ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

International Journal for Electronic Crime Investigation

(IJECI)

VOL: 7
ISSUE: 2 Year 2023

Email ID: ijeci@lgu.edu.pk

Digital Forensics Research and Service Center
Lahore Garrison University, Lahore, Pakistan.

International Journal for Electronic Crime Investigation
Volume 7(2) Year (2023)

SCOPE OF THE JOURNAL

The International Journal for Crime Investigation IJECI is an innovative forum for researchers, scientists and engineers in all the domains of computer science, white Collar Crimes, Digital Forensics, Nano Forensics, Toxicology and related technology, Criminology, Criminal Justice and Criminal Behaviour Analysis. Moreover, the scope of the journal includes algorithm, high performance, Criminal Data Communication and Networks, pattern recognition, image processing, artificial intelligence, VHDL along with emerging domains like quantum computing, IoT, Hacking. The journal aims to provide an academic medium for emerging research trends in the general domain of crime investigation.

SUBMISSION OF ARTICLES

We invite articles with high quality research for publication in all areas of engineering, science and technology. All the manuscripts submitted for publication are first peer reviewed to make sure they are original, relevant and readable. Manuscripts should be submitted via email only.

To submit manuscripts by email with attach file is strongly encouraged, provided that the text, tables, and figures are included in a single Microsoft Word/Pdf file.

Contact: For all inquiries, regarding call for papers, submission of research articles and correspondence,kindly contact at this address:
IJECI, Sector C, DHA Phase-VI Lahore, Pakistan
Phone: +92- 042-37181823
Email: IJECI@lgu.edu.pk

International Journal for Electronic Crime Investigation
Volume 7(2) Year (2023)

CONTENTS

International Journal for Electronic Crime Investigation
Volume 7(2) Year (2023)

Editorial

The Weakest Link in Cyber Security

Kaukab Jamal Zuberi

An organization can have the best IT security in the world but without user awareness about cyber threats and their safe behaviour, it can ultimately lead to failure. End users or "humans" are the weakest link in the security chain at any level. To further strengthen the cause of better end user awareness, research conducted by the University of Toronto which showed clearly that increased user awareness decreases the risk of cyber-attacks.

A cyber-attack is an event in which one or more persons gain unauthorized access to information systems. It can result in information damage, data theft, hardware theft, and temporary or permanent closure of the victim's operations. Reported statistics show that only small part of the victims is able to detect the attack and significantly less people take necessary actions to counteract it.

Cyberattacks can be one of the biggest threats to any business, no matter how small or large. While businesses need to protect themselves from potential cyber attacks by investing in the right security measures, another important factor to consider is ensuring that the members of staff are aware about the issues of cybersecurity and adhering to security protocols. Otherwise, chances are there that vulnerabilities will remain within an organization and it is definitely something which should not be overlooked.

Every time we open our web browsers we take risks, but some websites ask too much. At first sight it may seem just a small issue but it holds a serious risk for the users' systems.

Cybercriminals are now targeting employees on a larger scale and the damage due to corporate espionage has seen a steady increase. Recent reports say that cyber attacks via the supply chain are also increasing, where malicious code is hidden in hardware or software with the help of insiders. This means that end users are becoming increasingly important targets for cybercrime groups. The latest Symantec Internet Security Threat report shows that malicious actors have started to focus on these employees who handle critical business data and sensitive information. Cybercriminals use different methods to subvert them. For instance, they send fake emails claiming to be from the CEO's office asking for money transfers or login credentials or steal sensitive data on USB flash drives dropped into the company mailboxes.

Protecting your network is hard and is one of the major reasons why companies can't effectively protect their own networks is because

they don't have enough supporting technologies in place to do so. Cyber attackers are too sophisticated, even for most large cloud providers, so relying on vendors alone to prevent cyber attacks can be a big mistake. Therefore, it's important to fully understand what you're up against and what you can do to prevent attacks from striking your infrastructure in the first place. But achieving this goal is easier said than done...

There has been a lot in the news recently about computer programmers, IT consultants (both from U.S. and abroad) and I.T. specialists being placed on H-1B temporary work visas or permanent residencies by American companies so they can meet increased demand for their services. This has been dubbed the "brain drain" as these are positions that could be filled by U.S. citizens if not for a lack of interest in the field of computers and information technology, the high cost of education for these fields, and the fact that skilled individuals are outsourced from their home country for MUCH less than an American citizen would make, if not paid as an illegal alien (a common practice).

We have all heard the news about cyber attacks on organizations and individuals with the intention of stealing money, or personal information. The goal is to collect all the data on one system so that they can make money off of it. Over a period of time, monitoring cyber threats has become very important. The weakest link in a chain is the most vulnerable to breakage and the same applies to organizations which are susceptible to cyber attacks when it comes to user behaviour. Users are the weakest link in cyber security and hence proper training on email security should be at priority for anyone who uses email regularly even if you don't believe your employees can fall prey to such attacks.

The increasing need to raise awareness for creating cyber security awareness among the end users has led people to ignore this as an aspect that does not concern them. The cons of this attitude have made their way to the fore in a big way because of increased cyber attacks in the world.

According to reports published by government cyber authorities, over 80 percent of cyber attacks worldwide are attributed to the negligence of system administrators, with users being the weakest link in terms of cybersecurity. Unfortunately, in Pakistan, there is a significant lack of awareness among end users, posing a considerable risk to vital institutions and commercial organizations that form part of the critical infrastructure. To address this issue, the government should implement impactful strategies, policies, and incentives to enhance user awareness in both the public and private sectors. By doing so, the country's critical infrastructure can be effectively safeguarded.

Sarwar *et al. (IJECI)* 2023

(IJECI)
ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

International Journal for
Electronic Crime Investigation

DOI: https://doi.org/10.54692/ijeci.2023.0702151

Research Article

Vol. 7 issue 2 Year 2023

Optimizing Virtualization for Client-Based Workloads in Cloud Computing

Muhammad Imran Sarwar

Department of Computer Science & IT, The Superior University, Lahore, Lahore-54500, Pakistan

Corresponding author: info@imranchishty.com

Received: February 25, 2023; Accepted: March 09, 2023; Published: June 15, 2023

Abstract

Cloud computing has transformed the IT field by offering adaptable and versatile resources to cater to the increasing demands of businesses and organizations. Virtualization technologies are instrumental in facilitating the efficient deployment and management of resources within cloud environments. However, there are notable concerns regarding the security implications of virtualization in the cloud. This research paper thoroughly examines the security aspects of virtualization technologies in cloud computing, primarily focusing on identifying potential weaknesses, risks, and strategies to mitigate security threats. Additionally, the study investigates the security features and mechanisms provided by leading virtualization platforms and management tools. It scrutinizes access controls, isolation methods, network security, data protection, and integrity mechanisms offered by virtualization technologies to safeguard the cloud infrastructure and customer data. Furthermore, the paper addresses emerging security concerns associated with containerization technologies, encompassing vulnerabilities related to container escape, risks stemming from shared kernels, and issues concerning image integrity. It explores the effectiveness of container security measures, such as isolation, sandboxing, and access controls, in reducing these risks. Lastly, the paper summarizes the main findings and provides recommendations to enhance the security of virtualization technologies in cloud computing. It emphasizes the importance of continuous monitoring, regular security updates, robust access controls, and threat intelligence integration to mitigate security risks and uphold a secure cloud infrastructure.

Keywords: Cloud Computing, Virtualization Technologies, Security Analysis, Vulnerabilities, threats countermeasures, Hypervisor, Containerization, Access Controls.

1. Introduction

Cloud computing is a technology that enables users to access various services, applications, and data storage. It functions as a pool of resources characterized by two main features. The first characteristic is elasticity, which allows users to adjust and allocate

resources according to their specific needs dynamically. The second feature is multi-tenancy, which enables multiple users to access and store data on the same shared resources [1]. The primary goal of Software as a Service (SaaS) is to facilitate efficient enterprise or site search on the Internet. Quick and accurate access to information from databases and internal storage via website content is crucial in fast-paced organizations. As a search technology branch, SaaS provides significant benefits to various companies and external customers, catering to their specific requirements. Users access SaaS through a web browser, allowing them to manage, store, and process essential resources like software, operating systems, and applications in a cloud-based environment [2]. Platform as a Service (PaaS) allows clients to deploy and utilize diverse applications using programming languages, libraries, services, and tools to assist users.

Public cloud environments offer users easy access to computing resources, including hardware components (operating systems, central processors, memory, storage) and software (application servers, services). These public clouds primarily serve the purpose of application development and testing. In contrast, although more expensive than public clouds, private clouds provide an ideal solution for addressing security and privacy concerns within organizations [3]. Community clouds, unlike public clouds, provide cost-effective access without additional expenses. They enable multiple organizations to share computing resources. Hybrid cloud models combine private and public infrastructures and are often adopted by organizations to manage their IT infrastructure effectively [4]. Hybrid cloud architecture offers flexibility and cost-efficiency, making it a preferred choice for businesses and customers. Figure 1 illustrates the cloud model.

Figure 1: Cloud service models

Virtualization is a vital component in cloud computing, as it allows for isolating resources and services from physical infrastructure. Organizations are increasingly acknowledging the importance of cost efficiency and environmentally sustainable practices in their operations. Virtualization provides advantages such as increased capacity and initial cost savings. However, it also brings about high-security concerns. Figure 2 depicts a conceptual virtualization model.

Figure 2: Virtualization model

As virtualization gains practicality, new advancements and technologies emerge, each with advantages, disadvantages, and risks. Implementing these emerging virtualization technologies often poses challenges for project administrators and implementers. Virtualization encompasses various applications and executions beyond specific and centralized server systems. In today's context, virtualization is widely supported on readily available systems utilizing Intel architecture hardware. This is made possible by Intel Virtualization Technology, which offers hardware support for processor virtualization and facilitates advancements in virtual machine (VM) monitoring software. Consequently, the resulting virtual machine monitors (VMMs) can accommodate a broader range of legacy and future operating systems while maintaining high performance. Virtualization can be applied to hardware and software, and the progress in virtualization technology continues to evolve. Many organizations are adopting virtualization due to its cost-saving benefits, but it is essential to evaluate the associated risks. While server virtualization receives significant attention in the industry and literature, other areas of virtualization also need consideration [5].

1.1. Types of Virtualization

1. Storage virtualization involves consolidating physical storage from multiple devices into a centralized storage pool managed through a central console. This pooling of storage capacity enables software programs to identify available storage from physical devices and aggregate it into a virtual storage environment accessible by VMs. The virtualized storage appears as a unified storage entity, allowing read and write operations. In certain cases, even a RAID cluster can be considered a form of storage virtualization [6].
2. Storage virtualization offers numerous advantages, including increased productivity and enhanced security. It enables remote access, allowing users to work from any location and on any computer. This flexibility provides employees with convenience and resilience, enabling them to work from home or while on the move [7]. Additionally, storage virtualization helps protect sensitive data by storing it on a central server, minimizing the risk of loss or theft [8].
3. Server Virtualization: Server virtualization is the most widely recognized form of virtualization in the cloud. It offers benefits such as improved hardware utilization and application uptime. The main concept behind server virtualization is to combine multiple smaller physical servers into a single larger physical server to utilize the processor more effectively. Server virtualization can be further categorized into the following types:
4. Full Virtualization: In this type, the complete emulation of the real hardware allows the software to run an unmodified guest operating

system.

5. Paravirtualization: The software runs in a modified operating system as a separate system, but not in an unmodified form.
6. Partial Virtualization: This hardware virtualization may require software modifications to function.
7. Network Virtualization: This technology offers a virtual infrastructure that simplifies the administration of software and hardware resources within a network. It can be categorized into two types: external virtualization, which combines or divides networks into virtual units, and internal virtualization, where software incorporates network functionalities. External network virtualization is also referred to as virtual LAN [3].

1.2. Hypervisor

A hypervisor, a VMM, is a technology used to create and run VMs. It enables multiple operating systems to share a single host system and its hardware resources. The hypervisor, also called software virtualization, is responsible for partitioning and allocating resources such as CPU and RAM on the hardware [9]. There are two types of hypervisors:

1. Type 1: Native/Bare Metal Hypervisor: This hypervisor is installed directly on bare-metal hardware, functioning as a software layer. It operates independently and does not require a different operating system. Some level of external management is needed to oversee its operation.
2. Type 2: Hosted Hypervisor: This hypervisor is installed within an operating system. It runs as a software application on the host operating system, providing virtualization services. Examples of hosted hypervisors include VirtualBox and VM Workstation.

1.3. Virtualization Techniques

This technique depends on paired interpretation to trap just as to virtualize certain touchy and non-virtualizable guidelines with new arrangements of directions that have the proposed impact on the virtual equipment. The binary image is controlled at the runtime, and User level code is straightforwardly executed on the processor for superior virtualization. The mix of parallel interpretation just as immediate execution gives Full Virtualization as the visitor OS is decoupled from the essential equipment by the virtualization layer. Paravirtualization primarily alludes to correspondence between hypervisor just as visitor OS to improve productivity and execution. Paravirtualization additionally includes changing the OS part to supplant non-virtualizable Guidelines with hyper calls, which discuss straightforwardly with the virtualization layer hypervisor. Memory Virtualization. It familiarizes a course by decoupling the specialist's memory to give a passed-on, shared, or organized limit. It improves execution by giving more impor-

tance to memory limits without extension to the essential memory [10].

2. Related Work

[11] Storage virtualization refers to combining physical storage from multiple devices into a unified storage pool managed through a central control system. This allows software programs to recognize available storage from physical devices and merge it into a virtual storage environment that VMs can access. The virtualized storage is a single entity, enabling read and write operations. In some cases, even a RAID cluster can be considered a form of storage virtualization.

[12] The advantages of storage virtualization are significant, including increased productivity and enhanced security. It enables remote access, allowing users to work from any location and computer. This flexibility provides convenience and resilience for employees working remotely or while on the move [8]. Additionally, storage virtualization ensures the security of sensitive data by storing it on a central machine, minimizing the risk of loss or theft.

[13] Storage virtualization is the process of merging physical storage from various devices into a centralized storage pool controlled through a central console. This consolidation of storage capacity enables software applications to recognize and merge available storage into a virtualized storage environment that can be accessed by VMs. The virtualized storage is perceived as a unified entity, enabling read and write operations. In some instances, a RAID cluster can also be regarded as a type of storage virtualization.

[14] Storage virtualization offers numerous benefits, such as improved productivity and heightened security. It enables remote access, allowing users to work from any location and using any computer. This flexibility provides employees with convenience and adaptability, allowing them to work effectively from home or while traveling. Moreover, storage virtualization guarantees safeguarding sensitive data by storing it on a centralized machine, thereby minimizing the potential risks associated with data loss or theft.

[15] The presence of expensive and proprietary equipment, along with strict signaling protocols, poses challenges for current mobile core networks. When specific functionality is lacking, mobile operators are required to replace their hardware, even if it is sufficient for most purposes. This highlights the difficulty of implementing Network Function Virtualization (NFV) and emphasizes the need for dynamic designs to create and manage network capabilities. NFV's core concept revolves around deploying Virtualized Network Functions (VNFs) in a deployment diagram. The virtualization of the mobile core network using Cloud EPC can address the issue of costly control and maintenance of long-distance persistent tunnels for mobile operators. Technologies like MME pooling facilitate this approach. It's important to note that only a portion of the mobile core network can be virtualized if desired. Cloud EPC enables a

transition to a more intelligent, resilient, and scalable core architecture. By leveraging Cloud EPC, mobile carriers can expand their existing horizontal market business and explore vertical markets that were previously untapped. Service providers have the opportunity to offer home services through dedicated Customer Premises Equipment (CPE) supported by network-centric back-end systems.

[16] The VMM needs to offer a software interface to the VM that closely mimics the underlying hardware. However, it should also retain control over the machine and the ability to intervene in hardware access when necessary. When assessing these factors, the main design objectives for VMMs are compatibility, performance, and simplicity. Compatibility is of utmost importance because the key advantage of a VMM is its capability to run legacy software. Performance, which evaluates the effects of virtualization, aims to ensure that the VM operates at the same speed as software on real hardware.

[17] Virtualization improves the efficiency, ease of management, and reliability of centralized computer systems. It enables multiple users with different operating system requirements to effectively share a virtualized server. By coordinating operating system updates across VMs, downtime can be minimized. Additionally, failures in guest software are isolated to the specific VMs in which they occur. While these advantages have traditionally been associated with high-end server systems, recent academic research and the emergence of VM-based products indicate that the benefits of virtualization are applicable to a wider range of server and client systems. This cloud model emphasizes availability and encompasses five key attributes, three service models, and four deployment models.

[18] During the initial adoption of VMs, it was typical for a single organization to develop the VMM, hardware, and guest operating systems. These vertically integrated companies allowed experts to refine traditional virtualization techniques. One approach involved modifying guest operating systems to provide higher-level information to the VMM, taking advantage of the flexibility in the VMM/guest operating system interface. Another approach focused on exploiting the flexibility in the hardware VMM interface to enhance traditional VMMs. The VMM stored much of the privileged state of the guest in a hardware-defined structure and executed the SIE instruction to initiate interpretive execution. During interpretive execution, many guest operations, which would normally trap in a non-privileged environment, accessed shadow fields. Virtualizing the memory management unit (MMU) presented many complex scenarios when analyzing hardware VMMs. In this section, we explore future approaches in both hardware and software to bridge the performance gap with software VMMs. As hardware implementations advance, the overheads associated with hardware virtualization will diminish over time. Measurements were conducted on a desktop system using. It is important to note that for clarity, we have treated the software and hardware VMMs as separate entities. However, in VMware, both VMMs are part of

the same binary. We have conducted experiments with a hybrid VMM that dynamically selects the execution method based on heuristic algorithms driven by guest behavior. The goal is to leverage the superior system-level performance of the hardware VMM and the superior MMU execution of the software VMM.

[19] In NFV organizations, the task of assigning various administration chains to the physical network is essential. An administration chain comprises one or more services or virtual machines (VMs) interconnected to fulfill specific functionalities. These administration chains can be assigned in a hybrid network environment, utilizing either physical hardware or virtualized instances. When a service request is made, it can be allocated on dedicated hardware or through a VM provided by the service provider. Additionally, client VMs can be integrated into an administration chain. The main distinction between service and VM requests is that the client who initiated the service chain request manages the VM within the administration chain. Resource allocation for the first two deployment types resembles network-aware VM allocation in cloud environments, with the only difference being the consideration of CPU and memory requirements for VM deployment.

[20] In order to improve the networking performance of scaling up VMs, it is crucial to identify the specific system component that is causing limitations. To achieve this, we initiate an evaluation process that scrutinizes the four main systems involved. Through virtualization of the cloud platform, we enhance the availability of resources and enhance the flexibility of their management. This approach also leads to cost reduction by enabling hardware multiplexing and improves energy efficiency.

[21] The online nature of the cloud system exposes it to security issues commonly found on the web. Despite its differences from traditional computer systems, the cloud system can encounter similar security challenges. Security and data protection are major concerns in cloud computing. Traditional security issues like vulnerabilities, viruses, and hacking attacks can jeopardize the integrity of the cloud system and can have more severe consequences due to the nature of cloud computing. Unauthorized access by hackers and malicious intruders can compromise cloud accounts and steal sensitive data stored within cloud systems. Since data and business applications reside in the cloud center, it becomes crucial for the cloud system to implement robust security measures.

[22] A challenging and emerging development in cloud computing and data center architectures is NFV. As hosted applications in cloud systems increasingly have complex networking requirements, providers seek greater flexibility in managing the underlying infrastructure. This flexibility necessitates the redistribution of VMs, applications, and data storage based on the real-time status of the system. The complexity of this task requires advanced management techniques and the adoption of software-based approaches, particularly highly scalable and dynamic network virtualization methods. Network virtualization entails the separation of functionalities within

a networking environment by dividing the roles of Internet Service Providers (ISPs) into two components: infrastructure providers responsible for managing the physical infrastructure, and service providers responsible for creating virtual networks by pooling resources from multiple infrastructure providers and offering comprehensive connectivity services.

[23] Scaling resources manually by a human administrator may seem simple. Still, it is not a viable option considering the increasing cloud size and multiple web services sharing the same infrastructure and data. Automating resource allocation for web services becomes crucial, considering performance history, issues, SLA requirements, and resource security when scaling up or down. Such a system can rely on AI techniques to efficiently determine the required resources for the service. One commonly used machine learning algorithm is Support Vector Machines (SVM), employed in tasks like pattern recognition, spam filtering, and anomaly network intrusion detection. SVM can learn patterns and provide accurate classification by utilizing class labels. It finds the optimal global solution by finding a hyperplane that separates two classes. The data points closest to the hyperplane are known as support vectors, and based on their features, the predicted class is determined. In this study, a novel clustering selection algorithm has been proposed for collaborative range sensing. The algorithm focuses on selecting the most reliable cluster heads that can transmit their sensing decisions to the aggregation center. The scheme's performance has been evaluated by analyzing energy consumption and transmission delay, comparing it to the conventional model. Analytical and simulation results demonstrate the superior performance of the proposed method across different trust level values compared to the conventional model.

[24] Organizing unknown documents using ML techniques can be divided into two subsequent stages: training and testing. In the first stage, a designated set of documents, known as the training set, is provided to the system. Each document is then parsed, and a vector representing the document is extracted based on a predetermined vocabulary. These representative vectors and the corresponding known labels serve as input to a learning algorithm. By training on these vectors, the learning algorithm generates a classification model. To assess the performance of the computer processor, we utilize the sysbench stress test. This test is designed to challenge the central processor by calculating prime numbers. The algorithm divides the number using progressively increasing numbers and verifies that the remainder (estimate) is zero. In this particular case, we examine how power consumption increases as the number of virtual elements assigned to different physical cores increases, as explained in the context of computer pinning. This research paper provides a virtualization performance model for IT managers to follow before implementing Virtualization technology in their data centers.

[25] Virtualization allows existing operating systems to run on shared-memory multiprocessors. VMs can create diverse testing environments, facilitating innovative and effective quality assurance processes. Additionally,

virtualization can be leveraged to introduce new features to existing operating systems with minimal effort. It simplifies tasks such as system migration, backup, and recovery, making them more manageable and cost-effective. Virtualization provides a practical approach for achieving parallel compatibility across various hardware and software platforms, enhancing coherence among different aspects of the virtualization process.

[26] The cloud infrastructure is distributed among multiple instances operating within the cloud. When many VMs run on the same node, their performance can be negatively impacted. This is particularly evident when multiple machines are simultaneously utilizing the network, managing a significant volume of data. Similarly, access to the hard drive is shared among the virtual entities, and in the Amazon cloud, disk access, and network usage are treated as shared resources. The utilization of Mobile Edge Computing (MEC) systems also presents opportunities for further research in developing new services and applications to enhance network efficiency and improve the user experience. MEC can be leveraged to maintain network or service states for emerging applications, such as ensuring scalability by preserving critical parameters and providing backup support.

3. Proposed Methodology

The proposed methodology of this study is depicts in the Figure 3.

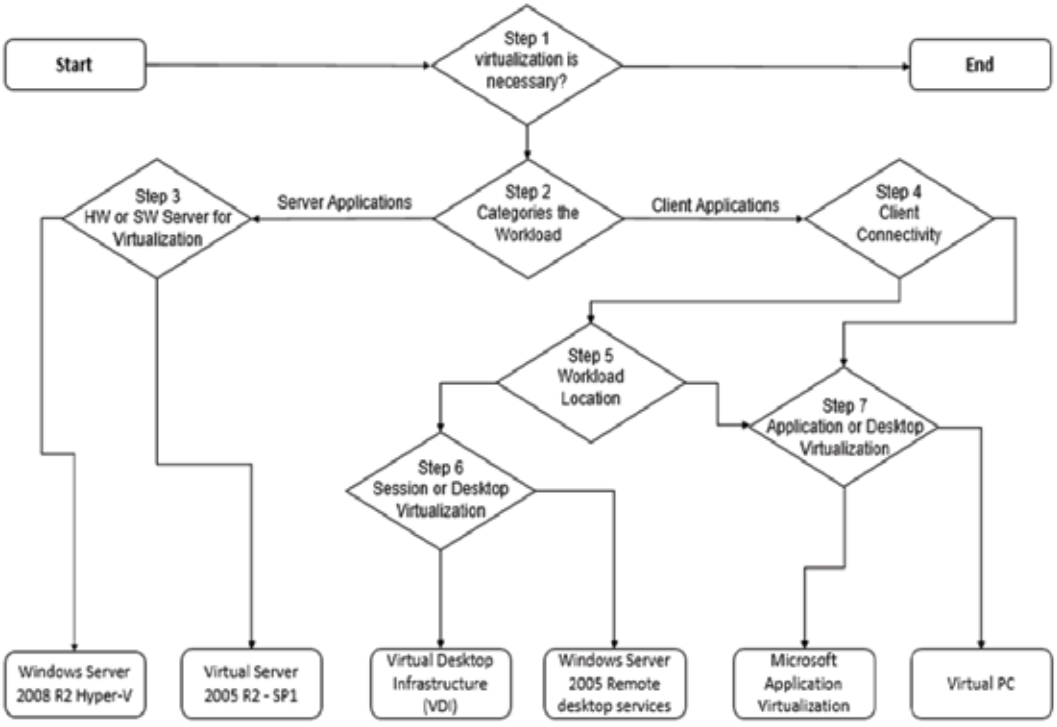


Figure 3: Research Methodology

Step 1: Is Virtualization Necessary:

This step aims to assess virtualization's suitability for a specific scenario. Several factors should be considered when considering virtualization to determine its compatibility with the needs. Compatibility refers to the ability of different components to work together [27]. Verifying whether the workload can be effectively executed in a virtualized environment is important. Workloads can include applications running on the client, server, or a combination. Susceptibility is another factor to consider. It involves evaluating whether the workload can run in a virtualized environment. It is advisable to review the support policies of non-Microsoft vendors to ensure compatibility across different virtualization technologies. Licensing is an important aspect to examine. It involves verifying if the necessary licenses are available to use the workload in a virtualized environment. Furthermore, it is crucial to determine the benefits of virtualizing the workload to the business. Assessing the business case for virtualization can reveal potential advantages such as cost savings, reduced implementation time, and low management costs [28].

Step 2: Categories the Workload:

Once the decision has been made to proceed with virtualization, the next step is to classify the workload into the appropriate category [29]. This step involves determining whether the workload is designed to run on a Windows Server-based server or a client device. Server workloads have distinct resource requirements and levels of interactivity compared to workloads specifically designed for server operating systems.

Step 3: Hardware Server or Software Server for Virtualization:

Microsoft provides two server virtualization products: Hyper-V, integrated into Windows Server 2008 R2, enabling hardware virtualization for servers, and Virtual Server 2005 R2 with SP1, offering software virtualization for servers [30]. The objective of this stage is to identify the product that best aligns with the specific technological needs of the environment for establishing most appropriate virtualization framework.

Step 4: Client Connectivity:

This step involves narrowing down the virtualization options based on the network requirements of the client computers. For server-based systems, computers already connected to the network can be utilized, whereas locally available applications will need to be relied upon for those not connected [31]. It is important to note that user specifications and application requirements can vary between different workloads, so the decisions made in this and subsequent sections should be validated for each workload being virtualized.

Option 1: Connected Clients, Option 2: Disconnected Clients

Step 5: Workload Location:

The subsequent phase for interconnected client

systems involves determining the execution location for the workload. Based on the previous technology choices, the virtualized workload can operate either in a centralized or decentralized manner.

Option 1: Workload Centralization, Option 2: Workload Decentralization

Step 6: Session or Desktop Virtualization: After deciding to centralize operations, the subsequent task involves choosing between session virtualization and desktop virtualization. In both scenarios, users establish connections to centralized workloads via a Remote Desktop Protocol (RDP) connection. Client computers can operate with a complete operating system and an installed RDP client, such as Windows 7 or any other compatible operating system [32]. Alternatively, they can be diskless and boot directly from the network, without storing any local data or programs.

Option 1: Virtualization of Sessions, Option 2: Virtualization of the Desktop

Step 7: Application or Desktop Virtualization: In this stage, the suitability of the workload for either application virtualization or desktop virtualization is determined.

Option 1: Virtualization of applications, Option 2: Virtual PC

App-V enables the installation of applications through MSI or on-demand streaming into a virtualized environment. In this setup, user machines are capable of handling application processing. To deploy and execute virtualized applications, client computers must have a compatible full client operating system that meets the hardware requirements specified by App-V. Additionally, a reliable network connection is necessary [33]. Windows Virtual PC allows users to run complete client operating systems on their local computers. The client computer must have sufficient CPU, memory, disk, and network resources for the base Windows operating system and each VM utilized to support this configuration. Windows Virtual PC facilitates the execution of legacy programs and operating systems, with Windows XP Mode in Windows 7 offering a tailored Windows XP VM running on Windows Virtual PC.

4. Simulations And Results

Checklist of items to consider:

1. Let's begin with the network.
2. Give the VM a name.
3. Choose a location for the VM.
4. Determine the VM's size.
5. Recognize the price model.
6. Storage for the VM.
7. Choose a computer operating system.

Figure 4 shows the Azure services interface.

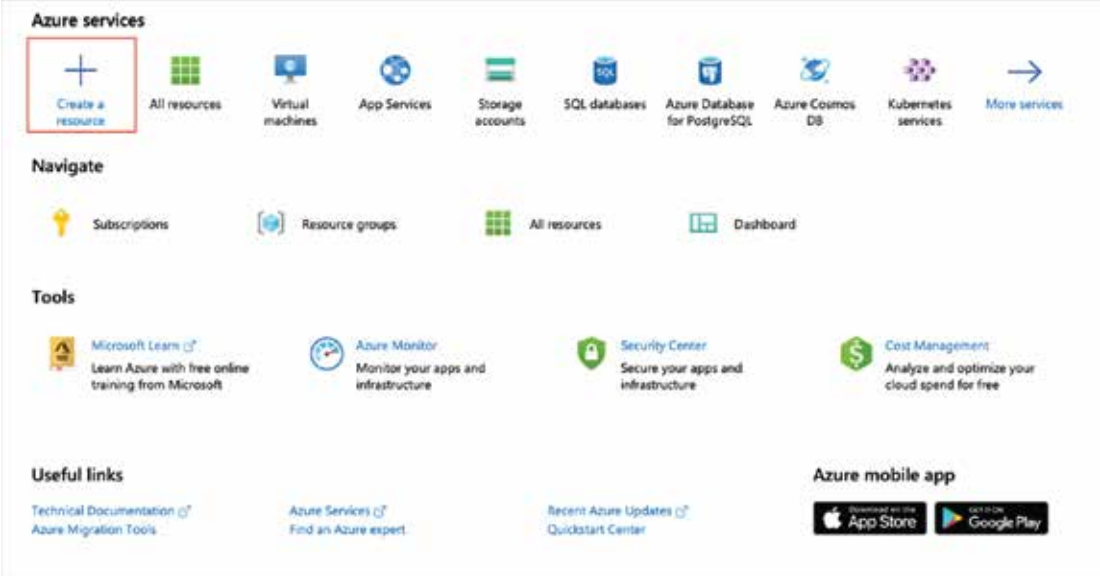


Figure 4: Azure services interface

Network: First consider the network, not the VM. Azure uses virtual networks (VNets) to provide secure connectivity between Azure VMs and other Azure services. VMs and services in the same virtual network can communicate [34].

Divide Network into Sections: Construct one or more subnets for the virtual network after choosing the virtual network address space(s). This allows to divide the network into more manageable portions. For example, that may give VMs 10.1.0.0, back-end services 10.2.0.0, and SQL Server VMs 10.3.0.0.

Keep the network Safe: By default, there is no inherent security boundary between subnets, allowing unrestricted communication between services. However, it is possible to establish Network Security Groups (NSGs) to control the flow of traffic to and from subnets and virtual machines (VMs) [35]. NSGs function as software firewalls at the network interface, enabling the application of customized rules to regulate inbound and outbound requests.

Plan for Each VM Deployment: Start with VMs that is to be built once it has mapped out with communication and network requirements. Selecting a server and taking an inventory is a good idea:

1. With whom does the server communicate?
2. What ports are available?
3. What operating system is being used?
4. How much disc space is currently occupied?
5. What kind of information is used in this? Are there any legal or other ramifications to not having it on-premises?
6. What is the server's CPU, memory, and disc I/O load like? Is there a need to account for surge traffic?

Giving Name to VM: The VM's name is one piece of information that many people overlook. The computer name configured as part of the operating system is the VM name. On a Windows VM, it can be given a name up to 15 characters; on a Linux VM, name can be given up to 64 characters [36]. In Azure, an Azure resource is a managed entity. VM, like physical computers in datacenter, require various components to function properly:

- Storage account for the discs in the VM
- Networking over the Internet (shared with other VMs and services)
- To communicate across a network, it needs a network interface.
- Secure network traffic with Network Security Groups Public Internet address

VM Location: Azure has servers and discs in data centers worldwide [37]. To offer redundancy and availability, these data centers are divided into geographic regions ('West US,' 'North Europe,' 'Southeast Asia,' and so on) - selection of VM location as shown in Figure 5.

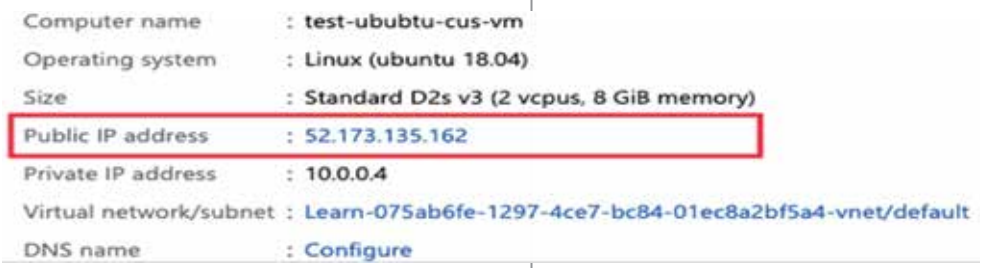


Figure 5: Location selection

Determine the VM's Size: After name and location have been decided, it will need to determine the size of the VM. Rather than specifying processor power, memory, and storage capacity separately, Azure offers a variety of VM sizes with different variants of these factors [38]. Azure offers a variety of VM sizes, allowing to choose the right combination of computation and memory.

Size Adjustment: When the current VM size no longer matches the requirements, Azure allows to adjust it. It can upgrade or downgrade the VM if the new size is compatible with the present hardware configuration [39]. This allows for a completely flexible and agile approach to VM management.

Select Price Model: The subscription will be charged two distinct costs for each VM: computation and storage. It may scale these costs independently and only pay for what is needed by separating them.

Storage for VM: All Azure VMs should have at least two virtual hard discs as a best practice (VHDs). The operating system is stored on the first disc, while temporary storage is kept on the second. Additional discs can be added to hold application data; the maximum number is governed by the VM size (typically two per CPU) [40].

Azure Storage: Microsoft's Azure Storage is a cloud-based data storage service. It can store practically any data, giving the secure, redundant, and scalable access. For a given subscrip-

tion, a storage account grants access to items in Azure Storage. Each attached virtual disc is always stored in one or more storage accounts on VMs [41].

Choosing a system: Different Windows and Linux flavors versions can be installed on the VM using Azure's OS images. As previously stated, the operating system will impact the hourly compute rate because Azure includes the cost of the OS license in the prices. It can be used the New-AzVM cmdlet to create a new Azure VM:

```
New-AzVm `
    -ResourceGroupName "TestResourceGroup" `
    -Name "test-wp1-eus-vm" `
    -Location "East US" `
    -VirtualNetworkName "test-wp1-eus-network"
    -SubnetName "default" `
    -SecurityGroupName "test-wp1-eus-nsg" `
    -PublicIpAddressName "test-wp1-eus-pubip" `
    -OpenPorts 80,3389
```

To create an Azure VM with the az vm create command:

```
az vm create \
    --resource-group TestResourceGroup \
    --name test-wp1-eus-vm \
    --image win2016datacenter \
    --admin-username jonc \
    --admin-password aReallyGoodPasswordHere
```

C# code to create an Azure VM using Microsoft. Azure. Management. Fluent NuGet package:

```
var azure = Azure
.Configure()
        .WithLogLevel(HttpLoggingDelegatingHandler.Level.Basic)
        .Authenticate(credentials)
        .WithDefaultSubscription();
// ...
var vmName = "test-wp1-eus-vm";

azure.VirtualMachines.Define(vmName)
WithRegion(Region.USEast)
WithExistingResourceGroup("TestResourceGroup")
WithExistingPrimaryNetworkInterface(networkInterface)
WithLatestWindowsImage("MicrosoftWindowsServer", "WindowsServer", "2012-R2-Datacenter")
WithAdminUsername("jonc")
WithAdminPassword("aReallyGoodPasswordHere")
WithComputerName(vmName)
WithSize(VirtualMachineSizeTypes.StandardDS1)
Create();
```

Snippet in Java using the Azure Java SDK:

```
String vmName = "test-wp1-eus-vm";
// ...
VirtualMachine virtualMachine = azure.virtualMachines()
    .define(vmName)
    .withRegion(Region.US_EAST)
    .withExistingResourceGroup("TestResourceGroup")
    .withExistingPrimaryNetworkInterface(networkInterface)
    .withLatestWindowsImage("MicrosoftWindowsServer", "WindowsServer", "2012-R2-Datacenter")
    .withAdminUsername("jonc")
    .withAdminPassword("aReallyGoodPasswordHere")
    .withComputerName(vmName)
    .withSize("Standard_DS1")
    .create();
```

5. Conclusion

In conclusion, the security analysis of virtualization technologies in cloud computing presents a comprehensive understanding of the key challenges and potential threats associated with adopting virtualization in the cloud environment. Through an in-depth examination of various virtualization techniques and their security implications, this analysis has shed light on the strengths, weaknesses, and best practices organizations should consider ensuring robust security in their cloud-based virtualized infrastructures. The analysis highlighted the benefits of virtualization, such as resource optimization, scalability, and cost-effectiveness. However, it also revealed several security concerns arising from the shared nature of virtualized environments, including the potential for information leakage, unauthorized access, and hypervisor vulnerabilities. These risks underscore the need for robust security measures to protect sensitive data and ensure the integrity and confidentiality of cloud-based services.

6. References

[1]. N. Tabassum, A. Namoun, T. Alyas, A. Tufail, M. Taqi, and K. Kim, “applied sciences Classification of Bugs in Cloud Computing Applications Using Machine Learning Techniques,” 2023.

[2]. M. I. Sarwar, Q. Abbas, T. Alyas, A. Alzahrani, T. Alghamdi, and Y. Alsaawy, “Digital Transformation of Public Sector Governance With IT Service Management–A Pilot Study,” IEEE Access, vol. 11, no. January, pp. 6490–6512, 2023, doi: 10.1109/ACCESS.2023.3237550.

[3]. T. Alyas, K. Ateeq, M. Alqahtani, S. Kukunuru, N. Tabassum, and R. Kamran, “Security Analysis for Virtual Machine Allocation in Cloud Computing,” Int. Conf. Cyber Resilience, ICCR 2022, no. Vm, 2022.

[4]. T. Alyas et al., “Performance Framework for Virtual Machine Migration in Cloud Computing,” Comput. Mater. Contin., vol. 74, no. 3, pp. 6289–6305, 2023.

[5]. T. Alyas, S. Ali, H. U. Khan, A. Samad, K. Alissa, and M. A. Saleem, "Container Performance and Vulnerability Management for Container Security Using Docker Engine," Secur. Commun. Networks, vol. 2022, 2022.

[6]. M. Niazi, S. Abbas, A. Soliman, T. Alyas, S. Asif, and T. Faiz, "Vertical Pod Autoscaling in Kubernetes for Elastic Container Collaborative Framework," 2023.

[7]. T. Alyas, A. Alzahrani, Y. Alsaawy, K. Alissa, Q. Abbas, and N. Tabassum, "Query Optimization Framework for Graph Database in Cloud Dew Environment," 2023.

[8]. T. Alyas et al., "Multi-Cloud Integration Security Framework Using Honeypots," Mob. Inf. Syst., vol. 2022, pp. 1–13, 2022.

[9]. T. Alyas, N. Tabassum, M. Waseem Iqbal, A. S. Alshahrani, A. Alghamdi, and S. Khuram Shahzad, "Resource Based Automatic Calibration System (RBACS) Using Kubernetes Framework," Intell. Autom. Soft Comput., vol. 35, no. 1, pp. 1165–1179, 2023.

[10]. G. Ahmed et al., "Recognition of Urdu Handwritten Alphabet Using Convolutional Neural Network (CNN)," Comput. Mater. Contin., vol. 73, no. 2, pp. 2967–2984, 2022.

[11]. M. I. Sarwar, K. Nisar, and I. ud Din, "LTE-Advanced – Interference Management in OFDMA Based Cellular Network: An Overview", USJICT, vol. 4, no. 3, pp. 96-103, Oct. 2020.

[12]. A. A. Nagra, T. Alyas, M. Hamid, N. Tabassum, and A. Ahmad, "Training a Feedforward Neural Network Using Hybrid Gravitational Search Algorithm with Dynamic Multiswarm Particle Swarm Optimization," Biomed Res. Int., vol. 2022, pp. 1–10, 2022.

[13]. T. Alyas, M. Hamid, K. Alissa, T. Faiz, N. Tabassum, and A. Ahmad, "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach," Biomed Res. Int., vol. 2022, pp. 1–10, 2022.

[14]. T. Alyas, K. Alissa, A. S. Mohammad, S. Asif, T. Faiz, and G. Ahmed, "Innovative Fungal Disease Diagnosis System Using Convolutional Neural Network," 2022.

[15]. H. H. Naqvi, T. Alyas, N. Tabassum, U. Farooq, A. Namoun, and S. A. M. Naqvi, "Comparative Analysis: Intrusion Detection in Multi-Cloud Environment to Identify Way Forward," Int. J. Adv. Trends Comput. Sci. Eng., vol. 10, no. 3, pp. 2533–2539, 2021.

[16]. S. A. M. Naqvi, T. Alyas, N. Tabassum, A. Namoun, and H. H. Naqvi, "Post Pandemic World and Challenges for E-Governance Framework," Int. J. Adv. Trends Comput. Sci. Eng., vol. 10, no. 3, pp. 2630–2636, 2021.

[17]. W. Khalid, M. W. Iqbal, T. Alyas, N. Tabassum, N. Anwar, and M. A. Saleem,

"Performance Optimization of network using load balancer Techniques," Int. J. Adv. Trends Comput. Sci. Eng., vol. 10, no. 3, pp. 2645–2650, 2021.

[18]. T. Alyas, I. Javed, A. Namoun, A. Tufail, S. Alshmrany, and N. Tabassum, "Live migration of virtual machines using a mamdani fuzzy inference system," Comput. Mater. Contin., vol. 71, no. 2, pp. 3019–3033, 2022.

[19]. M. A. Saleem, M. Aamir, R. Ibrahim, N. Senan, and T. Alyas, "An Optimized Convolution Neural Network Architecture for Paddy Disease Classification," Comput. Mater. Contin., vol. 71, no. 2, pp. 6053–6067, 2022.

[20]. J. Nazir et al., "Load Balancing Framework for Cross-Region Tasks in Cloud Computing," Comput. Mater. Contin., vol. 70, no. 1, pp. 1479–1490, 2022.

[21]. N. Tabassum, T. Alyas, M. Hamid, M. Saleem, S. Malik, and S. Binish Zahra, "QoS Based Cloud Security Evaluation Using Neuro Fuzzy Model," Comput. Mater. Contin., vol. 70, no. 1, pp. 1127–1140, 2022.

[22]. M. I. Sarwar, K. Nisar, and A. Khan, "Blockchain – From Cryptocurrency to Vertical Industries - A Deep Shift," in IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), September 20-23, 2019, Dalian, China, 2019, pp. 537–540. doi: 10.1109/ICSPCC46631.2019.8960795.

[23]. S. Malik, N. Tabassum, M. Saleem, T. Alyas, M. Hamid, and U. Farooq, "Cloud-IoT Integration: Cloud Service Framework for M2M Communication," Intell. Autom. Soft Comput., vol. 31, no. 1, pp. 471–480, 2022.

[24]. W. U. H. Abidi et al., "Real-Time Shill Bidding Fraud Detection Empowered with Fussed Machine Learning," IEEE Access, vol. 9, pp. 113612–113621, 2021.

[25]. M. I. Sarwar et al., "Data Vaults for Blockchain-Empowered Accounting Information Systems," IEEE Access, vol. 9, pp. 117306–117324, 2021, doi: 10.1109/ACCESS.2021.3107484.

[26]. N. Tabassum, T. Alyas, M. Hamid, M. Saleem, and S. Malik, "Hyper-Convergence Storage Framework for EcoCloud Correlates," Comput. Mater. Contin., vol. 70, no. 1, pp. 1573–1584, 2022.

[27]. N. Tabassum et al., "Semantic Analysis of Urdu English Tweets Empowered by Machine Learning," 2021.

[28]. N. Tabassum, A. Rehman, M. Hamid, M. Saleem, and S. Malik, "Intelligent Nutrition Diet Recommender System for Diabetic ' s Patients," 2021.

[29]. D. Baig et al., "Bit Rate Reduction in Cloud Gaming Using Object Detection Technique," 2021.

[30]. G. Ahmad et al., "Intelligent ammunition detection and classification system using convolutional neural network," Comput. Mater. Contin., vol. 67, no. 2,

pp. 2585–2600, 2021.

[31]. N. Tabassum et al., "Prediction of Cloud Ranking in a Hyperconverged Cloud Ecosystem Using Machine Learning," Comput. Mater. Contin., vol. 67, no. 3, pp. 3129–3141, 2021.

[32]. M. I. Tariq, N. A. Mian, A. Sohail, T. Alyas, and R. Ahmad, "Evaluation of the challenges in the internet of medical things with multicriteria decision making (AHP and TOPSIS) to overcome its obstruction under fuzzy environment," Mob. Inf. Syst., vol. 2020, 2020.

[33]. N. Tabassum, M. Khan, S. Abbas, T. Alyas, A. Athar, and M. Khan, "Intelligent reliability management in hyper-convergence cloud infrastructure using fuzzy inference system," ICST Trans. Scalable Inf. Syst., vol. 0, no. 0, p. 159408, 2018.

[34]. M. I. Sarwar, K. Nisar, S. Andleeb, and M. Noman, "Blockchain – A Crypto-Intensive Technology - A Review," in 35th International Business Information Management Association (IBIMA) Conference, November 4-5,2020, Seville, Spain, pp. 14803–14809.

[35]. M. A. Khan et al., "Effective Demand Forecasting Model Using Business Intelligence Empowered with Machine Learning," IEEE Access, vol. 8, pp. 116013–116023, 2020.

[36]. A. Amin et al., "TOP-Rank: A Novel Unsupervised Approach for Topic Prediction Using Keyphrase Extraction for Urdu Documents," IEEE Access, vol. 8, pp. 212675–212686, 2020.

[37]. S. Abbas, M. A. Khan, A. Athar, S. A. Shan, A. Saeed, and T. Alyas, "Enabling Smart City With Intelligent Congestion Control Using Hops With a Hybrid Computational Approach," Comput. J., vol. 00, no. 00, 2020.

[38]. M. Muhammad, T. Alyas, F. Ahmad, F. Butt, W. Qazi, and S. Saqib, "An analysis of security challenges and their perspective solutions for cloud computing and IoT," ICST Trans. Scalable Inf. Syst., p. 166718, 2018.

[39]. M. Mehmood et al., "Machine learning enabled early detection of breast cancer by structural analysis of mammograms," Comput. Mater. Contin., vol. 67, no. 1, pp. 641–657, 2021.

[40]. N. Iqbal, S. Abbas, M. A. Khan, T. Alyas, A. Fatima, and A. Ahmad, "An RGB Image Cipher Using Chaotic Systems, 15-Puzzle Problem and DNA Computing," IEEE Access, vol. 7, pp. 174051–174071, 2019.

[41]. A. Alzahrani, T. Alyas, K. Alissa, Q. Abbas, Y. Alsaawy, and N. Tabassum, "Hybrid Approach for Improving the Performance of Data Reliability in Cloud Storage Management," Sensors (Basel)., vol. 22, no. 16, 2022.

Hussain *et al. (IJECI)* 2023

(IJECI)
ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

International Journal for
Electronic Crime Investigation

DOI: https://doi.org/10.54692/ijeci.2023.0702152

Research Article

Vol. 7 issue 2 Year 2023

Blockchain Data Analytics: A Review

Rabia Aslam Khan[1], Muhammad Bilal But[2] and Sabreena Nawaz[3]

[1]University of Management and Technology, Lahore

[2]University of South Asia, Lahore

[3]University of Engineering and Technology, Lahore

Corresponding author: f2019288013@umt.edu.pk

Received: March 05, 2023; **Accepted**: March 19, 2023; **Published**: June 15, 2023

Abstract:

Blockchain technology has emerged as a transformative force with widespread applications across various industries. In particular, the analysis of blockchain data has become crucial for crypto businesses and financial institutions seeking to protect transactions from illicit activities, minimize the risk of financial crimes, and ensure compliance with regulations. This paper presents a comprehensive review of blockchain data analytics, focusing on its significance in these domains. The paper examines the advancements and possibilities in blockchain data analytics, shedding light on their transformative potential. It provides an overview of the techniques and tools used for analyzing blockchain data, including transaction tracing, pattern recognition, and anomaly detection. Moreover, it explores the challenges and opportunities associated with blockchain data analytics, such as scalability, privacy concerns, and regulatory frameworks.

1. Introduction

The past decade has witnessed significant advancements in Blockchain technology, shaping its development and impact. The shared ledger Blockchain has been distributed which records transactions flanked by two parties without a stable central authority. Two individuals may perform an irreversible transaction on the Blockchain that is permanently registered on the public ledger [1]. The first use of the Blockchain was Bitcoin's cryptocurrency. The success of Bitcoin has given way to an age called Blockchain 1.0. Above 1000 chain-based cryptocurrencies known as "'alt-coins" are currently in place. These developments have created public interest in technology for the Blockchain[2]. Several new applications have emerged based on the Blockchain which includes identity

systems (e.g. Hyper and Bitnation), copyright (e.g. Blockphase and LBRY) voting (e.g. Social Krona, FollowMyVote), and origin (such as chronicled and EverLedger).

Private blockchains have been built and only provide participants with the required permissions to read and write. In contrast, anonymous Blockchains, for instance, Bitcoin enables unrestricted network access to any block node without requiring permission. In all Block network nodes, all transactions can be detected. In this article, we concentrate on public Blockchains, where information is open to the public. While most Blockchain solutions adhere to a chain structure, it is worth noting that alternative data structures can also be employed. Analyzing this knowledge can yield fresh insights into emerging patterns, giving rise to numerous questions such as:

1) How will the data stored on Blockchains be interpreted and modeled?
2) What insights can be gleaned from shared Blockchain's transactions?
3) What are the cutting-edge computational, analytical tools, and techniques currently used for analyzing Blockchain data?

By providing a brief introduction to Blockchain analytics, we answer the above raised questions. First, we give a short history of shared blockchains. Following the examination of typical "data structure models" in Blockchain, this paper provides insights into essential analytical methods and tools utilized in the field. Finally, new research has been addressed using Blockchain cryptocurrency modelling, e-criminal identification, trade of human beings and illegal economic activity analytics.

2. Literature Review

Some of the developments we took for granted in their day were also revolutions. Remember about how much the way we live and operate on smartphones has changed. When people were out of the workplace, they went elsewhere, and they were connected to a location by telephone, not to an individual. Today, global nomads are beginning to create new companies directly on their phones. Smartphones were there just a decade ago. We are in the middle of yet another silent revolution: blockchain, a digital ledger with an ever-growing collection of documents or records called "blocks. Bitcoin, as the pioneering global blockchain innovation, initiated the digital currency experiment. Presently, the market capitalization of Bitcoin fluctuates between $10 and $20 billion, and it serves as a medium of exchange for millions of individuals, even within the vast and growing cash market.

The second invention was known as the blockchain, and was mostly the discovery that Bitcoin's code could be isolated from the money and used for all sorts of interorganisation. Current blockchain research is under way in almost every big financial institution in the world and 15 per cent of banks could use blockchain in 2017.

The third invention is known as the 'intelligent

contract,' embodied in a blockchain system of the second generation called 'Ethera,' which explicitly includes small computational programs in blockchain that allow for the presentation of financial instruments such as loans and bonds instead of only cash tokens of bitcoin. The intelligent contract network currently has a market value of about $1 billion, with over hundred proposals on the market.

The fourth big breakthrough is the "proof of stake," which is now the bleeding edge of blockchain thinking. Current blockchains of the age are guaranteed with "job evidence," of which the party with the highest overall computational capacity decides. These groups are known as miners and run huge data centers in return for cryptocurrency payments to provide this protection. This data centers are replaced by modern structures, with a comparable or even higher level of safety, with sophisticated financial instruments. Proof-of-stake applications will live later this year.

Blockchain scaling is the fifth key breakthrough on the horizon. Right now, every device processing each transaction on the network in the blockchain world. It's sluggish. A scaled blockchain accelerates this mechanism by determining the number of computers needed to verify each transaction and division of work effectively without losing protection. It is a tough but not unpleasant challenge to handle this without losing the iconic safety and strength of blockchain. It is anticipated that a scaled blockchain would be fast enough to speed up the Internet of Things and lead the world's big payment mixers (VISA and SWIFT) [3].

It is only ten years after an elite group of informatics, encrypters and mathematicians worked for this landscape of creativity. When this breakthrough's full potential affects civilization, things will surely be a little strange. Blockchain technology enables payments for utilities like charging stations and landing pads to be more efficient and reliable. Transactions in international currencies, which can currently take anywhere from days to mere minutes, could be streamlined and made more dependable with this emerging system. All these changes and the other reforms are a major reduction in transaction costs. If the costs of transactions fall below invisible limits, aggregations and dislocations of current business models can be abrupt, drastic, and difficult to foresee. For instance, auctions were formerly small and local instead of universal and national, as they are now available on platforms such as eBay. The scheme suddenly changed as the costs of contacting people fell. As many of them as e-commerce has since been invented in the late 1990s, Blockchain is fairly supposed to trigger.

2.1. Blockchain

Blockchain looks complex and may certainly be complicated, but it has a very basic central definition. A ledger sort is a blockchain. It helps to get a first understanding of blockchain, what a ledger is really. A database is a knowledge set that is maintained on a file server electronically. Data, or data, is usually

organized in table format in data base so that relevant information can be searched and filtered more easily. How does anyone use a Table to store information rather than a database? Spreadsheets intended for storage and retrieval information in limited quantities for an individual or a specific number of persons. In the other hand, a database contains much greater quantities of information that can be downloaded, filtered and exploited by a multitude of people efficiently and effectively. This is achieved by housing data on servers consisting of powerful machines in large databases. This servers can be installed with hundreds or thousands of processors, to allow multiple users to concurrently access the database through computing power and storage. Whilst any number of persons may have access to a spreadsheet or database, it is always maintained and run by a designated person, who is fully responsible for how it functions and the data in it.

2.2. Blockchain Storage Structure

The data structure is a vital difference between a conventional and a blockchain databases. A blockchain collects information in groups containing information known as blocks. Blocks have these capabilities and the "block-chain" data chain is linked to the previously filled block until they are loaded. The new information after the recently inserted block is compiled and attached to the chain until it is done. A database organises its data into tables, whereas a blockchain organises its data into chunks, as its name implies, are chained together. It makes a blockchain database, but it's not all blockchain databases. This method also creates an inherently immutable decentralised data timeframe. When a block is filled, it becomes part of this schedule and is placed in stone. When attached to the chain, an exact time stamp is given to any block in the chain.

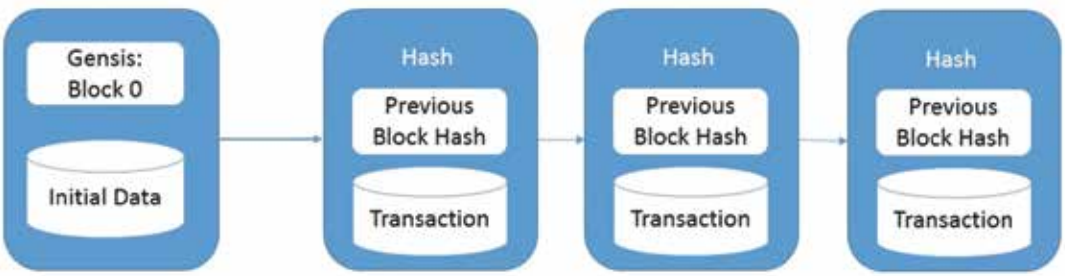
2.3. Blockchain Transaction Process

Blockchain is related list of the irreversible tamper-resistant blocks that has been stored at every node. A collection of transactions and related meta-data is recorded for each block. On the same ledger data deposited on each node, the transactions act Blockchain. First viewed as a peer-to-peer sharing mechanism by Satoshi Nakamoto[4]. Nakamoto referred to the transaction tokens traded as Bitcoins between customers in his scheme.

An immutable tamperproof block is the main element in every blockchain framework. A block in blockchain is encrypted data related to number of transactions in its simplest form. The fact that the block exists is a guarantor of the execution and verification of transactions. A current Blockchain is attached to a newly formed block. This Blockchain is mostly a related list that links one block to the other. The genesis block is the original block of such list. “Genesis Block” is a particular block which is numbered or labelled as zero. It is hard coded while programming the blockchain. Every other block has connections to an existing block. Thus, by adding new blocks to the current chains, a blockchain can expand[5].

Any OLTP transaction which operates on certain data shall be equivalent to a transaction

in a blockchain environment. Traditional blockchain implementations (like Bitcoin) provide transactions representing money sharing between two individuals (or users). Any valid transaction is registered for efficiency in a block that may consist of several transactions. Immutability is accomplished by the use of solid cryptographic characteristics including hazing. A blockchain is in fact a connected sequence, as each block stores in its chain the hatch of the previous block. The hash of its contents within the block is also digitally signed by each block. These hazards have cryptographical integrity, since any opponent who wishes to amend a block often has to change all previous blocks in a series, making the attack cryptographically impossible. One main design approach is to build a Merkle tree to store and verify hazelnuts efficiently. Thus the Merkle tree root is stored by each block, as the root makes the immutability simple to check. Figure 1 shows basic blockchain transaction processing is shown in block diagram.



Blockchain transactions are similar to conventional equivalents in the database. The clients transmit those transactions to the blockchain system servers[6]. This transfers are based on the data saved on all the servers involved. The blockchain transaction in its vanilla form can be used on the replicated distributed database as a series of read or write operations carried out on any of the node in blockchain. Any blockchain implementation uses a consensus protocol to specify the ordering for all incoming transactions.

2.4. Data Models of Blockchain

Public blockchains can be categorized primarily into two types: those utilizing the "Unspent Transaction Output" model, also known as UTXO, and those that operate based on accounts (such as Bitcoin, Litecoin, Ethereum etc.). A block of the data contain limited number of the transactions in all types of blockchains, but transactions vary. Below, these two types of blockchain transaction data are briefly discussed.

2.5. Blockchain Data Based on Upsent Transaction Output

The unused blockchains (UTXO) are first and the most important blockchains in terms of market capitalization: Bitcoin itself ranges from 45 - 60% of the entire market capitalisation of cryptocurrency. Each block of data includes a (financial) transaction, which covers the transference of coins between several

parties in UTXO blockchains. In each transaction, certain inputs are consumed and new outputs are produced (i.e. coins are directed to). There are three rules that emphasizes the shape of the data on the UTXO blockchain. This is because of Satoshi Nakamoto's design choices in Bitcoin[4].

2.5.1. Balance Rule

In the same transaction, all coins obtained from a single transaction shall be used. The transaction fee is any amount not sent to an output address and is collected by the miner who creates the block.The coin user will keep up the change by generating a new address (i.e. changing address) and submitting to this new address the balance remaining. Another alternative is to redirect the balance by the address of the user as one of the output addresses. This reprocess of donor address is discouraged. As a consequence, most of the nodes appear only twice, once when coins are received and again when they are spent. If a change address is generated, resulting in becoming new address of owner of the coin. Because of these principles, blockchains based on unspent transaction results should be considered as branching trees instead of networks. Non-transactional data are also stored in blockchains in UTXO. Nakamoto[4] Left the text message "The Times 3 January 2009 Chancellor on the verge of a second bailout for banks" was included in the first Bitcoin block. Metadata is disputed in Bitcoin transactions, and since 2014, every Bitcoin transaction has included an 80-byte (OP RETURN) field designed to store log information.

The blockchain was created in 2011 to store key-value pairs for a distributed namespace,enhancing the functionality of metadata. Namecoin data blocks are used to store ICANN-controlled registrations and updates for.bit domain names.[7].

2.5.2. Source Rule

Multiple transactions' input coins can be combined and consumed in single transaction, or they can be spent individually.

1.4.3. Mapping Rule

Payment of each coin must demonstrate proof of the funds through referring to a previous collection of outputs. While this helps us to track the past of payments, it is not always possible to pinpoint the origins of a particular coin. This is due to the fact that each transaction has its own set of inputs and outputs.

1.4.4. Blockchian Data-Based on Account

In the account-based blockchains, some of the coins are spent while retaining the rest. Ant transaction in blockchain has unerringly 1 input and 1 address While it is simple to create an address, most people use the same address for receiving and sending coins many times. Ethereum[8], is

actually the most valuable account-based blockchain established in 2015. Including Bitcoin, Ethereum has its own currency: Ether. The Ethereum project's main goal is to store data and software code on a Blockchain. The code (a smart contract) is written in the Ethereum Virtual Machine's proprietary Solidity coding language, which is compiled and executed as bytecode. In both code and agreements, smart contracts are self-executing contracts. MYSQL snippets in a database are an example. Intelligent contracts, on the other hand, ensure the non-stop, deterministic execution of code that can be publicly regulated.

Externally controlled addresses (governed by users) and contract addresses are used in account-based blockchains (governed by smart contract code). In a blockchain system, the process of uploading smart contract code involves the initiation of a transaction by an externally owned address or a contract address. The transaction is broadcasted to the network and propagated to all participating nodes, where it is validated and included in a block during the consensus process. Once confirmed, the smart contract code becomes part of the blockchain's immutable ledger and is distributed to all nodes, ensuring decentralization, redundancy, and transparency. This shared infrastructure enables trustless and transparent interactions with the contract, as users can examine the code's functionality and expect consistent outcomes. To put it another way, uploading the contract forces the code to be stored locally on other nodes. Each Ethereum transaction comprises of an input data field, similar to the log field in UTXO blockchains, which is used to transfer messages to the smart contracts.

When executing smart contract code in the blockchain, the code is invoked by calling stored functions with specified parameters. This process takes place across all nodes worldwide, establishing Ethereum as the "World Computer." The contract formation cost is borne by the contract holder, while other users or contracts interact with the contract by creating transactions directed to its address. The operations performed by the contract, such as multiplication (e.g., 5) and addition (e.g., 3), accumulate a computational cost known as "gas," which is measured in ethers and charged to the address initiating the transaction. Ether serves as the digital currency, acting as the fuel for the Ethereum World Computer. The advent of smart contracts has given rise to smart contract-based tokens, representing units of data that can be traded. These tokens enable users to access real-world services provided by businesses. For instance, the Storj token allows storage of files on personal hard drives and compensates users with Ethereum-based fees. Tokens

can be bought, sold, and function as stores of value within the global economy, with their exchange rates against fiat currencies viewable on platforms like coinmarket-cap.com.

There are two types of transactions in account-based blockchains. The first form of transaction involves sending a crypto-currency, such as Ether on Ethereum, from one address to another. A guided edge between the two addresses can be used to model this. Internal transactions, on the other hand, are made when smart contracts alter states associated with addresses.

1.5. Blockchain Data Analytics Tools and Methods

2.6.1. Tools

In blockchain systems, data blocks are typically stored in files on disk. For example, Ethereum utilizes levelDB files, while Bitcoin uses .dat files. However, the nature of storing data in this manner can result in time-consuming data querying processes. Although several blockchain query languages and analytics frameworks have emerged in recent years, their adoption remains relatively limited. [9]. In-house data querying and analysis tools have been developed by companies including Santiment.com and Chainalysis.com, but they are not yet available to the general public. Online explorers such as blockchain.com and etherscan.io provide limited access to analytics tools for the public. The BlockSci project [10] is a commonly used method in Bitcoin data analytics. Biva.1, a Bitcoin Network Visual Analytics application, is a related tool. In addition to transaction data related to financial interactions between addresses, the advent of Ethereum 2.0 and its inclusion of software code within blockchains has significantly contributed to the emergence of smart contract analysis as a vital path for data analytics. However, most of the existing research in this domain primarily focuses on static code analysis, which involves tasks like contract classification. Unfortunately, there is limited examination of the decisions made by the smart contracts under investigation. [11].

3. Methods

Early research works in the field of blockchain analysis have focused on analyzing UTXO data by constructing graphs using a single type of node, following established network analysis methodologies. Two prominent methods used in this context are the transaction graph and address graph approaches. The transaction graph method primarily considers transactions while ignoring addresses, forming edges between transaction nodes. It assumes that transactions are acyclic and prohibits the addition of new edges to a transaction node in the future. Conversely, the address graph method disregards transactions and connects

address nodes with edges. However, the presence of the Mapping Rule, which links a transaction's inputs to all its output addresses[12], can result in the formation of large cliques when transactions involve a high number of addresses. It is important to note that employing single node approaches alone may not adequately capture the intricacies of blockchain data, including the relationships between transactions and addresses. Therefore, further advancements and methodologies are required to comprehensively understand and analyze the complexities of blockchain data.

The loss of knowledge regarding addresses or transactions can have a significant impact on predictive models. When crucial information about addresses or transactions is missing or not considered, it can lead to incomplete or inaccurate predictions. The predictive models heavily rely on historical data and patterns to make informed projections or forecasts. If there is a loss of knowledge, such as incomplete transaction records or missing address data, it can undermine the model's ability to accurately capture trends, patterns, and relationships within the blockchain data. As a result, the predictive models may produce less reliable or misleading outcomes, potentially hindering decision-making processes reliant on the model's predictions[13].

As Bitcoin gained popularity, numerous studies emerged with the objective of predicting its price by examining various network characteristics. For instance, researchers explored network features such as mean account balance, the number of new edges, and clustering coefficients in the blockchain network [13]. In contrast, [14]have used network flows and network temporal activity as alternative price predictors, respectively. K-chainlets provide a lossless network encoding technique for the blockchain, using subgraphs composed of nodes that can represent addresses or chainlets. This model leverages the local higher-order structures present in the blockchain graph, treating subgraphs as the building blocks for analysis instead of individual edges or nodes. These subgraphs, known as chainlets, are formed based on a single judgment, allowing them to be treated as a single data unit. Unlike social networks, where nodes' proximity is influenced by their neighbors' behavior, the inclusion of input and output nodes within a chainlet is fixed and cannot be modified due to the irreversible nature of blockchain transactions. Chainlets offer a powerful means of capturing complex relationships and patterns within the blockchain, enabling more comprehensive analysis and understanding of transaction flows. By considering subgraphs as cohesive units, the granularity of analysis increases, providing a higher level of abstraction for studying the blockchain's structure and dynamics. [15].

Blockchain Data Analytics' Applications

Since the seminal Bitcoin paper, cryptocurrencies have been the most widely used Blockchain technology [4] in 2008. While there has recently been interest in analysing Blockchain platform data, Blockchain Data Analytics has

primarily focused on Bitcoin and a few other cryptocurrencies.. In general, studies look at the potential and weaknesses of coins in terms of providing a stable and open economic structure for all participants. Applications of the blockchain data analytics are described below:

4. Criminal Usage Detection

The utilization of Bitcoin in illicit activities, such as on SilkRoad.com, has been prevalent since its inception. While cryptocurrencies offer pseudonymity, as users are not required to disclose their identities to participate in the network, all transactions are visible on the public Blockchain. In order to maintain anonymity, criminals employ various tactics to separate their real-life and online identities. One such method is accessing the Blockchain network through privacy-enhancing distributed platforms like Tor. Additionally, criminals aim to make their activities on the Blockchain indistinguishable from those of regular users, attempting to create transactions that appear natural in terms of frequency, timing, and quantity. Cryptocurrencies are also utilized in illegal practices such as personal extortion, human trafficking, and ransomware payments. To combat these activities, law enforcement authorities may employ Blockchain Data Analytics software and algorithms to identify and analyze illegal behaviors [16-19]

5. Trade Finance

Companies have increasingly found traditional methods of commercial financing to be frustrating, as lengthy cycles often disrupt operations and create challenges in managing liquidity effectively. When it comes to cross-border trade, multiple variables need to be considered, including the country of origin and product details. Such transactions also generate a significant amount of paperwork. However, blockchain technology has the potential to revolutionize trade finance by simplifying transactions and enhancing cross-border management. It empowers companies to operate more efficiently across regions and overcome geographical boundaries, thus improving overall trade finance operations.

6. Money Laundering Protection

Blockchain technology, at its essence, offers immense value in the fight against money laundering. Its underlying technology provides strong support for the implementation of crucial mechanisms like 'Know Your Customer (KYC).' KYC enables businesses to uncover and verify the identities of their customers, serving as a vital tool in preventing money laundering activities. By leveraging the capabilities of blockchain encryption, organizations can enhance their safeguards against illicit financial transactions and bolster their overall money laundering protection measures.

7. Supply Chain Management

The inherent immutability of the blockchain ledger renders it highly suitable for various activities within supply chain management, including real-time product monitoring as goods move and change hands. Blockchain technology introduces numerous possibilities for businesses involved in shipping and logistics. By utilizing blockchain entries, events in the supply chain, such as the distribution of products across different containers until they reach their destination port, can be efficiently tracked and recorded. This empowers businesses with a modern and dynamic approach to organizing and leveraging tracking data, thereby enhancing overall supply chain management practices.

7.1. Real Estate

In the real estate industry, where homeowners typically sell their houses every five to seven years and individuals tend to move approximately 12 times in their lifetime, blockchain technology holds significant potential. With such high activity levels, blockchain can play a vital role in expediting domestic property sales by enabling swift monitoring of financial details. Additionally, it serves as a robust safeguard against encryption theft, ensuring the security of sensitive information. Moreover, blockchain brings transparency to the sales and procurement process, fostering trust and confidence among buyers, sellers, and other stakeholders in the real estate market.

8. Healthcare

Blockchain technology has great potential in the healthcare sector, particularly for storing and managing certain types of health information. General details such as age and sex, as well as basic medical data like immunization records or vital signs, can be effectively stored on a blockchain. These pieces of information, when isolated, do not reveal a patient's identity, thereby addressing privacy concerns. By utilizing a shared blockchain, accessible to a large number of authorized individuals, healthcare stakeholders can securely access and update relevant health information.

As connected medical devices gain popularity and become more integrated with patient records, blockchain can play a crucial role. It enables the seamless integration of specialized connected medical devices with individual health records. Data generated by these devices can be stored and added to personal medical records through the blockchain. Currently, the fragmentation of data from connected medical devices poses a significant challenge, but blockchain technology can serve as the bridge connecting these data silos. By leveraging blockchain, healthcare providers can enhance data interoperability and create a comprehensive view of a patient's health information, ultimately improving the quality and efficiency of healthcare delivery..

8.1. Insurance

Intelligent contracts represent a groundbreak-

ing application of blockchain technology in the insurance industry. They offer a transparent and secure way for customers and insurers to handle claims effectively. By registering both the contracts and claims on the blockchain, the network can validate the authenticity of claims, reducing the occurrence of fraudulent or duplicate claims associated with the same incident. An illustrative example of this is openIDL, a network developed in collaboration with the American Insurance Association and built on the IBM Blockchain platform. This platform automates the reporting of insurance regulations and simplifies compliance requirements, streamlining the overall insurance process. With intelligent contracts and blockchain technology, the insurance sector is empowered to enhance efficiency, trust, and accuracy in claims management and regulatory compliance.

9. Price Prediction

One of the most urgent issues is how Bitcoin acts as a class of financial assets, particularly as regards whether the transaction chart is linked to price formation, liquidity or a market crash. The analysis of transactions and addresses and the price of Bitcoin has become a prominent analytic subject. It is becoming increasingly necessary to build mathematical models that can forecast and assign price changes to transactions and transaction graph properties. Although simple transactional Blockchain features such as the average transaction amount display mixed results for cryptocurrency price prediction, a number of recent studies have shown that the global graphical features are useful for predicting price[13]. For example, [20] looked at the effects of average balance, clustering coefficient, and the number of new edges on Bitcoin price prediction, and Blockchain chainlets as a predictor. In [21] recently proposed two network flow tests to calculate the Bitcoin transaction network's dynamics and evaluate the relationship between flow sophistication and Bitcoin market variables.

10. Conclusions

This article has shed light on the research topics that encompass the prevalent challenges faced in data management and analytics within real-world blockchain applications. By exploring these issues, we aim to lay the groundwork for future research endeavors focused on identifying viable solutions to these open problems. It is our hope that this study serves as a stepping stone towards addressing and resolving the key obstacles in the field, leading to advancements and innovations in blockchain technology.

11. References

[1]. M. Andreessen, "Why Bitcoin Matters," *New York Times* vol. 21, 2014

[2]. M. Swan, *Blockchain: Blueprint for a New Economy*. O'Reilly Media, 2015.

[3]. V. Gupta, "A Brief History of Blockchain," ed. Harvard Business Review

2017

[4]. S. Nakamoto, "Bitcoin A Peer-to-Peer Electronic Cash System," 2008

[5]. R. W. Christian Decker, "Information Propagation in the Bitcoin Network," presented at the 13-th IEEE International Conference on Peer-to-Peer Computing 2013

[6]. F. Nawab, "Geo-Scale Transaction Processing," pp. 1-7, 2018.

[7]. H. Kalodner, M. C. , P. E. , J. B. , and A. N. , "An empirical study of namecoin and lessons for decentralized namespace design " 2015.

[8]. G. WOOD, "ETHEREUM: A Secure Decentralized Generalized Transaction Ledger," 2014.

[9]. M. Bartoletti, S. Lande, L. Pompianu, and A. Bracciali, "A general framework for blockchain analytics," pp. 1-6, 2017.

[10]. M. M. Harry Kalodner, Kevin Lee, Steven Goldfeder, Martin Plattner, Alishah Chator, Arvind Narayanan, "Blocksci: Design and applications of a blockchain analysis platform," 2017.

[11]. M. Bartoletti and L. Pompianu, "An Empirical Analysis of Smart Contracts: Platforms, Applications, and Design Patterns," vol. 10323, pp. 494-509, 2017.

[12]. F. M. Michele Spagnuolo, Stefano Zanero, "BitIodine: Extracting Intelligence from the Bitcoin Network," vol. 8437, 2014.

[13]. M. Faghih Mohammadi Jalali and H. Heidari, "Predicting changes in Bitcoin price using grey system theory," *Financial Innovation,* vol. 6, no. 1, 2020.

[14]. S. Y. Yang and J. Kim, "Bitcoin Market Return and Volatility Forecasting Using Transaction Network Flow Properties," pp. 1778-1785, 2015.

[15]. C. G. Akcora, A. K. Dey, Y. R. Gel, and M. Kantarcioglu, "Forecasting Bitcoin Price with Graph Chainlets," vol. 10939, pp. 765-776, 2018.

[16]. R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second generation Onion Router," 2004.

[17]. S. Phetsouvanh, F. Oggier, and A. Datta, "EGRET: Extortion Graph Exploration Techniques in the Bitcoin Network," pp. 244-251, 2018.

[18]. R. S. Portnoff, D. Y. Huang, P. Doerfler, S. Afroz, and D. McCoy, "Backpage and Bitcoin," pp. 1595-1604, 2017.

[19]. M. M. A. Danny Yuxing Huang, Vector Guo Li Luca Invernizzi, Kylie McRoberts, Elie Bursztein, Jonathan Levin Kirill Levchenko, Alex C. Snoeren, Damon McCoy, "Tracking ransomware end-to-end," 2018.

[20]. M. Sorgente and C. Cibils, "The Reaction of a Network: Exploring the Relationship between the Bitcoin Network Structure and the Bitcoin Price," 2014.

[21]. S. Y. Yang and J. H. Kim, "Bitcoin Market Return and Volatility Forecasting Using Transaction Network Flow Properties," 2014.

Humaira *et al. (IJECI)* 2023

(IJECI)
ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

International Journal for
Electronic Crime Investigation

DOI: https://doi.org/10.54692/ijeci.2023.0702153

Research Article

Vol. 7 issue 2 Year 2023

Analysis of Network Security in IoT-based Cloud Computing Using Machine Learning

Humaira Naeem

Department of computer science, Virtual university of Pakistan

Corresponding author: humairanaeem@vu.edu.pk

Received: March 07, 2023; **Accepted**: March 22, 2023; **Published**: June 15, 2023

Abstract:

Network security in IoT-based cloud computing can benefit greatly from the application of machine learning techniques. IoT devices introduce unique security challenges with their large-scale deployments and diverse nature. Machine learning can help address these challenges by analyzing IoT network traffic, detecting anomalies, identifying potential threats, and enhancing overall network security. The security of cloud networks is validated using binary classification to detect attacks. Random forest classifiers achieved an accuracy of 96%, while K nearest classifier had an accuracy of 93% and a precision value of 0.96. The proposed model ensures security of big data against intrusion attacks on the network. Although machine learning techniques can be powerful for protecting cloud computing networks, challenges still need to be addressed before widespread adoption. Understanding the potential and limitations of machine learning approaches to network security can help researchers and practitioners develop more effective strategies for safeguarding their systems in an increasingly interconnected world. Network security of big data in cloud computing can be enhanced by applying machine learning techniques. Machine learning algorithms can analyze large amounts of data to detect patterns, anomalies, and potential security threats. Here are several ways machine learning can be utilized to improve network security in the context of big data and cloud computing:

Keywords: Cloud computing ; NID; KNN; RF ; Machine learning

1. Introduction

The convergence of IoT (Internet of Things) and cloud computing has revolutionized the way we interact with and manage vast networks of interconnected devices. IoT-based cloud computing systems enable seamless communication, data storage, and analysis, facilitating a wide range of applications across industries. However, this rapid proliferation of IoT devices and the utilization of cloud services also bring forth significant network security challenges. The need to protect sensitive data, ensure device integrity, and

safeguard against emerging threats has become paramount [1].

To address these challenges, machine learning techniques have emerged as a powerful approach to enhance network security in IoT-based cloud computing environments. Machine learning algorithms have the capacity to analyze massive volumes of data generated by IoT devices and cloud services, enabling the identification of patterns, anomalies, and potential security threats. By leveraging machine learning capabilities, organizations can significantly strengthen their network security posture and mitigate the risks associated with IoT deployments[2].

This paper aims to explore the analysis of network security in IoT-based cloud computing using machine learning. We will delve into how machine learning techniques can be applied to bolster network security, such as intrusion detection, anomaly detection, device authentication, security analytics, threat intelligence, and privacy protection. Furthermore, we will examine the implications of machine learning in enabling predictive maintenance and proactive security measures within IoT-based cloud computing environments [3].

By harnessing the power of machine learning, organizations can better safeguard their IoT networks, protect sensitive data, and respond effectively to emerging threats. This paper will provide insights into the potential benefits, challenges, and considerations in implementing machine learning-based network security strategies in IoT-based cloud computing[4][5].

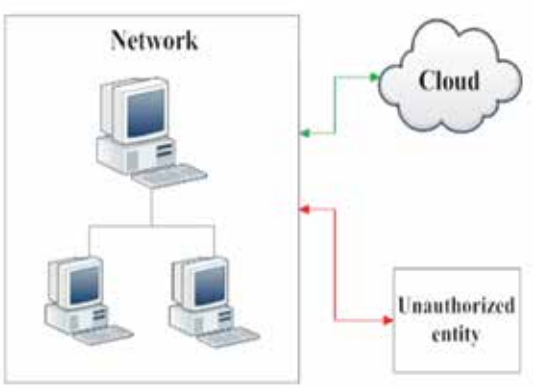


Figure 1: Unauthorized access of data

Machine learning algorithms can analyze network traffic data from IoT devices to identify patterns associated with known attacks or intrusions. By monitoring network packets, data payloads, and device behavior, machine learning models can detect and raise alerts for suspicious activities, helping prevent unauthorized access and data breaches [6].

IoT devices generate massive amounts of data, and machine learning can be utilized to identify abnormal patterns or behaviors. By training models on historical data, machine learning algorithms can learn the normal behavior of IoT devices and detect any deviations that may indicate potential security threats. For example, abnormal traffic patterns, unusual data transfers, or unexpected device behavior can be flagged for further investigation [7].

Machine learning can enhance device authentication mechanisms in IoT-based cloud computing. Machine learning models can differentiate between legitimate and unauthorised devices by analyzing device characteristics, communication patterns, and contextual

information. This helps enforce access control policies and prevent unauthorized access to cloud resources.]

Machine learning algorithms can be applied to analyze security-related data from IoT devices and cloud environments. Machine learning models can identify potential security events or trends by aggregating and correlating data from multiple sources, such as device logs, network traffic, and system logs. This enables proactive threat mitigation and aids in making informed security decisions [8].

Threat Intelligence and Response: Machine learning can assist in the analysis of threat intelligence feeds and security-related data to identify emerging threats and vulnerabilities in By integrating external threat intelligence sources with internal network data, machine learning models can provide real-time threat detection, enabling quick response and mitigation measures [9].

IoT devices often handle sensitive data, and machine learning can protect privacy. By applying machine learning techniques such as differential privacy, data anonymization, and encryption, IoT data can be secured while maintaining its utility for analysis and insights. Machine learning can help predict and prevent security incidents in IoT-based cloud computing environments. By analyzing historical data on device performance, maintenance logs, and security events, machine learning models can identify patterns that lead to security vulnerabilities or device failures. This enables proactive maintenance and security measures to prevent future incidents [10].

It's crucial to continuously train and update machine learning models in IoT-based cloud computing environments to adapt to evolving threats. Regular evaluation, monitoring, and collaboration with domain experts are necessary to ensure the effectiveness and accuracy of these models. Additionally, implementing security best practices such as secure device provisioning, network segmentation, and encryption protocols in conjunction with machine learning can create a comprehensive security framework for IoT-based cloud computing networks.

Figure 2: Cloud deployment Models

2. Related Work

In [11] proposed an intrusion detection system for IoT-based cloud computing using a deep learning approach. The study employed deep neural networks to analyze network traffic data and detect malicious activities. The model achieved high accuracy in identifying various types of attacks, highlighting the effectiveness of deep learning in IoT network security.

[12] presented a machine learning-based anomaly detection framework for IoT networks in cloud computing. The research focused on analyzing network traffic patterns to identify deviations from normal behavior. The proposed framework utilized machine learning algorithms, including clustering and classification, to detect anomalous activities in real-time.

In this paper [13] explored the use of machine learning techniques for threat intelligence in IoT-based cloud computing environments. The study integrated external threat intelligence feeds with network data to identify and classify emerging threats. Machine learning models were employed to analyze the data and provide timely threat detection and response.

Privacy preservation in IoT-based cloud computing has also received attention. Author proposed a privacy-preserving framework based on machine learning for IoT networks. The approach employed techniques such as differential privacy and data anonymization to protect sensitive data while still allowing effective analysis and insights. Predictive maintenance and security have been addressed through the application of machine learning in IoT-based cloud computing. The research focused on analyzing historical device performance data and maintenance logs to predict security vulnerabilities and device failures. Proactive maintenance measures and security enhancements were then implemented to prevent future incidents.

These studies highlight the diverse applications of machine learning in improving network security within IoT-based cloud computing. By leveraging machine learning algorithms, organizations can detect intrusions, identify anomalies, analyze threat intelligence, protect privacy, and implement predictive security measures. However, challenges such as data quality, model robustness, and scalability must be addressed to ensure the effectiveness and practicality of machine learning-based network security solutions in real-world deployments. The concepts of big data, the Internet of Things, and cloud computing are closely related and have the potential to bring significant benefits to industries such as healthcare and engineering. However, concerns related to their interconnectivity are also addressed in the proposed methodology outlined in [14].

Smart homes rely on the Internet of Things (IoT), which generates significant amounts of data from sensors and actuators for various activities. The utilization of IoT applications benefits homeowners and industrialists alike. A study proposed in paper [15] focuses on the use of cloud computing and fog nodes to store, network, and process IoT data. The methodology was validated using a Canadian smart home dataset and demonstrated promising results.

Cloud-based architectures provide computational models for performing big data operations, offering cost efficiency, elasticity, and virtualization benefits. Security concerns arise as data stored in the cloud is operated over the internet. The paper [16] proposes a big data-based security model for the cloud called

Big Cloud. The main goal is to address security concerns through automatic security evaluation. The system is validated using a case study of the Apache Hadoop stack, evaluating the strengths and weaknesses of the proposed methodology.

Big data faces several challenges due to data storage, data misuse, and illegal access. This research [17] addresses these issues and proposes an encryption technique for large-scale data storage in multiple cloud storages. The study's main objective is to provide a secure architecture that restricts unauthorized access to the system. The proposed framework involves processes such as uploading, slicing, indexing, distributing, decrypting, retrieving, and combining data. The study introduces a hybrid cryptographic method to secure vast amounts of data before storing them in multiple clouds.

The proposed methodology in this study [18] starts by examining the already in-use tiered cloud architectures before presenting a solution for storing massive data. The study then focuses on the usage of a P2P Cloud System (P2PCS) for processing and analyzing large data. Additionally, a case study is presented, which is related to the healthcare system, and offers a hybrid mobile cloud computing approach consisting of cloudlets. The Mobile Cloud Computing Simulator (MCCSIM) is used to simulate the model, and the hybrid cloud model outperforms classic cloud models by up to 75%. To validate the security and privacy safeguards, the system is evaluated against potential threats.

Data in big data may be organized, unstructured, or semi-structured, providing critical insights for businesses to make informed decisions. Many businesses rely on big data to manage and analyse their data stores. Storage management is crucial to big data management to ensure that data is stored properly, securely, and readily available. Despite the numerous benefits of big data technology, it still faces storage management challenges, especially when combined with cloud computing, such as the significant issue of data security. This paper [19] addresses the aforementioned security challenge.

Jaleel [20] proposed a fragmentation scheme based on columns, where the server side stores encrypted pieces of data. Each fragment is assigned a unique ID to support queries through two processing stages. First, the client sends the ID to fetch data from the server, and then the client decrypts the fragment before using it for the query. The result is returned after executing the query on the fetched fragment. However, this method may not be suitable for large datasets as it requires significant overhead for the entire database. Additionally, the need to fetch the entire database to the client side eliminates the benefits of cloud computing's storage capabilities, and the need to perform the query twice slows down the overall performance. Thus, this strategy is not an ideal solution for this problem as it hinders the primary benefit of cloud computing's storage.

Tabassum [21] proposed a solution to the challenge of data confidentiality when outsourcing a database. They argued that no

software-based solution could be entirely secure due to inconsistent internet security standards, so they suggested using a smart card as a mediator on the cloud side. The smart card would encrypt the data before loading it into the database and decrypt it before sending it to the user. This approach assumes that the client and the cloud communication occurs securely. The system has processing power and memory capacity limitations due to the smart card's limitations and the challenges of inserting it into the cloud provider's server. Therefore, it is not a practical solution for outsourcing sensitive data, and the situation worsens if the database system needs to be distributed.

Imran et al. [22] proposed a technique to secure data stored on untrusted cloud providers. The authors differentiated between private and public data by encrypting sensitive information and storing it on a smart USB key. Non-sensitive data was stored in plain text on a public cloud server. However, this approach is not practical for general usage as it necessitates the use of a USB key to access or query the data. To connect the two segments, a distributed protocol was suggested.

Tabassum et al. [23] proposed an approach to improve the security of data stored with untrusted cloud providers by utilizing encryption techniques, a proxy, and the user's application. The method consists of six encryption techniques to handle different types of queries that cannot be solved with a single encryption algorithm. Other research studies in this area are also available.

Peter et al. [24] propose a new approach to enhance security by combining encryption techniques and a fragmentation methodology. The scheme's architecture is illustrated in Figure 1. The public clouds are composed of a master cloud and several slave clouds. The master cloud stores an encrypted clone of the entire database while individual public clouds store extended columns. Column-based fragmentation technique consists of two parts: master cloud and slave clouds. The whole database is encrypted with a highly secure encryption method and stored in the master cloud without providing the encryption key to the master cloud provider during initial setup. None of the keys are shared with the cloud service providers.

Abadi's study [25] developed a technique for secure query processing on cloud-based data storage using an order-preserving homomorphism encryption approach. The algorithm uses a secret-splitting strategy to enable the CSP to handle complex SQL queries. The proposed PHE approach provides a balance between security and efficiency in real-world settings, and is both effective and cost-efficient. To evaluate the algorithm's effectiveness in terms of overhead, query processing capability, storage, and computational costs, CSPs and AUs will conduct further assessments [26]

3. Security Concerns In Cloud

Integrity: Ensuring the integrity of information stored in a system is crucial to guarantee that the requested information accurately represents the original data and hasn't been tampered with by unauthorized parties. To

protect against potential data loss, each network service typically has multiple backup systems in place. Regular backups of data are usually stored on removable media, which are kept off-site for additional security [27].

Availability: In the context of computer security, availability refers to the assurance that authorized users can access computational resources and services whenever they need to. This is crucial for ensuring the uninterrupted operation of mission-critical systems, as any unauthorized activity that hinders access to these resources can have serious consequences. Therefore, maintaining high availability is a fundamental aspect of a robust security strategy [28].

Confidentiality: Maintaining data confidentiality is critical to prevent unauthorized access to sensitive information by third parties. Unauthorized access can occur through various means such as social manipulation, technical vulnerabilities, or failure to encrypt communications between clients and servers. Social manipulation can lead to actual loss of confidentiality while technical vulnerabilities can result in compromised security. Therefore, it is important to adopt appropriate measures to ensure data confidentiality [29].

Figure 3: Security Concerns in Cloud

4. METHODOLOGY

The methodology of the proposed research is shown in figure 4 and is divided into the following:

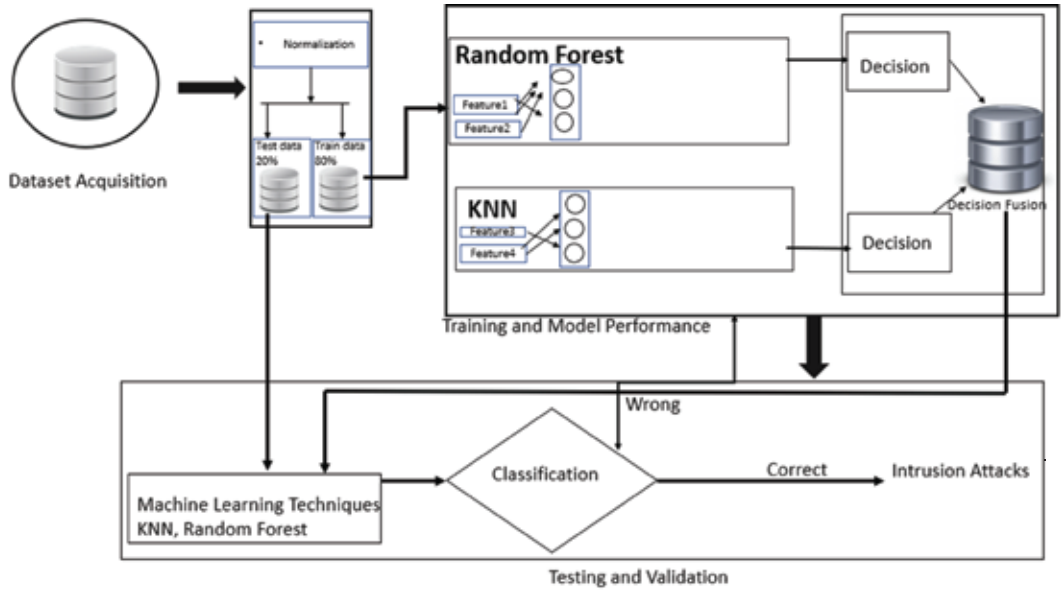


Figure 4: Proposed Methodology

4.1. Dataset Collection

The initial stage of the research involves gathering pertinent data. The UNR-IDD dataset is the focus of our investigation, which includes port statistics and TCP port data indicating changes in port statistics over a designated period. By examining network traffic at the port level where decisions are made, the port statistics offer a comprehensive analysis that enables the prompt detection of potential security breaches [30].

4.2. Preprocessing

This phase involves preprocessing operations on the dataset to remove artifacts such as null values and fabricated data. Proper preprocessing is fundamental as it greatly influences classification performance. If the data is not preprocessed correctly, the model will not produce the desired output [31].

4.3. Feature Extraction

The following data features have been captured in the targeted dataset:

Metrics and magnitudes are gathered from each port within the SDN during a simulated flow between two hosts in order to provide port statistics.

Delta Port Statistics: Change in collected metrics from each port within the SDN during a simulated flow between two hosts, observed over a 5-second interval for increased intrusion detection detail[32].

Flow Entry and Flow Table Statistics: Metrics that offer information on the network's switch conditions, gathered in any network environment, together with port statistics.

4.4. Training

In this phase, the model is trained for the task of evaluating network breaches. To address the challenge of tail classes, a targeted dataset is created with sufficient samples to enable machine learning classifiers to perform well. Furthermore, the dataset ensures completeness by having no missing data. In the proposed research, Random Forest and K-nearest neighbor are utilized to train the model [33].

4.5. Classification

The aim of this binary classification is to differentiate between normal operations and intrusions, with the ability to predict whether a network is under attack. However, the model does not provide information about the type or nature of the attack. Figure 5 shows data flow diagram of proposed systemc[34].

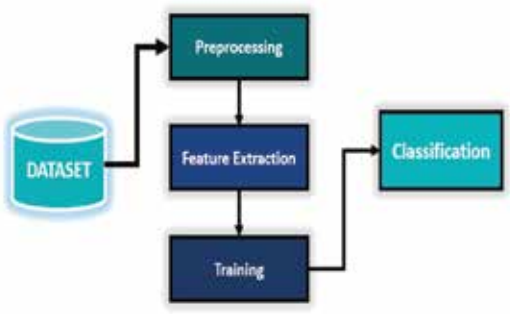


Figure 5: Data flow diagram

5. RESULTS

For the simulation of our proposed model we use Google Colaboratory where we implement our system using python. They diagram below shows the feature selection plot from our dataset [35].

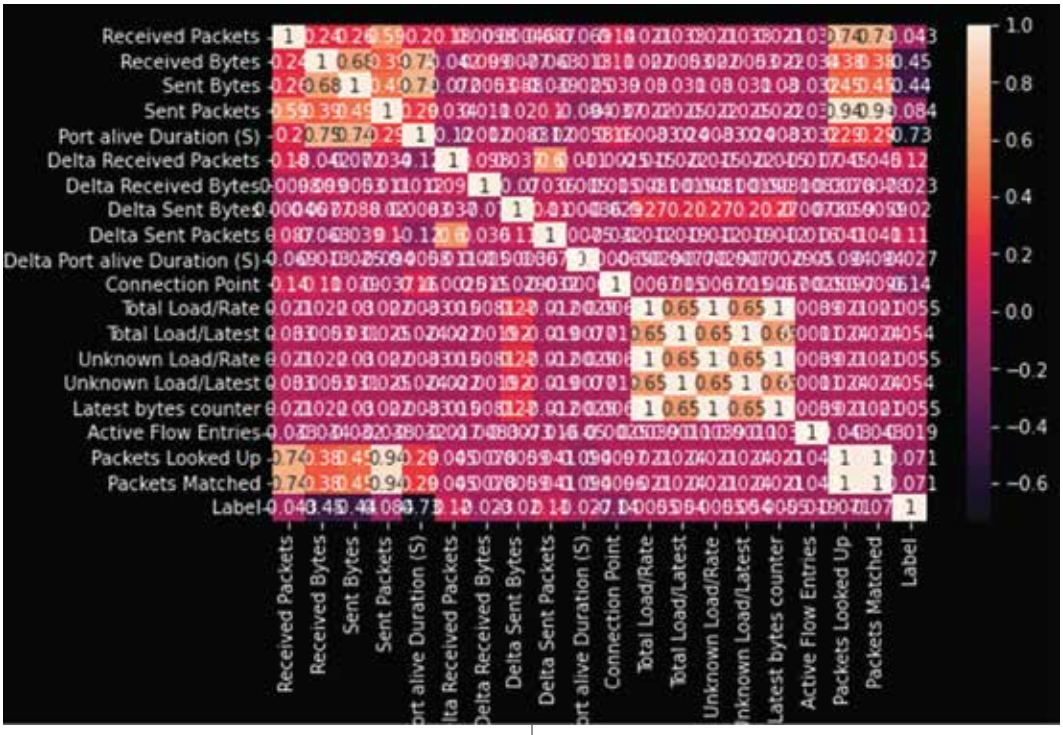


Figure 5: Feature Selection

A visual representation of the binary classifier labels is presented in Figure 6.

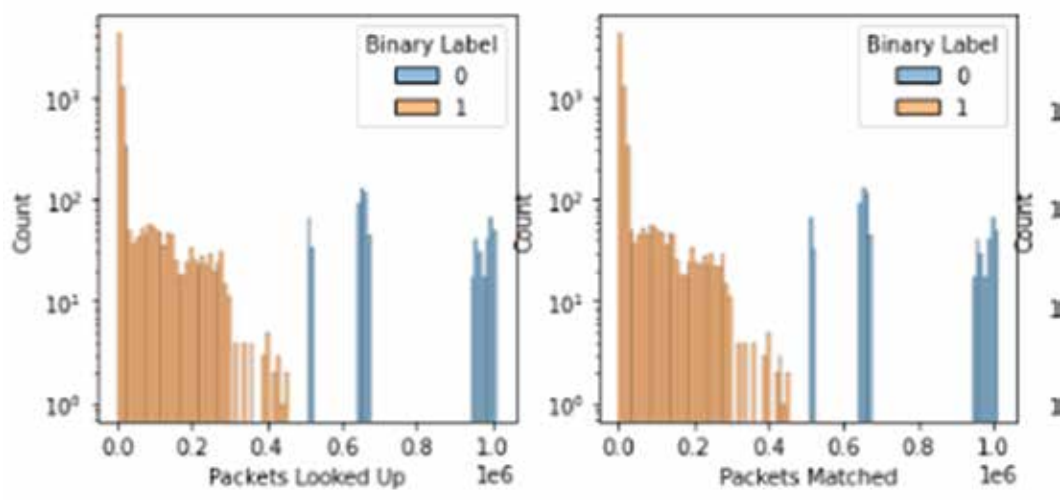


Figure 6: Packet Matching

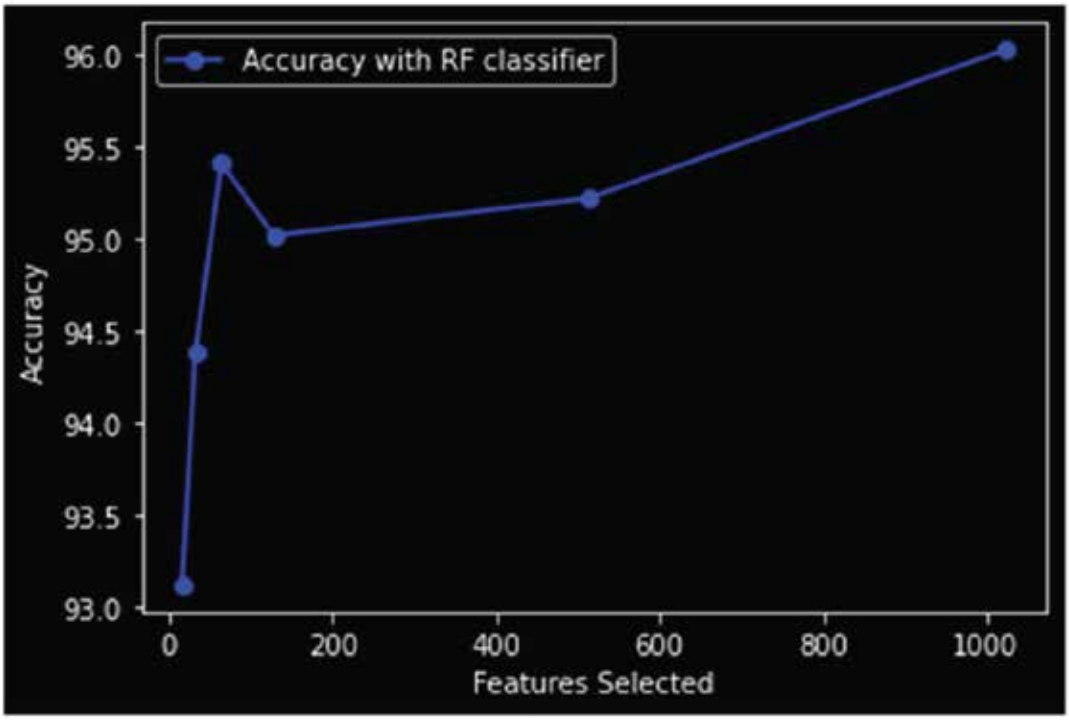


Figure 7: Accuracy Curve

In IoT-based cloud computing, network security is critical due to the large-scale deployment of IoT devices and the diverse nature of their communication. Machine learning techniques can play a crucial role in enhancing network security by leveraging the power of algorithms to analyze data and identify patterns or anomalies that indicate potential security threats [36].

One key application of machine learning is intrusion detection, where algorithms analyze network traffic data from IoT devices to identify known attack patterns. By examining network packets, data payloads, and device behavior, machine learning models can learn to recognize signatures of previous attacks and raise alerts when similar patterns are detected. This helps prevent unauthorized access and data breaches [37].

Another important aspect is anomaly detection. IoT devices generate massive amounts of data, and machine learning algorithms can be trained to understand the normal behavior of these devices. By analyzing historical data, machine learning models can identify deviations from the norm that may indicate security risks. For example, abnormal traffic patterns, unexpected data transfers, or unusual device behavior can be flagged for further investigation [38].

Machine learning can also contribute to device authentication and access control in IoT-based cloud computing. Machine learning models can differentiate between legitimate and unau-

thorised devices by analyzing device characteristics, communication patterns, and contextual information. This enables the enforcement of access control policies and prevents unauthorized access to cloud resources [39].

Security analytics is another area where machine learning can be applied effectively. Machine learning models can identify potential security events or trends by aggregating and correlating data from various sources, such as device logs, network traffic, and system logs. This enables proactive threat mitigation and facilitates informed security decisions [40].

Machine learning can also help in the analysis of threat intelligence feeds and security-related data to identify emerging threats and vulnerabilities in IoT-based cloud computing. By integrating external threat intelligence sources with internal network data, machine learning models can provide real-time threat detection, allowing quick response and mitigation measures [41].

Privacy and data protection are important considerations in IoT environments. Machine learning techniques, such as differential privacy, data anonymization, and encryption, can be employed to secure IoT data while preserving its utility for analysis and insights [42].

Furthermore, machine learning can facilitate predictive maintenance and security. By analyzing historical data on device performance, maintenance logs, and security events, machine learning models can identify patterns that lead to security vulnerabilities or device failures. This enables proactive maintenance and security measures to prevent future incidents [43].

The analysis demonstrates how machine learning techniques can be applied in IoT-based cloud computing to enhance network security. By leveraging the capabilities of machine learning algorithms, organizations can detect and prevent security threats, improve access control mechanisms, analyze security-related data, protect privacy, and implement proactive security measures. Continuous training, evaluation, and collaboration with experts are essential to ensure the effectiveness and accuracy of the machine learning models in evolving threat landscapes [44].

6. Conclusion

To validate the security of the system, binary classification is utilized for detecting attacks on cloud networks. Random forest classifiers achieved an accuracy of around 96.0%, while K nearest classifier had an accuracy of 93% and a precision value of 0.96. The proposed model ensures big data security against intrusion attacks on the network. While machine learning techniques can be powerful for protecting cloud computing networks, challenges still need to be addressed before these techniques can be widely adopted. Understanding the limitations and potential of machine learning approaches to network security can help researchers and practitioners develop more effective strategies for protecting their systems in a highly interconnected world.

7. References

[1]. N. Tabassum, A. Namoun, T. Alyas, A. Tufail, M. Taqi, and K. Kim, "applied sciences Classification of Bugs in Cloud Computing Applications Using Machine Learning Techniques," 2023.

[2]. M. I. Sarwar, Q. Abbas, T. Alyas, A. Alzahrani, T. Alghamdi, and Y. Alsaawy, "Digital Transformation of Public Sector Governance With IT Service Management–A Pilot Study," IEEE Access, vol. 11, no. January, pp. 6490–6512, 2023, doi: 10.1109/ACCESS.2023.3237550.

[3]. T. Alyas, K. Ateeq, M. Alqahtani, S. Kukunuru, N. Tabassum, and R. Kamran, "Security Analysis for Virtual Machine Allocation in Cloud Computing," Int. Conf. Cyber Resilience, ICCR 2022, no. Vm, 2022.

[4]. T. Alyas, "Performance Framework for Virtual Machine Migration in Cloud Computing," Comput. Mater. Contin., vol. 74, no. 3, pp. 6289–6305, 2023.

[5]. T. Alyas, S. Ali, H. U. Khan, A. Samad, K. Alissa, and M. A. Saleem, "Container Performance and Vulnerability Management for Container Security Using Docker Engine," Secur. Commun. Networks, vol. 2022, 2022.

[6]. M. Niazi, S. Abbas, A. Soliman, T. Alyas, S. Asif, and T. Faiz, "Vertical Pod Autoscaling in Kubernetes for Elastic Container Collaborative Framework," 2023.

[7]. T. Alyas, A. Alzahrani, Y. Alsaawy, K. Alissa, Q. Abbas, and N. Tabassum, "Query Optimization Framework for Graph Database in Cloud Dew Environment," 2023.

[8]. T. Alyas, "Multi-Cloud Integration Security Framework Using Honeypots," Mob. Inf. Syst., vol. 2022, pp. 1–13, 2022.

[9]. Alyas, N. Tabassum, M. Waseem Iqbal, A. S. Alshahrani, A. Alghamdi, and S. Khuram Shahzad, "Resource Based Automatic Calibration System (RBACS) Using Kubernetes Framework," Intell. Autom. Soft Comput., vol. 35, no. 1, pp. 1165–1179, 2023.

[10]. G. Ahmed et al., "Recognition of Urdu Handwritten Alphabet Using Convolutional Neural Network (CNN)," Comput. Mater. Contin., vol. 73, no. 2, pp. 2967–2984, 2022.

[11]. M. I. Sarwar, K. Nisar, and I. ud Din, "LTE-Advanced – Interference Management in OFDMA Based Cellular Network: An Overview", USJICT, vol. 4, no. 3, pp. 96-103, Oct. 2020.

[12]. A. A. Nagra, T. Alyas, M. Hamid, N. Tabassum, and A. Ahmad, "Training a Feedforward Neural Network Using Hybrid Gravitational Search Algorithm with Dynamic Multiswarm Particle Swarm Optimization," Biomed Res. Int., vol. 2022, pp. 1–10, 2022.

[13]. T. Alyas, M. Hamid, K. Alissa, T. Faiz, N. Tabassum, and A. Ahmad, "Empirical Method for Thyroid Disease Classification Using a Machine Learning Approach," Biomed Res. Int., vol. 2022, pp. 1–10, 2022.

[14]. T. Alyas, K. Alissa, A. S. Mohammad, S.

Asif, T. Faiz, and G. Ahmed, "Innovative Fungal Disease Diagnosis System Using Convolutional Neural Network," 2022.

[15]. H. H. Naqvi, T. Alyas, N. Tabassum, U. Farooq, A. Namoun, and S. A. M. Naqvi, "Comparative Analysis: Intrusion Detection in Multi-Cloud Environment to Identify Way Forward," Int. J. Adv. Trends Comput. Sci. Eng., vol. 10, no. 3, pp. 2533–2539, 2021.

[16]. S. A. M. Naqvi, T. Alyas, N. Tabassum, A. Namoun, and H. H. Naqvi, "Post Pandemic World and Challenges for E-Governance Framework," Int. J. Adv. Trends Comput. Sci. Eng., vol. 10, no. 3, pp. 2630–2636, 2021.

[17]. W. Khalid, M. W. Iqbal, T. Alyas, N. Tabassum, N. Anwar, and M. A. Saleem, "Performance Optimization of network using load balancer Techniques," Int. J. Adv. Trends Comput. Sci. Eng., vol. 10, no. 3, pp. 2645–2650, 2021.

[18]. T. Alyas, I. Javed, A. Namoun, A. Tufail, S. Alshmrany, and N. Tabassum, "Live migration of virtual machines using a mamdani fuzzy inference system," Comput. Mater. Contin., vol. 71, no. 2, pp. 3019–3033, 2022.

[19]. M. A. Saleem, M. Aamir, R. Ibrahim, N. Senan, and T. Alyas, "An Optimized Convolution Neural Network Architecture for Paddy Disease Classification," Comput. Mater. Contin., vol. 71, no. 2, pp. 6053–6067, 2022.

[20]. J. Nazir et al., "Load Balancing Framework for Cross-Region Tasks in Cloud Computing," Comput. Mater. Contin., vol. 70, no. 1, pp. 1479–1490, 2022.

[21]. N. Tabassum, T. Alyas, M. Hamid, M. Saleem, S. Malik, and S. Binish Zahra, "QoS Based Cloud Security Evaluation Using Neuro Fuzzy Model," Comput. Mater. Contin., vol. 70, no. 1, pp. 1127–1140, 2022.

[22]. M. I. Sarwar, K. Nisar, and A. Khan, "Blockchain – From Cryptocurrency to Vertical Industries - A Deep Shift," in IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), September 20-23, 2019, Dalian, China, 2019, pp. 537–540. doi: 10.1109/ICSPCC46631.2019.8960795.

[23]. S. Malik, N. Tabassum, M. Saleem, T. Alyas, M. Hamid, and U. Farooq, "Cloud-IoT Integration: Cloud Service Framework for M2M Communication," Intell. Autom. Soft Comput., vol. 31, no. 1, pp. 471–480, 2022.

[24]. W. U. H. Abidi et al., "Real-Time Shill Bidding Fraud Detection Empowered with Fussed Machine Learning," IEEE Access, vol. 9, pp. 113612–113621, 2021.

[25]. M. I. Sarwar et al., "Data Vaults for Blockchain-Empowered Accounting Information Systems," IEEE Access, vol. 9, pp. 117306–117324, 2021, doi: 10.1109/ACCESS.2021.3107484.

[26]. N. Tabassum, T. Alyas, M. Hamid, M. Saleem, and S. Malik, "Hyper-Convergence Storage Framework for EcoCloud Correlates," Comput. Mater. Contin., vol. 70, no. 1, pp. 1573–1584, 2022.

[27]. N. Tabassum et al., "Semantic Analysis of Urdu English Tweets Empowered by Machine Learning," 2021.

[28]. N. Tabassum, A. Rehman, M. Hamid, M. Saleem, and S. Malik, "Intelligent Nutrition Diet Recommender System for Diabetic ' s Patients," 2021.

[29]. D. Baig et al., "Bit Rate Reduction in Cloud Gaming Using Object Detection Technique," 2021.

[30]. G. Ahmad et al., "Intelligent ammunition detection and classification system using convolutional neural network," Comput. Mater. Contin., vol. 67, no. 2, pp. 2585–2600, 2021.

[31]. N. Tabassum et al., "Prediction of Cloud Ranking in a Hyperconverged Cloud Ecosystem Using Machine Learning," Comput. Mater. Contin., vol. 67, no. 3, pp. 3129–3141, 2021.

[32]. M. I. Tariq, N. A. Mian, A. Sohail, T. Alyas, and R. Ahmad, "Evaluation of the challenges in the internet of medical things with multicriteria decision making (AHP and TOPSIS) to overcome its obstruction under fuzzy environment," Mob. Inf. Syst., vol. 2020, 2020.

[33]. N. Tabassum, M. Khan, S. Abbas, T. Alyas, A. Athar, and M. Khan, "Intelligent reliability management in hyper-convergence cloud infrastructure using fuzzy inference system," ICST Trans. Scalable Inf. Syst., vol. 0, no. 0, p. 159408, 2018.

[34]. M. I. Sarwar, K. Nisar, S. Andleeb, and M. Noman, "Blockchain – A Crypto-Intensive Technology - A Review," in 35th International Business Information Management Association (IBIMA) Conference, November 4-5,2020, Seville, Spain, pp. 14803–14809.

[35]. M. A. Khan et al., "Effective Demand Forecasting Model Using Business Intelligence Empowered with Machine Learning," IEEE Access, vol. 8, pp. 116013–116023, 2020.

[36]. A. Amin et al., "TOP-Rank: A Novel Unsupervised Approach for Topic Prediction Using Keyphrase Extraction for Urdu Documents," IEEE Access, vol. 8, pp. 212675–212686, 2020.

[37]. S. Abbas, M. A. Khan, A. Athar, S. A. Shan, A. Saeed, and T. Alyas, "Enabling Smart City With Intelligent Congestion Control Using Hops With a Hybrid Computational Approach," Comput. J., vol. 00, no. 00, 2020.

[38]. M. Muhammad, T. Alyas, F. Ahmad, F. Butt, W. Qazi, and S. Saqib, "An analysis of security challenges and their perspective solutions for cloud computing and IoT," ICST Trans. Scalable Inf. Syst., pp. 166718, 2018.

[39]. M. Mehmood et al., "Machine learning enabled early detection of breast cancer by structural analysis of mammograms," Comput. Mater. Contin., vol. 67, no. 1, pp. 641–657, 2021.

[40]. N. Iqbal, S. Abbas, M. A. Khan, T. Alyas, A. Fatima, and A. Ahmad, "An RGB Image Cipher Using Chaotic Systems, 15-Puzzle Problem and DNA Computing," IEEE Access, vol. 7, pp. 174051–174071, 2019.

[41]. A. Alzahrani, T. Alyas, K. Alissa, Q. Abbas, Y. Alsaawy, and N. Tabassum, "Hybrid Approach for Improving the Performance of Data Reliability in Cloud Storage Management," Sensors (Basel)., vol. 22, no. 16, 2022.

Khurram *et al. (IJECI)* 2023

(IJECI)
ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

International Journal for
Electronic Crime Investigation

DOI: https://doi.org/10.54692/ijeci.2023.0702154

Research Article

Vol. 7 issue 2 Year 2023

The role of Artificial Intelligence in Cyber Security and Incident Response

Syed Khurram Hassan[1] and Asif Ibrahim[2]

[1] Institute of Quality and Technology Management, University of the Punjab, Lahore, Pakistan.

[2] Department of Mathematics, FC College University, Lahore.

Corresponding author: khuramshah6515@gmail.com

Received: March 10, 2023; **Accepted**: March 25, 2023; **Published**: June 15, 2023

Abstract:

The escalating number and intricacy of cyber attacks have underscored the urgent requirement for innovative solutions to bolster the security of digital infrastructure. Among these solutions, Artificial Intelligence (AI) emerges as a promising technology with the potential to significantly enhance cyber security and incident response. It has the aptitude to improve the speed and accuracy of threat detection, response and mitigation while also reducing the workload on security professionals. This research paper focuses on the role of AI in key areas of cyber security and incident response, specifically vulnerability assessment, intrusion detection and prevention, and digital forensics analysis. It elucidates how AI, with its innate capabilities, can be a game-changer by empowering organizations to detect, respond to, and mitigate threats more effectively. However, AI is not a silver bullet for Cyber Security. Like any technology, it possesses its limitations and potential vulnerabilities. Consequently, this paper also addresses the need for ongoing development in the field of AI to overcome these limitations and challenges. By recognizing the need for continuous advancements, the research paper emphasizes the importance of future research and development efforts to maximize the potential benefits of AI in the realm of cyber security.

Keywords: Innovative solutions, artificial intelligence, cybersecurity, incident response, machine learning, sophisticated attacks, vulnerabilities

1. Introduction

Cyber security is the safety of records/statistics, property, services, and systems of cost to reduce the possibility of loss, damage/corruption, compromise, or misuse to a stage commensurate with the cost assigned. As time-sharing structures emerged within the mid to past-due 1960s and more than one job and users have been capable of running on equal time, controlling the get admission to the facts in the system became a primary point of the subject. One answer that

turned into used become to manner categorized statistics one degree at a time and "sanitize" the device after the jobs from one stage have been run and earlier than the jobs for the subsequent stage were run. This approach to pc protection became referred to as durations processing because the jobs for every level had been all run over their particular length of the day. This becomes an inefficient manner to use the device, and an effort changed into made to locate greater green software solutions to the multilevel security problem. Another approach is including extra functions or mechanisms in a laptop gadget another manner of enhancing laptop security. The mechanisms offered in this phase are grouped into authentication mechanisms, get admission to control and inference manipulation. The other approach to improving the safety of a system is to difficulty the system to rigorous warranty strategies on the way to increase one's self-belief that the system will perform as preferred. Among those strategies are penetration analysis, formal specification and verification, and covert channel evaluation. None of these techniques assure a stable system. The best boom is one's self-belief inside the protection of the gadget [1].

During the Initial Response, the gathering of data regarding the incident that began inside the previous section maintains. The goal is to accumulate enough data to allow the formula of an adequate response method in the next step. Typically, the data this is amassed in this step includes interviews of any individuals concerned in reporting the suspected incident, and available network surveillance logs or IDS reviews, which can suggest that an incident took place. The aim of the formulation of the response strategy is "thinking about the totality of the occasions" that surround the incident. These occasions include the criticality of the affected systems or statistics, what sort of attacker is suspected, and what the overall harm would possibly amount to. A business enterprise's response posture, which defines its coverage concerning the response to pc protection incidents, might also have a big effect on the choice of a reaction method. During the research of the incident, exceptional varieties of proof relevant to the incident, e.g. Host- or network-based proof, are accumulated with the purpose to reconstruct the occasions that comprise the computer protection incident. This reconstruction ought to provide reasons for what came about, when, how, or why it occurred, and who is accountable. To gain this, an investigation is usually divided into two steps: Data Collection and Data Analysis. The cause of the Resolution section is to take the right measures to contain an incident, remedy the underlying troubles that brought on the incident, and take care that a similar incident will now not occur once more. All the important steps completed must be taken and their progress supervised to verify that they may be powerful. Adjustments to the affected systems must be best completed after amassing viable evidence, otherwise, that evidence is probably lost. After the resolution of the incident is entire, it may be necessary to update protection rules or the IR techniques, if the reaction to the incident uncovered a weak spot in contemporary exercise [2].

Artificial intelligence in cyber security is beneficial as it improves how safety professionals examine, look at, and understand cybercrime. It complements the cyber protection technology that businesses use to fight cybercriminals and assist keep groups and customers secure. On the opposite hand, artificial intelligence can be very aid extensive. It may not be sensible in all applications. More importantly, it can also serve as a new weapon inside the arsenal of cybercriminals who use the generation to hone and enhance their cyber attacks. Synthetic intelligence in cyber security is beneficial as it improves how safety professionals examine, look at, and understand cybercrime. It complements the cyber protection technology that businesses use to fight cybercriminals and assist keep groups and customers secure. On the opposite hand, artificial intelligence can be very aid extensive. It may not be sensible in all applications. More importantly, it can also serve as a new weapon inside the arsenal of cybercriminals who use the generation to hone and enhance their cyber attacks. Artificial intelligence is a developing area of interest and investment in the cyber protection community. Let's hash it out. How artificial intelligence cyber security features improve digital safety ideally, if you're like many modern-day corporations, you have more than one tier of protection in location — perimeter, community, endpoint, software, and statistics security measures. For example, you could have hardware or software firewalls and network security answers that track and determine which network connections are allowed and block others. If hackers make it past these defenses, then they'll be up against your antivirus and antimalware solutions. Then possibly they'll face your intrusion detection/intrusion prevention answers (IDS/IPS), and many others [3].

Not a lot of scarce literary resources describing attempts to apply Artificial Intelligence strategies in Incident Handling, however, based on our enjoyment of the introduction of Artificial Intelligence strategies in Tactical, and particularly, Operational Cyber Intelligence, we've got come to the conclusion that gift the primary characteristic of Artificial Intelligence in Incident Handling can be fixing a category challenge, i.e. The unambiguous reference of the modern-day incident to one of the elements of the Classification Scheme, where for every element applicable techniques and workflows have been developed [4].

For the long term, the IR technique has been driven and completed with the aid of people. Automation in the execution of cyber attacks has significantly expanded the tempo with which assaults are now carried out, making it difficult for human analysts to follow. Alert fatigue is a commonplace problem among safety teams that are overwhelmed with the aid quantity and pace of in recent times automated cyber assaults. AI rises as a method to address this problem, being already gifted within the discipline of cyber security, both in literature and security products. AI is also used as an offensive device for carrying out cyber assaults, leading to the necessity of leveraging AI for protection as a way of tackling the speed and volume of such assaults. It is equally important, even though, to not forget the AI it

as a goal for the cyber attack. For the long term, the IR technique has been driven and completed with the aid of people. Automation in the execution of cyber attacks has significantly expanded the tempo with which assaults are now carried out, making it difficult for human analysts to follow. Alert fatigue is a commonplace problem among safety teams that are overwhelmed with the aid quantity and pace of in recent times automated cyber assaults. AI rises as a method to address this problem, being already gifted within the discipline of cyber security, both in literature and security products. AI is also used as an offensive device for carrying out cyber assaults, leading to the necessity of leveraging AI for protection as a way of tackling the speed and volume of such assaults. It is equally important, even though, to not forget the AI as a goal for cyber attack [5].

2. Vulnerability Assessment

In the contemporary interconnected and digitized world, the cybersecurity landscape has grown increasingly intricate and sophisticated. Organizations now confront a myriad of perils posed by cyber criminals who exploit vulnerabilities in their systems and networks, aiming to illicitly access sensitive information, disrupt operations, or inflict financial losses. To confront and mitigate these risks, organizations employ a range of security measures, among which vulnerability assessment emerges as a pivotal component of their comprehensive cybersecurity and incident response strategies. Undoubtedly, vulnerability assessment assumes paramount importance within the realm of cybersecurity administration. It entails the meticulous identification of vulnerabilities present in software and systems, constituting a proactive process of scanning and scrutinizing potential targets and emerging threats with the aim of averting malicious attacks [6]. The domain of Vulnerability Assessment has reached a considerable level of maturity; however, keeping up with the wide range of computing and digital devices requiring scrutiny poses a significant challenge [7]. This practice revolves around a methodical approach to pinpointing and assessing vulnerabilities existing within an organization's IT infrastructure, applications, and systems. It encompasses proactive scanning, testing, and analysis of potential weaknesses that may be exploited by malicious individuals. Conventional approaches to vulnerability assessment have predominantly relied on manual techniques and static rule-based systems, which frequently struggle to match the pace of the evolving threat landscape and the relentless growth in both the volume and intricacy of vulnerabilities [8]. The advent of artificial intelligence (AI) has brought about a transformative shift in the realm of cybersecurity, encompassing vital aspects such as vulnerability assessment and incident response. AI introduces fresh capabilities and efficiencies that hold the potential to greatly enhance the effectiveness and efficiency of these pivotal security processes. Through harnessing machine learning algorithms, natural language processing (NLP), and deep learning methodologies, AI-powered vulnerability assessment empowers organizations to identify, analyze, and address vulnerabilities in a more proactive,

precise, and timely manner. As highlighted by Cybersecurity Ventures, a staggering 111 billion lines of new software code are generated worldwide on an annual basis (Ventures, 2017). By employing automated mechanisms to aid in vulnerability detection prior to system deployment, product teams can dedicate more attention to feature development and performance enhancement. The proliferation of devices and applications being deployed presently not only amplifies the risks associated with networked systems but also furnishes a rich trove of training data for utilization in conjunction with artificial intelligence techniques [9]. The role of AI in vulnerability assessment assumes a multifaceted nature [10].

a) Artificial intelligence (AI) possesses the ability to automate and streamline the entire vulnerability assessment process, mitigating the need for manual efforts and empowering security teams to focus on tasks of greater value. Through the utilization of machine learning algorithms, AI can analyze extensive datasets comprising system logs, network traffic, and historical vulnerability information. This analysis facilitates the identification of patterns and anomalies that may signify potential vulnerabilities. Furthermore, AI can continuously learn and adapt to emerging threats and attack techniques, thereby bolstering the overall resilience of the vulnerability assessment process.

b) AI serves as a catalyst for more advanced and sophisticated vulnerability detection and analysis. Leveraging deep learning techniques, such as Convolution Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs), AI models can extract insightful information from complex datasets, including unstructured sources like security reports, blogs, and research papers. This capability empowers organizations to identify previously unknown vulnerabilities and effectively detect emerging threats.

c) AI-based vulnerability assessment significantly contributes to incident response by expediting the identification of vulnerabilities with greater accuracy. Consequently, security teams can allocate resources and prioritize tasks accordingly. By reducing the time between vulnerability detection and remediation, organizations can substantially diminish their exposure to potential attacks and minimize the impact of security incidents.

Below is a diagram illustrating the vulnerability management life cycle, outlining the optimal steps to assess vulnerabilities within a system:

1. Identify and Uncover Neglected Devices and Assets: Thoroughly examine the network to identify any devices or assets that may have been overlooked or forgotten.

2. Prioritize and Sequence Assets: Evaluate the importance and value that each asset contributes to the company, and prioritize them accordingly.

3. Comprehensive Scanning: Even after

prioritization, leave no stone unturned. Conduct a meticulous scan of every component within the system.

4. Effective Reporting: Establish a streamlined reporting mechanism to promptly communicate any ambiguities or concerns to higher-level staff.

5. Vulnerability Assessment and Ticket Assignment: Assess the vulnerabilities discovered and assign tickets based on the level of risk acceptance and urgency.

6. Solution Verification and Remediation: Verify the effectiveness of applied solutions and ensure they successfully mitigate the identified vulnerabilities.

7. Continuous Improvement: Embrace an iterative approach by repeating the improvement cycle to enhance the assessment process continually.

Figure 1: Vulnerability Assessment Cycle

2.1 How AI can be used for Vulnerability

Security vulnerabilities encompass various flaws and weaknesses found within information technology and its associated products, spanning across different levels and components of an information system. These deficiencies directly impact the smooth operation of the entire information system. When maliciously exploited, they can gravely compromise the integrity, confidentiality, and availability of the system. Consequently, the study of security vulnerabilities stands as a fundamental aspect within the realm of information security research [11]. In light of the escalating complexity of cyber threats, traditional security techniques are no longer sufficient to safeguard against these ever-evolving risks. Consequently, businesses are turning to artificial intelligence (AI) to bolster their cybersecurity strategies. AI offers enhanced capabilities for detecting and responding to threats, bolstering vulnerability management, and improving compliance and governance practices. By leveraging AI technologies such as machine learning, natural language processing, behavioral analytics, and deep learning, organizations can fortify their cyber defenses and shield themselves against a wide array of cyber threats, including malware, phishing attacks, and insider threats. AI has numerous applications in the cyber security industry, including [10].

2.1.1. Threat Detection and Response

AI plays a pivotal role in cyber security by enabling efficient threat detection and response. By leveraging machine learning techniques and natural language processing, organizations can analyze vast amounts of data to identify patterns and anomalies indicative of cyber threats. Intrusion detection systems

powered by AI algorithms monitor network traffic, detecting trends and abnormalities that may signify a security breach. Additionally, AI-driven cyber threat hunting helps uncover and track advanced persistent threats (APTs) lurking within networks. Predictive analytics further empowers organizations to proactively identify and address potential threats before they materialize, bolstering proactive defense strategies [10].

2.1.2. Vulnerability Management

AI is instrumental in effective vulnerability management, offering robust solutions for vulnerability scanning and prioritization. AI-enabled tools assist businesses in identifying and prioritizing issues that require remediation. Vulnerability management encompasses automating tasks such as penetration testing, security policy enforcement, and patch administration. Through AI, penetration testing can be automated, simulating attempts to exploit vulnerabilities and assessing the efficacy of existing security measures [10].

2.1.3. Compliance and Governance

AI finds valuable applications in ensuring compliance and governance within organizations. It aids in risk detection, monitoring adherence to regulations and policies, and enforcing compliance. For instance, AI automates compliance reporting and monitoring, ensuring companies adhere to regulations like HIPAA and GDPR. By analyzing extensive data sets, AI can assess risks, identify potential threats and weaknesses, and provide recommendations for suitable mitigation strategies. Furthermore, AI can automatically detect and prevent policy violations, ensuring policy compliance across the organization [11].

2.2. Identifying Vulnerabilities

There are a lot of ways that we can use in order to automate the process of identifying the vulnerabilities. Some of these ways are listed and explained below:

2.2.1. Automated Code Analysis

Utilizing AI algorithms, software code can undergo comprehensive analysis to unveil potential vulnerabilities. This approach facilitates the early identification of vulnerabilities. Static analysis techniques examine code without executing it, seeking out known code patterns, unsafe practices, or insecure coding methodologies that may give rise to vulnerabilities. Dynamic analysis techniques, on the other hand, involve executing the code in controlled environments, closely monitoring its behavior, and uncovering any security weaknesses. It's worth noting that dynamic analysis, in contrast to static analysis, conducts its evaluation during runtime on a live system. This entails executing the code with specific test cases to fulfill defined coverage criteria, albeit this process tends to be time-intensive [12].

2.2.2. *Network Traffic Analysis*

AI plays a pivotal role in scrutinizing network traffic data to discern anomalies or patterns that may indicate potential vulnerabilities. By monitoring the flow of network traffic, AI algorithms can identify suspicious activities like port scanning, atypical packet behaviors,

or attempted network intrusions. The surge in network traffic coupled with the evolution of artificial intelligence necessitates novel approaches to intrusion detection, malware behavior analysis, and the categorization of internet traffic and other security aspects. Machine learning (ML) exhibits impressive capabilities in addressing these network-related challenges [13].

2.2.3. *Vulnerability Scanning*

vulnerability scanners can autonomously conduct comprehensive scans of systems, networks, or applications to unveil known vulnerabilities. These scanners harness AI techniques to compare the gathered data against established vulnerability databases, exploit frameworks, or attack signatures, discerning the presence of any vulnerabilities [14].

1.1.4. *Behavior Monitoring and Anomaly Detection*

AI algorithms possess the ability to learn and understand typical system or user behaviors, allowing them to identify deviations that could potentially indicate vulnerabilities. Through the analysis of system logs, user activities, or system behaviors, AI systems have the capacity to detect anomalies that may serve as red flags for unauthorized access attempts, privilege escalation, or other security breaches [10].

1.1.5. *Natural Language Processing (NLP)*

Leveraging the power of NLP techniques, textual sources such as security advisories, vulnerability reports, or user feedback can undergo thorough analysis. AI algorithms excel at extracting and scrutinizing pertinent information, recognizing vulnerability-specific keywords, and comprehending the contextual nuances surrounding reported vulnerabilities [10].

3. Machine Learning-based Classification

Through the application of machine learning algorithms, datasets labeled with vulnerability information can serve as training material for code, network traffic, or system log classification. These algorithms acquire the ability to discern whether a given instance is vulnerable or non-vulnerable by assimilating patterns and indicators extracted from historical data. This knowledge empowers them to effectively identify new instances of vulnerabilities based on their learned expertise [10].

4. DEEP LEARNING

Harnessing the potential of deep learning techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), proves valuable in scrutinizing intricate and unstructured data to uncover vulnerabilities. For instance, CNNs excel at processing images depicting software interfaces or network diagrams, while RNNs excel at analyzing sequences of events or logs, enabling the detection of vulnerability-related patterns [10].

5. DATA FUSION

AI systems excel at merging data from diverse sources, such as vulnerability databases, security feeds, or system logs, to construct a comprehensive and holistic perspective of potential vulnerabilities. By correlating information gleaned from these distinct sources, AI algorithms bolster the accuracy and dependability of vulnerability identification, enabling more robust cyber security measures.

6. Intrusion Detection And Prevention

Intrusion detection and prevention involve the continuous monitoring of system logs and also the network traffic to identify potential security breaches. A crucial role in this process by collecting and analyzing large amounts of data in real-time is played by automated security tool. These tools employ various techniques such as signature-based detection, anomaly detection, and behavior-based analysis to identify suspicious activities or patterns that may indicate unauthorized access attempts or other security threats. However, despite the automation provided by these tools, the expertise and judgment of human analysts remain essential. Human analysts are responsible for reviewing the findings and analysis generated by the automated systems. They assess the severity and context of the detected threats, investigate any false positives or false negatives, and determine the appropriate response strategy. Human analysts bring their knowledge, experience, and critical thinking skills to interpret the data, validate the findings, and make informed decisions about how to respond effectively to the detected threats [15].

While automation streamlines the detection process and provides initial insights, human analysts add a layer of intelligence and contextual understanding that cannot be replicated by machines alone. Their involvement ensures that the response to detected threats is tailored to the specific circumstances, aligns with organizational policies and priorities, and minimizes the risk of false positives or unnecessary disruptions to legitimate network activities. Human analysts also play a crucial role in adapting the intrusion detection and prevention systems to evolving threats by continuously learning from new attack techniques and adjusting the system configurations accordingly [15].

In this topic, we will explore the initial three subtopics: Network-based Intrusion Detection and Prevention (NIDP), Host-based Intrusion Detection and Prevention (HIDP), and Intrusion Detection and Prevention Systems (IDPS).

6.1 Network-based Intrusion Detection and Prevention (NIDP)

Network-based Intrusion Detection and Prevention (NIDP) entails the surveillance of network traffic to identify and respond to potential intrusions. NIDP utilizes various techniques to analyze packet-level data and identify abnormal or malicious behavior. A fundamental approach in NIDP is packet analysis, which involves scrutinizing network

packet headers and contents to identify patterns or anomalies indicating potential intrusions. Common techniques employed in packet analysis include deep packet inspection (DPI) and protocol analysis [16].

Anomaly detection is another crucial aspect of NIDP, involving the establishment of baseline behavior for comparison against current network activity to identify deviations. Statistical methods, machine learning algorithms, and behavioral analysis are frequently employed in anomaly detection to identify anomalies. By comparing present network traffic patterns to historical data or predefined thresholds, NIDP systems can generate alerts or implement preventive measures [17].

Signature-based detection is a well-established technique in NIDP, which entails comparing network traffic against a database of known attack signatures. If a match is found, the system raises an alert. Although signature-based detection efficiently identifies known attacks, it may struggle with detecting novel or previously unseen attack patterns. To overcome this limitation, intrusion detection and prevention systems often combine signature-based detection with anomaly-based approaches for heightened security [18].

Network traffic monitoring is an integral part of NIDP, encompassing the collection and analysis of network flow data, including source and destination IP addresses, ports, protocols, and session duration. Through network flow analysis, security administrators can identify suspicious patterns such as abnormal data volumes or unusual communication patterns. Network flow data can also be utilized to visualize network activity and detect patterns that may not be discernible through other analysis techniques [19].

6.2 Host-based Intrusion Detection and Prevention (HIDP)

Host-based Intrusion Detection and Prevention (HIDP) focuses on monitoring activities and events on individual hosts or endpoints to protect against internal network-based attacks. HIDP techniques provide detailed visibility into host-level activities, playing a vital role in safeguarding systems. Log analysis is a key component of HIDP, as system logs contain valuable information regarding host activities, including login attempts, file accesses, system calls, and configuration changes. Analyzing log files enables security analysts to identify suspicious or unauthorized activities. Automated log analysis tools aid in detecting patterns or events of interest, facilitating efficient intrusion detection [20].

System call monitoring is another important HIDP technique that involves capturing and analyzing system calls made by programs or processes running on a host. By monitoring system calls, HIDP systems can detect malicious or abnormal behavior, such as unauthorized access attempts, privilege escalation, or file manipulation. Anomalies detected through system call monitoring can trigger alerts or proactive measures to mitigate potential risks. File integrity checking is a mechanism employed to ensure the integrity of critical system files. HIDP systems often main-

tain hash or checksum values for each file and periodically verify their integrity by recalculating the hash and comparing it with the stored value. The detection of discrepancies indicates potential file modifications or tampering, which could signify a security breach [21].

Behavior-based detection techniques in HIDP involve the continuous monitoring and analysis of process and application behavior running on hosts. This approach focuses on identifying deviations from expected behavior patterns, allowing for the detection of abnormal or potentially malicious activities. In conclusion, NIDP, HIDP, and IDPS form essential subtopics in intrusion detection and prevention. By utilizing techniques such as packet analysis, anomaly detection, signature-based detection, log analysis, system call monitoring, and behavior-based detection, organizations can enhance their ability to identify and prevent intrusions, safeguarding their networks and systems from malicious activities [22].

6.3 Intrusion Detection and Prevention Systems (IDPS)

Intrusion Detection and Prevention Systems (IDPS) play a vital role in the identification and response to intrusions in computer networks and systems. These systems are designed to continuously monitor network traffic, host activities, and system logs, offering real-time capabilities for detecting and preventing threats. IDPS can operate in different modes, including network-based, host-based, or a combination of both, to provide comprehensive security coverage. IDPS are built on a combination of technologies, methodologies, and algorithms to recognize and mitigate security threats. To find malicious actions and potential vulnerabilities, they use cutting-edge detection techniques like signature-based detection, anomaly detection, and behavior-based analysis. [23].

Signature-based detection in IDPS involves comparing network traffic, host data, or system logs against known attack signatures or patterns. These signatures are derived from previously identified and documented malicious activities. If a match is found, the IDPS generates an alert, enabling security personnel to take appropriate actions. Signature-based detection is effective in identifying known attacks but may face challenges in detecting new or unknown attacks that lack pre-existing signatures. Anomaly detection is another essential component of IDPS. This technique involves establishing a baseline of normal behavior for the network or host and comparing ongoing activities against this baseline. Any deviation or anomaly from the established norm may indicate a potential intrusion. Anomaly detection algorithms utilize statistical methods, machine learning, and behavioral analysis to identify unusual patterns, network traffic spikes, or abnormal system behavior. By generating alerts based on detected anomalies, IDPS can aid in the detection of previously unseen or evolving threats. Behavior-based analysis is a proactive approach employed in IDPS to identify malicious activities based on the observed behavior of network traffic, applications, or system processes. By analyzing the sequence of actions, resource access patterns, or commu-

nication behavior, IDPS can detect deviations from expected behavior and raise alerts. Behavior-based analysis is particularly effective in detecting sophisticated attacks that may evade signature-based detection [24].

Apart from detection, IDPS also prioritize prevention and response. When a potential intrusion or suspicious activity is detected, IDPS can take various actions to prevent further harm or reduce the impact. These actions may involve blocking network traffic, isolating compromised hosts, resetting user sessions, or notifying security personnel for further investigation. IDPS can also integrate with other security systems, such as firewalls, to automatically enforce access control policies or update rule sets to enhance overall security.

6.4. *AI and Intrusion Detection and Prevention*

AI can augment the capabilities of human analysts and traditional security tools in several ways. Here are some examples:

6.4.1. Real-time monitoring

AI algorithms can analyze network traffic and system logs in real-time, allowing them to quickly identify and respond to potential threats. This is particularly useful in large or complex networks, where it may be difficult for human analysts to keep track of all the activity. AI can also flag potential threats that might otherwise go unnoticed by human analysts, such as low-level attacks that are designed to evade detection [25].

6.4.2. Anomaly detection

AI can be trained to recognize normal patterns of network activity, and to flag any deviations from these patterns that might indicate the presence of a cyber threat. For example, AI can detect unusual login activity, identify attempts to exploit known vulnerabilities and alert security teams to potential threats that might otherwise go unnoticed. By detecting potential threats at an early stage, AI can help to minimize the damage caused by a cyber-attack [26].

6.4.3. Automated response

Automated response in cyber security refers to the use of AI-powered tools and algorithms to automatically perform certain actions in response to detected threats or security incidents. These automated actions help to prevent the spread of cyber-attacks and mitigate their impact. Let's explore an example to better understand how automated response works. Imagine a large organization with a sophisticated AI-powered intrusion detection system in place. This system continuously monitors the network for any suspicious activities or potential cyber threats. One day, the intrusion detection system identifies a series of network packets exhibiting patterns indicative of a DDoS (Distributed Denial of Service) attack. Upon detecting this potential threat, the AI-powered security tool automatically springs into action. It analyzes the incoming network traffic, identifies the malicious packets, and determines the best course of action to mitigate the attack. In this case, the AI system decides to block the IP addresses associated with the attacking packets. Using its

automated response capabilities, the AI tool sends instructions to the organization's network infrastructure, specifically the firewalls or routers. These instructions result in the immediate blocking of the identified IP addresses, effectively stopping the malicious traffic from reaching the organization's network resources. Simultaneously, the AI system also initiates actions to isolate any infected systems within the organization's network. It identifies the compromised devices, such as computers or servers that may be participating in the DDoS attack, and quarantines them from the rest of the network. By isolating the infected systems, the AI tool prevents the attack from spreading further and causing additional damage to other network components [27].

In this scenario, the automated response capabilities provided by AI-powered security tools play a vital role in containing and mitigating the DDoS attack. By automatically blocking suspicious traffic and isolating infected systems, the AI system helps prevent the attack from disrupting the organization's network services and causing significant downtime. Furthermore, by automating these routine response tasks, the AI system reduces the workload on human analysts. Instead of spending time manually identifying and blocking malicious traffic, analysts can focus on more complex and strategic security tasks, such as investigating the root cause of the attack, identifying potential vulnerabilities, or fine-tuning the AI system's response algorithms.

Overall, the example highlights how automated response, facilitated by AI, can enhance an organization's ability to respond quickly and effectively to cyber threats. By leveraging AI's speed and precision, organizations can reduce response times, minimize the impact of attacks, and improve the efficiency of their security operations [28].

6.4.4. Predictive analysis

AI can also be used for predictive analysis, which involves using historical data to identify potential future threats. By analyzing patterns and trends in network activity over time, AI algorithms can identify potential vulnerabilities and anticipate potential threats before they occur. This can help organizations to proactively mitigate these threats before they can cause any damage.

However, it's important to remember that AI is not a panacea for all cyber security challenges, and it should be used in conjunction with other tools and techniques. For example, AI algorithms may not be able to detect advanced persistent threats (APTs) or zero-day vulnerabilities, which require human expertise and intuition to identify. Additionally, AI algorithms may be susceptible to false positives or false negatives, which can lead to unnecessary alerts or missed threats [29].

7. Incident Response

Although they use different process methodologies, incident response and computer forensics have similar goals. While both situations' primary goals are to investigate computer

security incidents and contain their effects, incident response is more focused on bringing things back to normal while computer forensics is more focused on producing evidence that can be used in court.

An organization's response to improper or undesirable behavior using a computer or network component is known as an incident response. A methodical and well-planned approach should be employed to react rather than being caught off guard and launching a disorderly and potentially disastrous response. As a result, events are typically handled by a team known as the Computer Security Incident Response Team, or CSIRT, which is made up of individuals who possess the various certifications required for the response procedure [30].

7.*1. Real time analysis of security events*

The gathering, storing, and analyzing of all data relating to the incident that has occurred or is still occurring is one of the key activities in dealing with cyber security incidents [31].

Detecting security threats in real time is the responsibility of the security operations centre (SOC), a centralised organisation. It is an essential part of a CSIRT (Corporate Security Incident Response Team). A key piece of technology used in SOCs, SIEM systems collect security events from various sources within enterprise networks, normalise the events to a standard format, store the normalised events for forensic analysis, and correlate the events to detect malicious activities in real time. The authors of this essay emphasise the critical role SIEM systems play for SOCs, address current operational barriers to properly employing SIEM systems, and identify upcoming technical problems that SIEM systems will need to overcome to remain relevant [32].

7.2. *Automated Incident Triage*

Recent years have seen a dramatic rise in the number of computer security incidents across all industries. Even small businesses suffer significant financial and reputational losses as a result of these accidents. Naturally, there has been a rise in demand for incident management relating to computers. Today, incident handling is still a challenging job that is primarily carried out by human expert teams. It is exceedingly expensive to retain such a team on call around-the-clock, especially in large organizations with extensive networks. Consequently, it is highly desirable to have automated incident handling. It was extremely difficult to automate this process due to its complexity and reliance on humans [33].

Data triage is used by Security Operation Centers to separate the real "signals" from a lot of noisy alerts and "connect the dots" to answer some higher-level questions about the activities of the attack. This work intends to naturally produce information emergency robots straightforwardly from network safety investigators' activity follows. Data triage automatons that are currently in use, such as SIEMs and Security Information and Event Management systems (SIEMs), require expert analysts to dedicate time and effort to the creation of event correlation rules [34].

7.3. *Role of AI in Incident Response*

Artificial intelligence (AI) has a significant role to play in incident response, particularly in the early detection and rapid response to security incidents. AI-powered systems can monitor and analyze vast amounts of data and quickly identify anomalous behaviors or patterns that may indicate a potential security breach.

Here are some ways in which AI can help with incident response:

7.3.1. Early detection

Early detection is a crucial aspect of cyber security as it allows organizations to identify potential threats and take proactive measures to mitigate them. Artificial Intelligence (AI)-powered systems play a significant role in enhancing early detection capabilities by monitoring network traffic, endpoints, and critical infrastructure for any signs of unusual activity or behavior. AI-powered systems leverage advanced algorithms and machine learning techniques to analyze vast amounts of data in real-time. By establishing a baseline of normal network behavior, these systems can identify anomalies that may indicate the presence of a threat. These anomalies could be deviations from typical patterns, such as unexpected network traffic spikes, unauthorized access attempts, or unusual data transfers. One of the significant advantages of AI-powered systems is their ability to detect threats that may go unnoticed by human analysts. While human analysts play a critical role in cybersecurity, they are limited by their capacity to process large volumes of data and to recognize subtle patterns or anomalies. AI systems, on the other hand, can analyze massive amounts of data quickly and efficiently, allowing them to identify potential threats in near real-time. To achieve early detection, AI systems employ various techniques. One common approach is anomaly detection, where AI algorithms learn from historical data to establish normal patterns of network behavior. They then continuously monitor incoming data and compare it to the established baseline. Any deviation from the norm triggers an alert, indicating a potential security threat. Another technique used by AI-powered systems is behavioral analysis. These systems monitor and analyze the behavior of endpoints, such as individual devices or users, to identify any abnormal activities. By learning from historical data and establishing typical user behaviors, AI algorithms can identify behavior that deviates from the norm, which may suggest malicious intent or compromised endpoints [35].

7.3.2. Rapid response

AI systems play a crucial role in alerting security teams to potential security incidents, enabling them to respond promptly and mitigate the impact of the incident. Through continuous monitoring and analysis of network traffic, endpoints, and critical infrastructure, AI-powered systems can quickly identify anomalies and suspicious activities that may indicate a security breach or cyber attack. When an AI system detects unusual activity or behavior, it generates an alert that is immediately relayed to the security team. These alerts serve as early warnings, providing crucial

information about potential threats before they can cause significant harm. By leveraging advanced algorithms and machine learning techniques, AI systems can differentiate between normal and abnormal patterns, helping to identify potential security incidents in real-time. The quick alerting capability of AI systems is beneficial for several reasons. First, it allows security teams to respond swiftly, minimizing the time window for attackers to exploit vulnerabilities or escalate their activities. By receiving alerts in near real-time, security professionals can take immediate action to investigate and contain the incident, preventing further compromise of systems and data. Second, early detection and rapid response help mitigate the impact of security incidents. By identifying threats at an early stage, organizations can limit the potential damage caused by unauthorized access, data breaches, or malicious activities. Security teams can implement appropriate countermeasures, such as isolating affected systems, blocking malicious traffic, or initiating incident response protocols to contain and mitigate the incident swiftly [36].

7.3.3. Automated investigation

AI can help automate the process of investigating security incidents. This can help reduce the time and resources required to identify and remediate security issues.

7.3.4. Threat intelligence

AI can analyze vast amounts of threat intelligence data and provide insights into emerging threats and vulnerabilities. This can help security teams stay ahead of the curve and proactively address potential security risks.

7.3.5. Behavioral analysis

AI can analyze user behavior and identify anomalous patterns that may indicate insider threats or other malicious activity [37].

8. Forensics Analysis

The development of digital technology over the past ten years has had a significant impact on our day-to-day lives and business practices. As a result, the digital forensics field will face numerous challenges as this evolution continues [38].

The goal of forensic analysis is to uncover and interpret evidence that can help investigators understand what happened, identify potential suspects or perpetrators, and provide evidence for use in court. Forensic analysts may work for law enforcement agencies, government agencies, or private companies, and their work may be used in criminal investigations, civil lawsuits, and other legal proceedings. Therefore, Digital forensics is a complex and evolving field. To conduct effective forensic analysis in cyber security, analysts must have a deep understanding of computer systems, network protocols, and cyber threats. They must also be familiar with the legal and regulatory requirements for handling digital evidence, as well as the ethical considerations involved in handling sensitive data [39].

9. How AI Can Assist in Forensics Analysis

Compared to other application domains, digital forensics appears to have used automation and AI less frequently[40]. Artificial intelligence (AI) has the potential to significantly aid forensic analysis in a number of ways. Here are a few instances:

9.1. Data Analysis

AI can analyze vast amounts of data collected during forensic investigations, including network traffic logs, system logs, and other digital evidence. With machine learning algorithms, AI can identify patterns and anomalies in the data, which may be indicative of a cyber attack or other malicious activity [37].

9.2. Image and Audio Analysis

When it comes to image analysis, AI algorithms can be trained to identify and classify objects, faces, and other visual elements within images. By utilizing deep learning models and neural networks, AI can accurately detect and recognize specific objects or individuals. This capability proves invaluable in forensic investigations where identifying suspects or potential evidence is crucial. AI systems can rapidly process large volumes of images and flag relevant information, significantly reducing the time and effort required for manual examination. Furthermore, AI can assist in facial recognition, comparing faces captured in images or video footage against databases of known individuals. This technology can help identify persons of interest or locate missing individuals by matching faces from surveillance footage, social media images, or other sources. AI-powered facial recognition systems have been instrumental in solving numerous criminal cases by linking suspects to evidence or establishing the presence of certain individuals at crime scenes. In the context of video analysis, AI algorithms can analyze video content to extract meaningful information. This includes tracking the movement of objects or individuals, detecting specific activities or behaviors, and identifying important events within the footage. AI can also perform forensic video enhancement, enhancing the quality of low-resolution or poorly captured videos to improve visibility and aid in identifying key details. These capabilities enable investigators to reconstruct events, identify patterns, and gather evidence from video recordings more efficiently [41].

9.3. Predictive Analytic

Predictive analytic is a type of data analysis that uses machine learning algorithms to analyze historical data and identify patterns and trends that can be used to predict future events. In the context of cyber security, predictive analytic can be used to identify potential security threats or vulnerabilities by analyzing historical data from previous incidents. Predictive analytic models driven by AI can examine a large amount of data from a variety of sources, including system logs, network traffic logs, and other digital evidence. The models can spot trends and oddities in the data that might point to a security risk, such a cyberattack attempt or a system weakness that could be used by hackers. By using these predictive models, security teams can be alerted to potential security breaches in real-time, allowing them to take proactive steps to prevent or

mitigate the damage caused by a cyber attack. For example, if a predictive model identifies a potential threat in real-time, security teams can investigate the issue and take steps to prevent the attack before it causes any damage. The use of predictive analytics in cyber security can help organizations to stay ahead of potential security threats and to anticipate new attack methods, allowing them to implement proactive security measures to prevent cyber attacks. Additionally, predictive analytic can be used to identify vulnerabilities in systems and applications, enabling organizations to take corrective action to secure their infrastructure and reduce the risk of a successful attack [42].

9.4. Natural Language Processing (NLP)

AI-powered NLP algorithms can analyze text data, such as emails, chat logs, and social media posts, to identify keywords or phrases that may be related to an incident. This can help investigators identify potential suspects or gain insights into the motives behind an attack [43].

9.5. Malware Analysis

AI can help in analyzing malware by detecting and classifying malicious code. It can also identify patterns in the behavior of malware to help investigators identify its origin and the extent of the damage caused. The makers of the Magnet Axiom forensic examination tool, Magnet Forensics, included machine learning in their Magnet [44].

10. Identifying The Source And Cause Of A Security Incident

Forensic analysis plays a critical role in determining the origin and cause of a security incident. It involves a systematic examination of digital evidence to understand what happened, how it occurred, who was responsible, and the extent of the damage [45]. Below are steps involved in conducting forensic analysis to identify the source and cause of a security incident:

1. ***Secure the Affected System:*** The initial step is to isolate and secure the affected system to prevent further harm or data loss. This may entail disconnecting the system from the network or taking it offline.
2. ***Document the Incident:*** Promptly document the incident by taking comprehensive notes, photographs, or videos of the affected system. Capture relevant information like error messages, timestamps, or any unusual behavior observed.
3. ***Preserve Evidence:*** To maintain the integrity of the evidence, create a forensic copy of the affected system's storage media. This involves making a bit-by-bit replica of the entire storage device or disk partition. The copy will be used for analysis while leaving the original evidence untouched.
4. ***Conduct Initial Analysis:*** Analyze system logs, network traffic logs, firewall logs, intrusion detection system (IDS) logs, and other relevant data sources to gather initial information about the incident. Look for signs of unauthorized access, unusual activities, or anomalies.
5. ***Recover Deleted or Hidden Data:***

Employ forensic tools and techniques to recover deleted or concealed data that may provide valuable insights into the incident. This may involve examining temporary files, registry entries, or system artifacts that can shed light on the source and cause.

6. ***Perform Malware Analysis:*** If malware is suspected, conduct a detailed analysis of suspicious files or software. Use specialized tools to analyze the malware's behavior, identify its characteristics, and determine its origin.
7. ***Network Traffic Analysis:*** Scrutinize network traffic logs, packet captures, or firewall logs to identify any suspicious or unauthorized network activity. Look for indicators of unauthorized access, data exfiltration, or communication with known malicious entities.
8. ***Timeline Reconstruction:*** Create a timeline of events based on the gathered evidence. This timeline should outline the sequence of actions leading up to and following the incident. It can help identify the initial compromise and the attacker's activities throughout the attack.
9. ***User and System Analysis:*** Analyze user accounts, system configurations, and access controls to identify potential vulnerabilities or weaknesses that may have been exploited during the incident. Look for signs of unauthorized access or privilege escalation.
10. ***Collaboration and Expert Consultation:*** In complex cases, collaborate with other experts such as network administrators, incident response teams, or law enforcement agencies. Their expertise and resources can assist in the investigation and analysis process.
11. ***Report Findings:*** Prepare a detailed report summarizing the forensic analysis findings. Include a description of the incident, the methods used for analysis, the identified source and cause of the incident, and recommendations for preventing future incidents.

It's important to note that forensic analysis is a specialized field, and it is advisable to involve experienced professionals or a dedicated incident response team to ensure a comprehensive and accurate investigation.

12. Data Carving

Data carving is a fundamental technique employed in the field of digital forensics to retrieve fragmented or deleted files from storage media. It involves the identification and reconstruction of files based on their distinct signatures or patterns, circumventing the structure of the file system. Data carving proves particularly valuable when conventional file recovery methods are ineffective or when dealing with intentionally erased or damaged files [46].

The process of data carving entails scouring the raw binary data of a storage device in search of specific file headers, footers, or other data patterns. These patterns serve as indicators suggesting the presence of a particular file

type, such as documents, images, videos, or archives. By recognizing these signatures, data carving tools can extract and reconstruct files from the scattered or unallocated space on the storage medium [47].

Data carving algorithms typically function by scrutinizing the binary data and identifying distinct patterns or structures that signify the beginning and end of a file. Once a potential file is detected, the carving tool proceeds to extract the file by copying the corresponding data blocks into a separate file, ultimately generating a reconstructed version of the original file. One of the primary challenges encountered in data carving involves handling fragmented files. Due to factors like partial overwriting or deletion, files on a storage device are often stored in non-contiguous clusters or sectors. Data carving algorithms must possess the ability to identify and assemble these dispersed fragments in order to accurately reconstruct the complete file [48].
Another obstacle involves the potential occurrence of false positives or false negatives during the data carving process. False positives arise when the carving tool incorrectly identifies non-file data as a file, which can lead to the recovery of irrelevant or corrupted data. Conversely, false negatives occur when a carving tool fails to identify and recover a valid file.To enhance the accuracy and efficiency of data carving, a range of techniques and heuristics have been developed. These include advanced signature matching algorithms, file format-specific carving, entropy analysis, and error correction mechanisms [49].
Data carving plays a critical role in digital forensics, enabling investigators to retrieve valuable evidence from storage media, even in cases where the file system has been compromised or intentionally tampered with. It is an indispensable tool in investigations related to cyber crime, data breaches, intellectual property theft, and other digital offenses [50].

12. CONCLUSION

The role of artificial intelligence (AI) in cyber security and incident response is constantly evolving and holds great potential for future developments. Looking ahead, the future of cyber security will likely be shaped by emerging technologies such as quantum computing, 5G networks, and the increasing integration of AI and automation. These advancements bring new opportunities but also introduce novel security risks and challenges that will require proactive measures and innovative solutions.

13. REFERENCES

[1]. Kemmerer, R. A. (2003). Cyber security. 25th International Conference on Software Engineering, 2003.

[2]. C.Felix . Freiling Laboratory for Dependable Distributed Systems University of Mannheim, Bastian Schwittay Symantec (Deutschland) GmbH,

[3]. LJUBOMIR LAZIĆ Belgrade Metropolitan University, Faculty of Information Technologies, BENEFIT FROM AI IN CYBERSECURITY the 11th International Conference on Business Information Security, 18th October

2019, Belgrade, Serbia

[4]. R.Trifonov, R.Yoshinov, S.Manolov, G.Tsoche, & G.Pavlova. Artificial Intelligence methods are suitable for Incident Handling Automation. MATEC Web of Conferences, 292, 01044.

[5]. Vasileios Anastopoulos, PhD Davide Giovannelli, LL. M./05/2022/Automated/Autonomous Incident Response[Online]. Available: https://www.bath.ac.uk/publications/library-guides-to-citing-referencing/attachments/ieee-style-guide.pdf

[6]. B. Seumo, "Cyber Security Administration Introduction," in Conf. Introduction to Cyber Security Administration by Dr. Blondel Seumo, Dubai, United Arab Emirates, 2023, pp.2.

[7]. D. R. McKinnel, T. Dargahi, A. Dehghantanha, K. -K. R. Choo, "A systematic literature review and meta-analysis on artificial intelligence in penetration testing and vulnerability assessment," C&EE, vol.75, pp. 175-188, 2019.

[8]. J. Agarwal, M. Liu, D. Blockley, "Vulnerability, Uncertainty, and Risk Analysis, Modeling and Management," in Conf. Proceedings of First International Conference on Vulnerability and Risk Analysis and Management, Hyattsville, Maryland, 2011, pp. 230-237.

[9]. S. Kommrusch, "Artificial Intelligence Techniques for Security Vulnerability Prevention," Fort Collins, USA, 2018.

[10]. S. A. Jawaid, "Artificial Intelligence with respect to Cyber Security," Vienna, USA, 2023.

[11]. Q. Zhu, L. Liang, "Research on Security Vulnerabilities Based on Artificial Intelligence," in ICIC, 2019, pp. 377-387.

[12]. D. Baca, B. Carlsson, K. Petersen, L. Lundberg, "Improving software security with static automated code analysis in an industry setting," Softw. Pract. Exper., 2012.

[13]. N. Alqudah, Q. Yaseen, "Machine Learning for Traffic Analysis: A Review," in Conf. International Workshop on Data-Driven Security (DDS 2020), Warsaw, Poland, 2020, pp. 911-916.

[14]. N. Schagen, K. Koning, H. Bos, C. Giuffrida, "Towards Automated Vulnerability Scanning of Network Servers," Portugal, 2018

[15]. Y. S. Park, C. S. Choi, C. Jang, D. G. Shin, G. C. Cho and H. S. Kim, "Development of Incident Response Tool for Cyber Security Training Based on Virtualization and Cloud," 2019 International Workshop on Big Data and Information Security (IWBIS), Bali, Indonesia, 2019, pp. 115-118, doi: 10.1109/IWBIS.2019.8935723.

[16]. L. A. H. Ahmed and Y. A. M. Hamad, "Machine Learning Techniques for Network-based Intrusion Detection System: A Survey Paper," IEEE Xplore, Mar. 01, 2021. https://ieeexplore.ieee.org/document/9428827 (accessed Oct. 28, 2022)

[17]. D.-S. Kim and Jong Chun Park, "Network-Based Intrusion Detection

with Support Vector Machines," pp. 747–756, Feb. 2003, doi: https://-doi.org/10.1007/978-3-540-45235-5_73 .

[18]. P. Ioulianou, V. Vassilakis, I. Moscholios, and M. Logothetis, "This is a repository copy of A Signature-based Intrusion Detection System for the Internet of Things. A Signature-based Intrusion Detection System for the Internet of Things," 2018. Available: https://eprints.whiterose.ac.uk/133312/1/ictf_2018_IoT.pdf

[19]. P. M and S. Bose, "Design of Intrusion Detection and Prevention System (IDPS) using DGSOTFC in collaborative protection networks," IEEE Xplore, Dec. 01, 2013. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6921946 (accessed Dec. 14, 2021).

[20]. A. N. Jaber, M. F. Zolkipli, H. A. Shakir, and M. R. Jassim, "Host Based Intrusion Detection and Prevention Model Against DDoS Attack in Cloud Computing," Advances on P2P, Parallel, Grid, Cloud and Internet Computing, pp. 241–252, Nov. 2017, doi: https://-doi.org/10.1007/978-3-319-69835-9_23 .

[21]. "(PDF) Host-based Intrusion Detection and Prevention System (HIDPS)," ResearchGate. https://www.researchgate.net/publication/271070098_Host based_Intrusion_Detection_and_Prevention_System_HIDPS

[22]. A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," Cybersecurity, vol. 2, no. 1, Jul. 2019, doi: https://-doi.org/10.1186/s42400-019-0038-7.

[23]. K. Scarfone and P. Mell, "Special Publication 800-94 Guide to Intrusion Detection and Prevention Systems (IDPS) Recommendations of the National Institute of Standards and Technology," 2007. Available: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-94.pdf

[24]. A. Sharifi, F. F. Zad, F. Farokhmanesh, A. Noorollahi, and J. Sharif, "An Overview of Intrusion Detection and Prevention Systems (IDPS) and Security Issues," IOSR Journal of Computer Engineering, vol. 16, no. 1, pp. 47–52, 2014, doi: https://-doi.org/10.9790/0661-16114752.

[25]. Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," IEEE Access, vol. 10, pp. 93104–93139, 2022, doi: https://-doi.org/10.1109/access.2022.3204051.

[26]. "Artificial Intelligence in CyberSecurity," IEEE Access, Mar. 11, 2019. https://ieeeaccess.ieee.org/closed-special-sections/artificial-intelligence-in-cybersecurity/

[27]. M. Khanbhai, P. Anyadi, J. Symons, K. Flott, A. Darzi, and E. Mayer, "Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review," BMJ Health & Care Informatics, vol. 28, no. 1, p. e100262, Mar. 2021, doi: https://doi.org/10.1136/bmjh-

ci-2020-100262.

[28]. V. Mathane and P. V. Lakshmi, "Predictive Analysis of Ransomware Attacks using Context-aware AI in IoT Systems," International Journal of Advanced Computer Science and Applications, vol. 12, no. 4, 2021, doi: https://doi.org/10.14569/ijacsa.2021.0120432

[29]. [1] F. C. Freiling and Bastian Schwittay, "A Common Process Model for Incident Response and Computer Forensics.," ResearchGate, 2007. https://www.researchgate.net/publication/221002732_A_Common_-Process_Model_for_Incident_Response_and_Computer_Forensics

[30]. [2] M. Evans et al., "Real-Time Information Security Incident Management: A Case Study Using the IS-CHEC Technique," IEEE Access, 2019. https://www.semanticscholar.org/paper/Real-Time-Information-Security-Incident-Management%3A-Evans-He/49dbe6d1ddc0862726351c939ebedf5811e15bf5 (accessed May 02, 2023).

[31]. [3] S. Bhatt, P. K. Manadhata, and L. Zomlot, "The Operational Role of Security Information and Event Management Systems," IEEE Security & Privacy, vol. 12, no. 5, pp. 35–41, Sep. 2014, doi: https://doi.org/10.1109/msp.2014.103.

[32]. [4] S. H. Hashemi, M. Babaeizadeh, M. Nowruzi, H. H. Jazi, M. Shahmoradi, and E. B. Beigi Samani, "A comprehensive semi-automated incident handling workflow," IEEE Xplore, Nov. 01, 2012. https://ieeexplore.ieee.org/document/6483144 (accessed May 02, 2023).

[33]. [5] C. Zhong, J. Yen, P. Liu, and R. F. Erbacher, "Automate Cybersecurity Data Triage by Leveraging Human Analysts' Cognitive Process," IEEE Xplore, Apr. 01, 2016. https://ieeexplore.ieee.org/document/7502316 (accessed May 02, 2023).

[34]. [6] Z. Chen, Y. Kang, P. Zhao, B. Qiao, and Q. Lin, "Towards intelligent incident management: why we need it and how we make it," Semantic Scholar, 2020, doi: https://-doi.org/10.1145/3368089.3417055.

[35]. [7] B. Ai, B. Li, S. Gao, J. Xu, and H. Shang, "An Intelligent Decision Algorithm for the Generation of Maritime Search and Rescue Emergency Response Plans," IEEE Access, vol. 7, pp. 155835–155850, 2019, doi: https://-doi.org/10.1109/ACCESS.2019.2949366.

[36]. [8] J. Williams, "A SANS Survey Written by Alissa Torres," 2014. Available: https://csbweb01.uncw.edu/people/cummingsj/classes/mis534/Articles/Ch3_IR_SANSSurvey.pdf

[37]. N. M. Karie and H. S. Venter, "Taxonomy of Challenges for Digital Forensics," Journal of Forensic Sciences, vol. 60, no. 4, pp. 885–893, Jul. 2015, doi: https://-doi.org/10.1111/1556-4029.12809.

[38]. M. Pollitt, "A History of Digital Forensics," Advances in Digital Forensics VI, vol. 337, pp. 3–15, 2010, doi: https://-doi.org/10.1007/978-3-642-15506-2_1.

[39]. A. Jarrett and K. R. Choo, "The impact

of automation and artificial intelligence on digital forensics," WIREs Forensic Science, Apr. 2021, doi: https://-doi.org/10.1002/wfs2.1418.

[40]. S. Ikram and H. Malik, "Digital audio forensics using background noise," IEEE Xplore, Jul. 01, 2010. https://ieeexplore.ieee.org/abstract/document/5582981 (accessed Mar. 27, 2022).

[41]. V. R. Kebande et al., "Towards an Integrated Digital Forensic Investigation Framework for an IoT-Based Ecosystem," IEEE Xplore, Aug. 01, 2018. https://ieeexplore.ieee.org/abstract/document/8465532/

[42]. D. O. Ukwen and M. Karabatak, "Review of NLP-based Systems in Digital Forensics and Cybersecurity," 2021 9th International Symposium on Digital Forensics and Security (ISDFS), Jun. 2021, doi: https://-doi.org/10.1109/isdfs52919.2021.9486354.

[43]. S. W. Hall, A. Sakzad, and K. R. Choo, "Explainable artificial intelligence for digital forensics," WIREs Forensic Science, Jun. 2021, doi: https://-doi.org/10.1002/wfs2.1434.

[44]. N. H. Ab Rahman and K.-K. R. Choo, "A survey of information security incident handling in the cloud," Computers & Security, vol. 49, pp. 45–69, Mar. 2015, doi: https://-doi.org/10.1016/j.cose.2014.11.006.

[45]. B. B. Meshram and D. N. Patil, "Digital Forensic Analysis of Hard Disk for Evidence Collection," International Journal of Cyber-Security and Digital Forensics (IJCSDF), vol. 7, no. 2, pp. 100–110, 2018, Accessed: May 24, 2023. [Online]. Available: http://sdiwc.net/digital-library/digital-forensic-analysis-of-hard-disk-for-evidence-collection

[46]. D. Povar and V. K. Bhadran, "Forensic Data Carving," Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp. 137–148, 2011, doi: https://-doi.org/10.1007/978-3-642-19513-6_12.

[47]. A. Pal and N. Memon, "The evolution of file carving," IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 59–71, Mar. 2009, doi: https://-doi.org/10.1109/msp.2008.931081.

[48]. M. Meyers and M. Rogers, "Computer Forensics: The Need for Standardization and Certification," International Journal of Digital Evidence Fall, vol. 3, no. 2, 2004, Available: https://www.utica.edu/academic/institutes/ecii/publications/articles/A0B7F51C-D8F9-A0D0-7F387126198F12F6.pdf

[49]. G. Cantrell, "Teaching Data Carving Using The Real World Problem of Text Message Extraction From Unstructured Mobile Device Data Dumps," The Journal of Digital Forensics, Security and Law, 2019, doi: https://-doi.org/10.15394/jdfsl.2019.1603.

[50]. K. Zhang and A. B. Aslan, "AI technologies for education: Recent research & future directions," Computers and Education: Artificial Intelligence, vol. 2, p. 100025, 2021, doi: https://-doi.org/10.1016/j.caeai.2021.100025.

Ashar *et al. (IJECI) 2022*

International Journal for
Electronic Crime Investigation

(IJECI)
ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

DOI: https://doi.org/10.54692/ijeci.2023.0702155

Research Article

Vol. 7 issue 2 Year 2023

Detecting Phishing Websites using Decision Trees: A Machine Learning Approach

Ashar Ahmed Fazal[1] and Maryam Daud[2]

[1]Department of Criminology and Forensic Sciences, Lahore Garrison University, Lahore

[2]University of Engineering and Technology, Lahore

Corresponding author: asharahmed.ash@gmail.com

Received: March 12, 2023; **Accepted**: March 29, 2023; **Published**: June 15, 2023

Abstract

This study emphasises the value of feature selection and preprocessing in improving model performance and demonstrates the efficiency of decision trees in identifying phishing websites. Internet users are significantly threatened by phishing websites, hence a strong detection strategy is required. The Phishing Websites Dataset from the UCI Machine Learning Repository, which contains 30 website-related features, is used in the study together with a decision tree classifier from the scikit-learn package. The dataset is preprocessed to remove invalid and missing values, and the most pertinent features are chosen for model training. 80% of the dataset is utilised to train the model, while the remaining 20% is used for testing. The findings demonstrate the decision tree classifier's precision in detecting phishing websites, scoring 95.97% accurate and showing a high true positive rate (96.64%) and a negligible (3.04%) false positive rate using the confusion matrix. This study highlights the significance of feature selection and preprocessing for optimal model performance in addition to validating the efficacy of decision trees in phishing detection. The method described here can be helpful for businesses and individuals looking to protect themselves from phishing assaults, and the given data visualisations make it easier to understand datasets and assess models.

1. Introduction

1.1. *Background and Motivation*

Phishing attacks are a serious threat to online security, with the potential to cause significant financial and personal harm to users. Phishing attacks involve the use of deceptive emails or websites that are supposed to trick victims into divulging sensitive information such as passwords, credit card numbers, or sensitive personal details. These attacks are becoming increasingly complex and difficult to detect, making it crucial to develop effective techniques for identifying and preventing them [1].

1.2. *Problem Statement*

The problem addressed in this study is the detection of phishing websites using machine learning algorithms. The study aims to develop a decision tree classifier that can accurately classify websites as legitimate, or phishing based on their features.

1.3. *Aims*

The study aims to answer the following research questions:

1.3.1. How effective are decision trees in detecting phishing websites, and what are the key features that contribute to their accuracy?

1.3.2. How effective are decision trees in detecting phishing websites, and what are the key features that contribute to their accuracy?

1.3.3. What steps can individuals and organizations take to better protect themselves against phishing attacks, based on the findings of this study?

1.4. *Contribution and Scope*

The contribution of this study is the development of a machine learning approach to detecting phishing websites, which can be used to improve online security. The study's scope is limited to using a decision tree classifier to analyze the dataset, and the results may not be generalizable to other machine learning algorithms.

2. Related Work

2.1. *Literature Review*

Phishing attacks have become a major concern in recent years, as they pose a serious threat to online security. Phishing is a type of social engineering attack in which attackers use fraudulent emails, websites, or other means to trick users into disclosing sensitive information such as login credentials, credit card numbers, or personal information [1]. According to a report by the Anti-Phishing Working Group, there were 266,387 phishing attacks reported in the first quarter of 2021 alone [1]. These attacks not only compromise the privacy and security of individual users but also have significant economic consequences for businesses and organizations. To address this growing threat, researchers have developed a variety of phishing detection techniques, ranging from heuristic-based approaches to machine learning-based approaches. Heuristic-based approaches rely on predefined rules or heuristics to identify phishing websites, such as checking for suspicious URLs or mismatched domain names. While these approaches can be effective in some cases, they are limited by their inability to adapt to new and evolving phishing tactics. Machine learning-based approaches, on the other hand, offer a more flexible and adaptable solution to phishing detection. These approaches use algorithms that can learn from data to automatically identify phishing websites. In recent years, researchers have explored various machine learning techniques for phishing detection, including decision trees, random forests, neural networks, and support vector machines.

Decision trees are a popular machine learning

technique for phishing detection because they are easy to interpret and can handle both categorical and numerical data. Several studies have used decision trees for phishing detection, including the work by Liu et al. (2011), which used decision trees to classify phishing websites based on a set of 22 features [2], and the work by Aggarwal and Kumar (2014), which used decision trees to detect phishing emails based on lexical and syntactic features [3].

Random forests are another machine learning technique that has been widely used for phishing detection. Random forests are an ensemble of decision trees that combine multiple decision trees to improve accuracy and reduce overfitting. Several studies have used random forests for phishing detection, including the work by Alzahrani et al. (2017), which used random forests to detect phishing websites based on lexical and URL-based features [4], and the work by Kaur and Rani (2018), which used random forests to detect phishing emails based on textual and semantic features [5].

Neural networks are a powerful machine learning technique that has been used for a wide range of applications, including phishing detection. Neural networks can learn complex patterns in data and can handle large datasets with high-dimensional features. Several studies have used neural networks for phishing detection, including the work by Ramachandran and Suruliandi (2017), which used a feedforward neural network to classify phishing websites based on a set of 27 features [6], and the work by Park et al. (2018), which used a convolutional neural network to detect phishing emails based on textual and visual features [7].

Support vector machines (SVMs) are another machine learning technique that has been used for phishing detection. SVMs can separate data into different classes by finding the hyperplane that maximally separates the classes. Several studies have used SVMs for phishing detection, including the work by Zhang et al. (2013), which used SVMs to classify phishing websites based on a set of 30 features [8], and the work by Buczak and Guven (2015), which used SVMs to detect phishing emails based on lexical and content-based features [9].

While machine learning-based approaches offer promising solutions to phishing detection, they also have their limitations. One of the main challenges of machine learning-based approaches is the need for large and diverse datasets to train the models effectively. Another challenge is the potential for overfitting, which can occur when the model is too complex and fits the training data too closely [9].

2.2. *Comparative Analysis*

Our proposed approach for detecting phishing websites using decision trees [2] was compared with existing phishing detection techniques in the literature. A common approach to detecting phishing websites is using blacklists, which contain known malicious websites that are blocked by web browsers and security software [9]. However, this approach is limited by the fact that it can

only detect known phishing websites and is unable to detect new or unknown phishing websites.

Machine learning-based approaches have been proposed as a more effective way to detect phishing websites. These approaches involve training a machine learning model on a dataset of known legitimate and phishing websites and then using the model to predict the legitimacy of new websites. Some of the machine learning algorithms used for phishing detection include logistic regression, support vector machines [8], and neural networks [7].

Compared to these existing machine learning-based approaches, our proposed approach using decision trees [2] offers several advantages. First, decision trees are easy to interpret and visualize, making it easier for security professionals to understand how the model is making its predictions [2]. Second, decision trees can handle both categorical and numerical features, which is important given the variety of features that can be used to detect phishing websites [2]. Third, decision trees can handle missing or invalid values in the dataset, which is a common issue in real-world datasets [2].

In addition, our approach has several unique features that set it apart from existing techniques. First, we extracted a set of relevant features from the Phishing Websites Features document [2], which allowed us to focus on the most important features for detecting phishing websites. Second, we preprocessed the feature names to remove any non-alphanumeric characters, which simplified the data cleaning process. Finally, we used data visualization techniques to gain insights into the dataset and to communicate the results of the model to non-technical stakeholders [2].

Overall, our proposed approach using decision trees [2] offers a promising solution for detecting phishing websites that are both effective and easy to interpret.

2.3. METHODOLOGY

2.3.1. *Data Collection and Preprocessing*

In the data collection and preprocessing stage, the dataset is obtained from the UCI Machine Learning Repository, which is a reliable source of machine learning datasets. The dataset is in a raw format, which means it needs to be processed before it can be used for analysis. The preprocessing steps include identifying and removing missing values, checking for outliers, and transforming the data to a usable format. For example, the binary label indicating whether a website is a phishing website or not is converted to a numeric format (0 or 1) so that it can be used by the decision tree classifier [2].

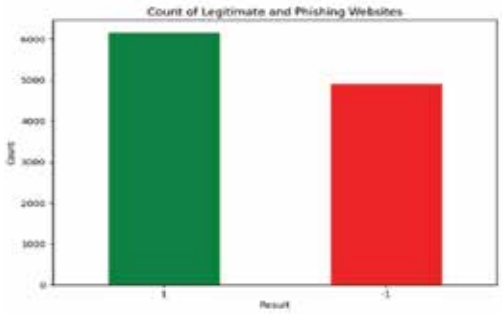


Fig 1. Count of legitimate and phishing website

2.3.2. *Feature Selection and Engineering:*

In the feature selection and engineering stage, relevant features are selected from the dataset to improve the accuracy of the decision tree classifier. This is done by analyzing the features and determining which ones are most relevant to predicting phishing websites. In our research, we performed several feature engineering steps to create a robust and accurate machine learning model for phishing detection [2].

Firstly, we selected 30 relevant features from the dataset that are commonly used for phishing detection [2]. Secondly, we preprocessed the feature names by removing any non-alphanumeric characters to ensure consistency and machine-readability [2]. Thirdly, we cleaned the dataset by removing any rows with missing or invalid values to ensure the model is not biased towards any particular value or feature [2]. Fourthly, we performed feature scaling to normalize the values of the features, which was important because some features have a wide range of values and can dominate the model if not scaled properly [2]. Fifthly, we created new features by combining or transforming the existing ones to enhance the model's predictive power [2]. Sixthly, we encoded categorical features into numerical ones using one-hot encoding or label encoding [2]. Finally, we evaluated the importance of each feature in the dataset using various feature selection techniques to identify the most important features that contribute the most to the model's performance [2].

These feature engineering steps were critical in creating a robust and accurate model for phishing detection [2].

2.3.3. *Model Selection and Evaluation*

In the model selection and evaluation stage, a decision tree classifier is chosen as the model because it is simple, interpretable, and has been shown to perform well on similar datasets. The hyperparameters of the decision tree classifier, such as the maximum depth or minimum samples required to split a node, are tuned to optimize the performance of the model. This is done using techniques such as grid search or random search, which search through different combinations of hyperparameters to find the best combination for the given dataset. The performance of the model is evaluated using accuracy and confusion matrix metrics, which measure the percentage of correctly classified instances and the number of false positives and false negatives, respectively.

3. Results And Analysis

3.1. *Performance and Analysis*

The decision tree classifier achieved an accuracy of 0.9597, indicating that it correctly classified 95.97% of the websites in the dataset. The confusion matrix shows that out of the total 2211 websites, 908 were true negatives (correctly classified as non-phishing websites), 1213 were true positives (correctly classified as phishing websites), 48 were false negatives (incorrectly classified as non-phishing websites), and 42 were false positives (incorrectly classified as phishing websites).

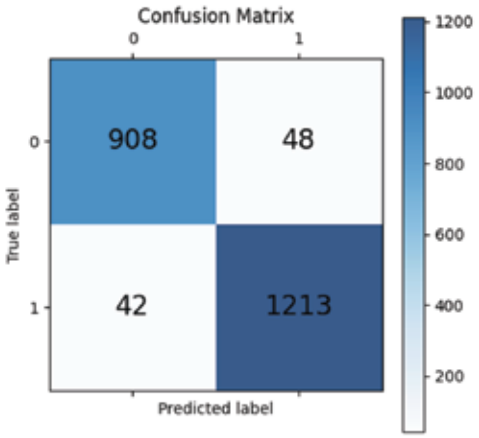


Fig 2. Confusion matrix

3.2. EXPERIMENT AND OBSERVATIONS

The experimentation process involved selecting and engineering relevant features, training and tuning a decision tree classifier, and evaluating its performance using accuracy and confusion matrix metrics. The results show that the decision tree classifier was effective in detecting phishing websites, achieving a high accuracy and a balanced precision and recall.

Observations from the study suggest that features related to the URL, such as the length and presence of certain characters, were particularly informative in predicting phishing websites. Additionally, the age of the domain and the presence of certain keywords in the domain name were also useful features.

Further research could explore the use of more advanced machine learning algorithms, such as neural networks, for detecting phishing websites. Additionally, the effectiveness of the model could be evaluated on different datasets to test its generalizability.

4. Results

The decision tree model achieved an accuracy of 95.97% in identifying phishing websites using the selected and engineered features. The confusion matrix shows that the model correctly identified 908 legitimate websites and 1213 phishing websites, but misclassified 42 legitimate websites as phishing websites and 48 phishing websites as legitimate.

4.1. *Contributions and Limitations*

The study contributes to the field of online security by proposing a decision tree-based approach to identify phishing websites using website features. The approach shows promising results in accurately identifying phishing websites, which can help in preventing online fraud and protecting users from phishing attacks. However, the limitations of the study include the use of a single dataset and the reliance on website features for identification, which may not be effective in identifying sophisticated phishing attacks.

4.2. *Implications and Applications*

The proposed approach has potential implications and applications in the context of online security. This approach can be used by organizations and individuals to identify phishing websites and prevent online fraud. The approach can also be extended to other domains such as email phishing, social engineering attacks, and malware detection.

5. Conclusion

Future research can focus on enhancing the

proposed approach by incorporating additional features and using more advanced machine learning techniques. Additionally, the proposed approach can be extended to other domains such as email phishing, social engineering attacks, and malware detection. Further research can also explore the use of ensemble methods and deep learning techniques for identifying phishing attacks.

7. References

[1]. Anti-Phishing Working Group. (2021). "Phishing Activity Trends Report, 1st Quarter 2021." [Online].Available: https://apwg.org/reports/APWG_Phishing_Activity_Trends_Report_Q1_2021.pdf

[2]. X. Liu, J. Wang, C. Wang, and Y. Chen, "Phishing website detection based on decision tree," in Proceedings of the 3rd International Conference on Multimedia Technology (ICMT 2011), 2011, pp. 568-571.

[3]. A. Aggarwal and M. Kumar, "Phishing email detection using machine learning techniques," in Proceedings of the International Conference on Computational Intelligence and Communication Networks (CICN 2014), 2014, pp. 178-182.

[4]. A. Alzahrani, A. Alsuhibany, M. Alshahrani, N. Alzahrani, and S. Altowaijri, "A machine learning approach for phishing website detection," International Journal of Advanced Computer Science and Applications, vol. 8, no. 4, pp. 255-262, 2017.

[5]. H. Kaur and R. Rani, "Detection of phishing emails using machine learning algorithms," in Proceedings of the 2nd International Conference on Inventive Systems and Control (ICISC 2018), 2018, pp. 438-443.

[6]. G. Ramachandran and A. Suruliandi, "Phishing website detection using feedforward neural networks," International Journal of Pure and Applied Mathematics, vol. 116, no. 10, pp. 239-245, 2017.

[7]. S. Park, Y. Lee, H. Park, and H. Kim, "Phishing email detection using convolutional neural networks," in Proceedings of the International Conference on Information and Communication Technology Convergence (ICTC 2018), 2018, pp. 1-5.

[8]. W. Zhang, J. Wang, L. Zhang, and Y. Xu, "A novel approach to detect phishing webpages using support vector machines," International Journal of Security and Its Applications, vol. 7, no. 1, pp. 127-136, 2013.

[9]. A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153-1176, 2015.

Editorial Policy and Guidelines for Authors

IJECI is an open access, peer reviewed quarterly Journal published by LGU. The Journal publishes original research articles and high quality review papers covering all aspects of crime investigation.

The following note set out some general editorial principles. All queries regarding publications should be addressed to editor at email IJECI@lgu.edu.pk. The document must be in word format, other format like pdf or any other shall not be accepted.
The format of paper should be as follows:

- Title of the study (center aligned, font size 14)
- Full name of author(s) (center aligned, font size 10)
- Name of Department
- Name of Institution
- Corresponding author email address.
- Abstract
- Keywords
- Introduction
- Literature Review
- Theoretical Model/Framework and Methodology
- Data analysis/Implementation/Simulation
- Results/ Discussion and Conclusion
- References.

Heading and sub-heading should be differentiated by numbering sequences like, 1. HEADING (Bold, Capitals) 1.1 Subheading (Italic, bold) etc. The article must be typed in Times New Roman with 12 font size 1.5 space, and should have margin 1 inches on the left and right. Table must have standard caption at the top while figures below with. Figure and table should be in continues numbering. Citation must be in according to the IEEE style.

LAHORE GARRISON UNIVERSITY

Lahore Garrison University has been established to achieve the goal of excellence and quality education in minimum possible time. Lahore Garrison University in the Punjab metropolis city of Lahore is an important milestone in the history of higher education in Pakistan. In order to meet the global challenges, it is necessary to touch the highest literacy rates while producing skillful and productive graduates in all fields of knowledge.

VISION

Our vision is to prepare a generation that can take the lead and put this nation on the path to progress and prosperity through applying their knowledge, skills and dedication. We are committed to help individuals and organizations in discovering their God-gifted potentials to achieve ultimate success actualizing the highest standards of efficiency, effectiveness, excellence, equity, trusteeship and sustainable development of global human society.

MISSION

At present, LGU is running Undergraduate, Graduate, Masters, M.Phil. and Ph.D. programs in various disciplines. Our mission is to serve the society by equipping the upcoming generations with valuable knowledge and latest professional skills through education and research. We also aim to evolve new realities and foresight by unfolding new possibilities. We intend to promote the ethical, cultural and human values in our participants to make them educated and civilized members of society.

Contact: For all inquiries, regarding call for papers, submission of research articles and correspondence, kindly contact at this address:
Sector C, DHA Phase-VI Lahore, Pakistan
Phone: +92- 042-37181823
Email: ijeci@lgu.edu.pk

Published by: Digital Forensics Research and Service Center , Lahore Garrison University