Standards and Technology [Niˇst] Special Publication.

Publication, N. I. (2013). *Risk Management Framework for Information Systems and Organizations.* National Institute of Standards and Technology [Niˇst] Special Publication.

Q. A. Al-Haija, M. Smadi, & S. Zein-Sabatto. (2020). “Multi-Class Weather Classification Using ResNet-18 CNN for Autonomous IoT and CPS Applications. *IEEE 7th Annual Conference.* Las Vegas, USA: IEEE.

R. A. Muzaki, O. C. Briliyant, & M. A. Hasditama. (2020). Improving Security of Web-Based Application Using ModSecurity and Reverse Proxy in Web Application Firewall. *International Workshop on Big Data and Information Security (IWBIS)*, 85-90.

S. N. Abd, Alsajri, M., & R. Ibraheem, H. (2020). Rao-SVM Machine Learning Algorithm for Intrusion Detection System. *” Iraqi Journal for Computer Science and Mathematics*, 23–27.

Scarfone, K. &. (2007). *Guide to Intrusion Detection and Prevention Systems (IDPS).* NIST Special Publication.

Shafahi, A. e. (2020). AI-Driven Network Security: Leverage Machine Learning for Advanced Network Threat Detection. *2020 IEEE International Conference on Electro Information Technology (eit)* (pp. 0821-0826). IEEE.

Sweeney, L. e. (2013). Discrimination in Online Ad Delivery. *Communications of the ACM*.

Today, I. P. (2021). What is an ASIC? *IT Pro Today*. Retrieved from IT Pro Today: IT Pro Today

Utsav Upadhyay, A. K. (2023). Defending the Cloud: Understanding the Role of Explainable AI in Intrusion Detection Systems. *16th International Conference on Security of Information and Networks (SIN).* IEEE.

Wang, W. e. (2020). *Applications of Deep Learning in Network Security.* IEEE.

Wang, Y. Z. (2022). *AI-Driven Network Segmentation for Dynamic Security Management.* IEEE.

Xu, W. e. (2023). *Adversarial Attacks on Machine Learning for Network Intrusion Detection: A Survey.* arXiv preprint arXiv:2301.09452.

Yin, M. e. (2020). A Survey on Explainable Artificial Intelligence for Network Security. *Journal of Network and Computer Applications*, 161.

Yuan, S. Z. (2019). *User and entity behavior analytics for insider threat detection: A survey.* ACM Computing Surveys (CSUR).