big data and cognitive computing

Special Issue Reprint

Managing Cybersecurity Threats and Increasing Organizational Resilience

Edited by
Peter R.J. Trim and Yang-Im Lee

mdpi.com/journal/BDCC

MDPI

Managing Cybersecurity Threats and Increasing Organizational Resilience

Managing Cybersecurity Threats and Increasing Organizational Resilience

Editors

Peter R.J. Trim

Yang-Im Lee

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Editors

Peter R.J. Trim
Birkbeck Business School
Birkbeck, University of London
London, UK

Yang-Im Lee
Westminster Business School
University of Westminster
London, UK

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Big Data and Cognitive Computing* (ISSN 2504-2289) (available at: https://www.mdpi.com/journal/BDCC/special_issues/Cybersecurity_2nd).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-0365-9644-0 (Hbk)
ISBN 978-3-0365-9645-7 (PDF)
doi.org/10.3390/books978-3-0365-9645-7

Contents

About the Editors

Peter R.J. Trim

Peter R.J. Trim is a Reader in Marketing and Security Management at Birkbeck, University of London, and holds degrees from various institutions including City University (City, University of London), Cranfield Institute of Technology (Cranfield University), and the University of Cambridge. He is a Fellow of the Higher Education Academy and the Royal Society of Arts. He has published 68 academic journal articles and 12 books and is the co-editor of two government reports, the editor of a journal special section, and the co-editor of three journal special issues. In addition, he has authored 65 chapters in books and given a large number of conference presentations. Peter has been involved in a number of funded research projects and has worked in several industries and overseas. He is the co-author with Yang-Im Lee of a book entitled: Strategic Cyber Security Management, Routledge, which draws on a social science perspective and links cyber security management with resilience and business continuity planning and other subject areas. Currently, Peter is involved in various aspects of research including cybersecurity management, social inclusion, social impact, and online marketing. Peter is actively involved in a number of cybersecurity initiatives, including conferences and workshops. He has also undertaken a number of projects involving government, academia, and industry.

Yang-Im Lee

Yang-Im has studied and worked in Korea, Japan, and the UK. She undertook postgraduate studies at the School of Oriental and African Studies, University of London, and was awarded a scholarship by Stirling University to undertake a Ph.D. at the institution. Yang-Im is currently a senior lecturer in Marketing at Westminster Business School, University of Westminster, where she teaches various aspects of marketing. Yang-Im has published over 30 articles in a range of academic journals, has co-authored books, and presented a number of conference papers. She is also the co-editor of three journal special issues. Yang-Im is a Fellow of the Royal Society of Arts and has been a visiting fellow at Birkbeck, University of London. Yang-Im has a deep interest in education and the use of technology and in the past provided support for the Information Assurance Advisory Council, where she worked as their academic liaison panel co-ordinator for a number of years. Yang-Im has been involved in a number of funded research projects in the UK and is currently undertaking research into online marketing, social inclusion, and cybersecurity management.

Preface

In order to effectively deal with the range of cybersecurity issues and challenges confronting society, it is essential that university, industry, and government researchers pool their knowledge and expertise and work on joint research projects. Should this be the case, a holistic view of the problem will be established, and it will be possible to implement solutions that thwart the actions of those who carry out cyber-attacks. The papers included in this reprint are testimony of the collective action of researchers and proof that knowledge can be developed and shared within the wider community. We are grateful, therefore, and thank the contributors for their commitment to this Special Issue, which will make a considerable contribution to raising the profile of cybersecurity.

Peter R.J. Trim and Yang-Im Lee
Editors

big data and cognitive computing

MDPI

Editorial

Managing Cybersecurity Threats and Increasing Organizational Resilience

Peter R. J. Trim [1],* and Yang-Im Lee [2],*

[1] Birkbeck Business School, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK
[2] Westminster Business School, University of Westminster, 35 Marylebone Road, London NW1 5LS, UK
* Correspondence: p.trim@bbk.ac.uk (P.R.J.T.); y.lee@westminster.ac.uk (Y.-I.L.)

Cyber security is high up on the agenda of senior managers in private and public sector organizations and is likely to remain so for the foreseeable future. Because cyber-attacks are increasing in sophistication and are of a persistent nature, it is clear that those undertaking research into counteracting cyber threats should familiarize themselves with the types of vulnerability that are likely to be exploited and develop workable solutions. This means working with likeminded people that are intent on ensuring that those carrying out such attacks do not succeed. It is because of the complexity and width of the problem that it is unlikely that those working in a single discipline will be able to solve the recurring problems that managers face. Indeed, the nature of connectivity and interactivity requires that cyber security researchers adopt an inter-disciplinary and/or multi-disciplinary approach to solving cyber security problems, and also that academic and industry researchers cooperate in order to work on cyber security solutions that can be applied across all industry sectors.

This Special Issue draws on the knowledge of various cyber security experts from a range of disciplines who address a number of issues and put forward solutions that utilize cyber security intelligence, with the aim of making organizations more resilient and able to withstand different types of cyber-attack. This means that studying the problem from various perspectives and establishing the breadth and depth of the problem are key priorities. The collection of papers in this Special Issue will help broaden the scope of the subject matter and through interpretation will offer recommendations for dealing with known cyber threats.

This volume of papers complements the existing literature and places cybersecurity within a wider context so that various concepts, models, and strategies can be applied to solving cybersecurity threats as and when they occur. Indeed, in [1] the vulnerability of individuals is made clear, and this is due to the increasing use of social media platforms and the increase in electronic risks that have allowed cybercrime to thrive. To counteract phishing and various other forms of cyber-attack, attention to data privacy is essential. This means that cyber security awareness is given priority and ways are found to reduce individuals' vulnerabilities. In [2], reference is made to ransomware attacks, which result in monetary losses and reputational damage, and it is for these reasons that current trends need to be monitored and ransomware detection deployed. By having an overview of ransomware attacks, intelligence can be established that identifies and minimizes the actions of those carrying out such attacks.

As regards security threats and requirements, Ref. [3] advocates a data processing approach that outlines the steps incorporated within a centralized contact tracing system that can prove beneficial in terms of collecting and sharing information relating to an event/outcome. By mapping security and privacy threats, the security requirements for each type of threat can be made known and the centralized contact tracing system can be viewed as effective. Acknowledging that decision support tools play a useful role vis à vis intelligent sociotechnical systems [4], a number of challenges can be overcome. The emphasis is on the ability to analyze and process various forms of information. Defeasible

Citation: Trim, P.R.J.; Lee, Y.-I. Managing Cybersecurity Threats and Increasing Organizational Resilience. *Big Data Cogn. Comput.* **2023**, *7*, 177. https://doi.org/10.3390/bdcc7040177

Received: 15 November 2023
Accepted: 16 November 2023
Published: 22 November 2023

logic programming (DeLP) can be utilized, a P-DAQAP framework can be developed, and a preliminary empirical evaluation undertaken.

Reflecting on the fact that security controls help to safeguard software [5], it is essential to find novel alternatives such as Security Chaos Engineering (SCE) that can be used to protect assets. A defensive security strategy can harness ChaosXploit, which will help identify and correct software misconfigurations sooner rather than later. Accepting that 5G communications systems are vulnerable [6], it can be argued that it is necessary to establish and measure the primary indicators in relation to the effectiveness of a security system, devise a list of cybersecurity KPIs, and model matters accordingly. Additionally, critical infrastructure can be better protected through the deployment of Threat Hunters that are able to detect anomalies [7]. Artificial intelligence (e.g., machine learning) and visualization techniques can enhance Cyber Situational Awareness (CSA) and manifest in the protection of critical infrastructure.

A particle swarm optimization (PSO)-driven selection approach to identify the optimum feature subsets and hybrid ensemble can help to enhance anomaly-based intrusion detection systems [8]. Research [9] undertaken into deep learning to detect and protect against botnet threats in relation to flying ad hoc networks (FANETs) utilizes the hybrid shark and bear smell optimization algorithm (HSBSOA). The outcome is the hybrid shark and bear smell-optimized dilated convolutional autoencoder (HSBSOpt_DCA).

In [10], a solution is provided for the transfer between blockchain-based heterogeneous cryptocurrencies and central bank digital currencies (CBDCs). The researchers focus on and draw from existing interoperability studies and solutions. An interoperable architecture involving heterogeneous blockchains is used, and a decentralized peer-to-peer (P2P) service model is proposed. In addition, security threats to the proposed service model are made known and, most importantly, security requirements to counteract security threats are detailed. In [11], attention is given to how managers can better appreciate the role that sociocultural intelligence plays and utilize artificial intelligence more to facilitate cyber threat intelligence (CTI). The intelligence cycle (IC) and the critical thinking process (CTP) are described and combined, and a cyber threat intelligence cycle process (CTICP) is developed that aids the resilience-building process.

Reflecting on the above set of papers, it can be argued that much has been achieved as regards counteracting the actions of those intent on carrying out cyber-attacks, but there is still a lot more work to be done. Clearly, the benefits of adequate cyber security provision are clear to see, and the holistic picture derived from this Special Issue will help to identify new areas of research and foster continued cooperation among cybersecurity researchers. This is important for strengthening the academic base of the subject and encouraging researchers from academia, industry, and government to pool resources and find novel solutions to current and emerging forms of cyber-attack.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alrobaian, S.; Alshahrani, S.; Almaieh, A. Cybersecurity awareness assessment among trainees of the technical and vocational training corporation. *Big Data Cogn. Comput.* **2023**, *7*, 73. [CrossRef]
2. Alraizza, A.; Algarni, A. Ransomware detection using machine learning: A survey. *Big Data Cogn. Comput.* **2023**, *7*, 143. [CrossRef]
3. Park, S.; Youm, H.-Y. Security and privacy threats and requirements for the centralized contact tracing systems in Korea. *Big Data Cogn. Comput.* **2022**, *6*, 143. [CrossRef]
4. Leiva, M.A.; García, A.J.; Shakarian, P.; Simari, G.I. Argumentation-based query answering under uncertainty with application to cybersecurity. *Big Data Cogn. Comput.* **2022**, *6*, 91. [CrossRef]
5. Chavarro, S.P.; Nespoli, P.; Díaz-López, D.; Roa, Y.N. On the way to automatic exploitation of vulnerabilities and validation of systems security through security chaos engineering. *Big Data Cogn. Comput.* **2023**, *7*, 1.
6. Odarchenko, R.; Iavich, M.; Iashvili, G.; Fedushko, S.; Syerov, Y. Assessment of security KPIs for 5G network slices for special groups of subscribers. *Big Data Cogn. Comput.* **2023**, *7*, 169. [CrossRef]
7. Lozano, M.A.; Llopis, I.P.; Domingo, M.E. Threat hunting architecture using a machine learning approach for critical infrastructure protection. *Big Data Cogn. Comput.* **2023**, *7*, 65. [CrossRef]

8. Louk, M.H.L.; Tama, B.A. PSO-driven feature selection and hybrid ensemble for network anomaly detection. *Big Data Cogn. Comput.* **2022**, *6*, 137. [CrossRef]
9. Abdulsattar, N.F.; Abedi, F.; Ghanimi, H.M.A.; Kumar, S.; Abbas, A.H.; Abosinnee, A.S.; Alkhayyat, A.; Hassan, M.H.; Abbas, F.H. Botnet detection employing a dilated convolutional autoencoder classifier with the aid of hybrid shark and bear smell optimization algorithm-based feature selection in FANETs. *Big Data Cogn. Comput.* **2022**, *6*, 112. [CrossRef]
10. Park, K.; Youm, H.-Y. Proposal of decentralized P2P service model for transfer between blockchain-based heterogeneous cryptocurrencies and CBDCs. *Big Data Cogn. Comput.* **2022**, *6*, 159. [CrossRef]
11. Trim, P.R.J.; Lee, Y.-I. Combing sociocultural intelligence with artificial intelligence to increase organizational cyber security provision through enhanced resilience. *Big Data Cogn. Comput.* **2022**, *6*, 110. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

big data and
cognitive computing

MDPI

Article

Cybersecurity Awareness Assessment among Trainees of the Technical and Vocational Training Corporation

Shouq Alrobaian [1], Saif Alshahrani [2] and Abdulaziz Almaleh [3,*]

[1] Technology and Information Security Department, Jazan University, Jazan 82817, Saudi Arabia
[2] Technical and Vocational Training Corporation, Bisha College, Bisha 67714, Saudi Arabia
[3] Information Systems Department, King Khalid University, Abha 62529, Saudi Arabia
* Correspondence: ajoyrulah@kku.edu.sa; Tel.: +966-533-212-174

Abstract: People are the weakest link in the cybersecurity chain when viewed in the context of technological advancement. People become vulnerable to trickery through contemporary technical developments such as social media platforms. Information accessibility and flow have increased rapidly and effectively; however, due to this increase, new electronic risks, or so-called cybercrime, such as phishing, scams, and hacking, lead to privacy breaches and hardware sabotage. Therefore, ensuring data privacy is vital, particularly in an educational institute where students constitute the large majority of users. Students or trainees violate cybersecurity policies due to their lack of awareness about the cybersecurity environment and the consequences of cybercrime. This paper aims to assess the level of awareness of cybersecurity, users' activities, and user responses to cybersecurity issues. This paper collected data based on a distributed questionnaire among trainees in the Technical and Vocational Training Corporation (TVTC) to demonstrate the necessity of increasing user awareness and training. In this study, quantitative research techniques were utilized to analyze the responses from trainees using tests such as the Chi-Squared test. Proof of the reliability of the survey was provided using Cronbach's alpha test. This research identifies the deficiencies in cybersecurity awareness among TVTC trainees. After analyzing the gathered data, recommendations for tackling these shortcomings were offered, with the aim of enhancing trainees' decision-making skills regarding privacy and security using the Nudge model.

Keywords: cybersecurity; awareness; survey; information security; cybercrime

Citation: Alrobaian, S.; Alshahrani, S.; Almaleh, A. Cybersecurity Awareness Assessment among Trainees of the Technical and Vocational Training Corporation . *Big Data Cogn. Comput.* **2023**, *7*, 73. https://doi.org/10.3390/bdcc7020073

Academic Editors: Peter R.J. Trim and Yang-Im Lee

Received: 24 February 2023
Revised: 8 April 2023
Accepted: 11 April 2023
Published: 12 April 2023

1. Introduction

The internet has become significantly connected to our lives as our economy and infrastructure have become heavily dependent on internet networks and modern technology [1]. The use of the internet has spread, especially with the digital transformation that depends on managing operations for the public and private sectors by integrating modern technology and taking advantage of it in all aspects of life and social circles [2]. The digital transformation has caused a vast revolution, especially among educational circles, mainly through the use of technology to obtain and disseminate information, which has led to an increase in the use of the internet [3].

The ease of sharing and finding personal information via social media or online searches has increased, but without adequate cybersecurity awareness, users may encounter challenges in determining whether to disclose their data. Factors such as cognitive biases, time limitations, and emotional influences can complicate the selection process of appropriate privacy protection options. This is especially true when interacting with user interfaces on websites that necessitate registration or involvement [4].

Therefore, users will not have complete control over the privacy of their data, which may lead to its violation [5]. Conversely, internet usage may involve certain processes or elements that necessitate user consent, often without them being fully aware that some of these aspects could be detrimental to their personal data [6]. Therefore, there are so-called "Service Terms" that are included in every service provided to the user, whether in the social or educational aspect, and they explain to the user how to benefit from and control their data when using this service. It is often ignored and unread by the user, usually due to a lack of awareness on behalf of the user in protecting their data [7].

As a result of the increased use of the internet, cybercrime and electronic fraud cases have increased. Cybercrimes are similar traditional crimes in terms of different aspects and groups, but their development is related to computer use and geographical diversity [8]. They are carried out by programmers called hackers, and they are divided according to their actions, which may be on a personal level, i.e., for personal benefit by causing harm to others, or for the general use, for example, for testing systems or trying something to help [9]. Hackers have developed new methods and techniques that may lead to financial gain and psychological harm, or they may just sabotage for fun [3]. These cyber attacks are cheaper and less dangerous than physical attacks, in addition to some other advantages, such as the irrelevance of the distance to or place of an attack and the difficultly in identifying and prosecuting the attacker. Accordingly, cyber attacks may continue to increase [1], which may lead to violations of cybersecurity systems that protect the automation of the economy and infrastructure [3].

Cybersecurity includes the process of providing protection for cyberspace and organizing all resources and processes related to cyber attacks [10]. The primary cause leading to the increase in cyber attacks is the failure to follow the cybersecurity guidelines offered by organizations. In [3], the authors stress the critical nature of implementing and adhering to cybersecurity guidelines across all divisions of an organization. They highlight the need to focus on the organization's members, representing the most vulnerable point in the security chain. This underscores the significance of cultivating strong cybersecurity practices among employees to bolster overall organizational security. The authors of [4] also emphasize the value of gently motivating users to make optimal choices regarding the sharing of their personal data in the context of online privacy and security. By utilizing non-intrusive interventions, individuals can be guided toward making better-informed decisions about their data protection and online safety.

As cyber attacks have increased around the world, cybersecurity has become a priority in many countries. Accordingly, the Kingdom of Saudi Arabia has strengthened its investments and efforts to develop cybersecurity and its related procedures in the public and private sectors by 2030. The Kingdom of Saudi Arabia has established the National Authority for Cybersecurity (NCA) [11] to strengthen the position of cybersecurity and control the procedures and operational processes associated with it. The Saudi Federation for Cybersecurity and Drones (SAFCSP) [12] is another Saudi association that applies international standards, regulations, and practices to help improve the cybersecurity of the Kingdom of Saudi Arabia for it to become one of the leading countries in the technology revolution [13].

The rapid development of technology has led to an increase in the use of intelligent devices connected to the internet, especially in the educational sector. The number of smart devices exceeded 4 billion in 2020, leading to increased cyber attacks and new challenges [14].

The main reason for the increase in cybercrime in the educational sector is the poor awareness among users, as experts have shown in [15]. Cybersecurity awareness and policies in Saudi Arabia have not received enough attention among university students and institute trainees. This entails protecting individuals and university and school students by raising awareness about cybersecurity, providing training programs and educational means on the challenges of cybersecurity and the consequences of information crimes, and increasing the knowledge of risks of losing sensitive information [3,15]. This work assesses the level of awareness of cybersecurity and users' activities and their reaction to cybersecurity aspects. The contribution of this paper is as follows:

- The level of cybersecurity awareness is explored among trainees at the TVTC by evaluating and measuring many security factors while using the internet.
- Gaps are found in awareness of trainees at the educational organization (TVTC) after examining and analyzing the results and strategies are proposed to enhance this awareness.
- Awareness about cybersecurity is enhanced by presenting a theoretical framework appropriate for the TVTC to educate trainees about the risks and consequences of cybercrime.
- The approach is developed in the TVTC and proposals are made commensurate with the gaps we found through analyses of the results to improve the security environment and the decision-making process of individuals and the organization.

The rest of this paper is structured as follows: the relevant works are put forward in Section 2. Section 3 presents the methodology for assessing cybersecurity awareness among trainees and describes the dataset collected in this study. The results are shown in Section 4 based on the analysis and examination of the data. This paper concludes with a review of the study's data and findings in Section 5, followed by Section 6 with a conclusion.

2. Related Work

Few studies have covered cybersecurity awareness in the educational community and among students, which depends on people's understanding and knowledge of cybersecurity or information security and the consequent risks and methods of protection from them [16]. Many relevant works have determined the awareness level by assessing the understanding of cybersecurity concepts among students. Alharbi et al. [3] showed how Majmaah University students [17] understand cybersecurity, cyberattacks, and their consequences. Based on the questionnaire conducted by researchers, they found that awareness about cybersecurity must be increased among university students, advanced educational methods should be used and combined with traditional methods, and videos and games can be used to provide awareness to students. However, the length of the questions in the questionnaire was one of the defects of this study, which may have led to ambiguity in understanding basic terms and concepts.

Khader et al. [18] suggested a theoretical cybersecurity awareness framework that directs the implementation of programs to raise graduates' cybersecurity awareness in any academic setting. The CAFA [19] can be a jumping-off point for educational institutions looking to establish new policies and procedures.

The study in [20] aimed to determine the level of understanding of threats related to online security and comprehension of the preventive measures used to protect young people from online dangers. Data were collected from youths enrolled in classes of children aged eleven and higher at random. According to the survey findings, most young people are unaware of internet security risks and hazards. This survey sample did not adhere to the universal frameworks used to produce acceptable results, which can be enhanced to reflect solid findings [21].

Another work performed by Taha et al. [22] compared college students' knowledge and behavior regarding information security awareness. The main objective was to compare students' understanding of information security when using smartphones versus computers to see where there are differences. As a result of their work, they encourage academic institutions to exercise caution and run information security awareness campaigns. The creation of the necessary level of awareness among all Jordanian students would be facilitated by including an information security course as a university requirement. However, the survey question count needed to be improved, which resulted in inaccurate measurements of all relevant factors considering cyber attack evolution and the tools available to defend against them.

The authors of [23] assessed students' cybersecurity knowledge in developing countries by examining the understanding of the effects of software and email security. The study was conducted through a scientific questionnaire containing eleven questions, which could be considered as of the defects of this study, as the number of questions needed to be increased to include all essential aspects of cybersecurity. However, through this questionnaire, the researchers found that awareness of email security increases awareness about cybersecurity more than software security.

Likewise, researchers [24] investigated the increasing awareness of cybersecurity with the spread of social engineering attacks targeting users as they are the weakest link according to their level of understanding about this type of attack. The researchers discovered that education programs are an effective method to raise awareness among users and employees. Nevertheless, the work could have included the study and comparison of laws and regulations legislated by governments.

3. Materials and Methods

3.1. Research Method

This study used a survey method to gather qualitative data about the Technical and Vocational Training Corporation trainees and assess their level of cyber security knowledge. The survey was conducted online to efficiently and ethically collect a sizable sample of male and female trainees. There were 40 questions in total, covering a variety of cybersecurity topics, such as demographics (4 questions), technical information (2 questions), internet usage (2 questions), information about prior hacks (1 question), use of security tools such as antivirus [25], two-factor authentications (2FA) [26], firewalls [27] (9 questions), password policy (9 questions), browser security (3 questions), social networking (5 questions), and cybersecurity knowledge (9 questions). The survey questions were chosen based on mechanisms designed by other cybersecurity researchers [3,23].

The internet serves as a worldwide platform for information and commerce, offering numerous benefits to users. However, as individuals spend more time online, they may encounter various infringements, including privacy concerns that necessitate increased awareness of responsible internet usage [28]. To better understand this phenomenon, questions were designed to gather insights into the online behavior of trainees, ultimately shedding light on their internet usage habits and potential vulnerabilities.

Awareness questions about security tools, which in turn help individuals to protect themselves from cybercrime-related threats during personal use of the internet, noting that it is not enough to rely on them alone [29], have been created to examine the current security practices among Technical and Vocational Training Corporation trainees.

The browser security segment questions aim to assess the trainees' comprehension of how secure their standard web browser is. A web browser is the gateway to information and services via the internet, through which accounts are accessed via e-mail, social media, and downloading various files. Hence, it counts as a sensitive gateway to attack and cybercrime [30].

The networking and cyberspace knowledge questions assess the trainees' understanding of the dangers of accessing a variety of social networks, as it is the main basis for communication between individuals and access to various websites, which increases the risk of attacks on their personal data and information accessed through it. The questions also assessed the trainees understanding of how to respond to cybercrime events [31]. Therefore, we examined the trainees' cybersecurity knowledge, abilities, behavior, beliefs, and self-perception.

The questionnaire was selected from other survey questions created by other researchers in [3], with adjustments to reduce the number of questions (which is mentioned as a limitation in [3]) according to a random sample of 50 male and female trainees who recommended reducing the number of questions to maintain some degree of satisfaction.

3.2. Study Model

The survey depends on the scientific questionnaire standards used in related works [3] and [23] with a few modifications in several questions due to limitations in previous works, such as responses of a random sample of trainees. The modified questions were reviewed and analyzed based on the questionnaire standards [32]. The survey questions also include additional scientific explanations for each section to make it easier for non-technical trainees to understand the questions. The first page of the questionnaire also contained the aim of the study, explaining the meaning and some basic information to the user. After obtaining the required approval from the TVTC, the survey was distributed through the questionnaire link among trainees with the help of heads of department. The sample size of this study followed the standard guidelines [21], which resulted in 739 complete responses from TVTC trainees with limited responses to one answer for each sample by requesting signing in to a Google account.

3.3. Data Collection

The data were collected in electronic form by sharing an official link through the organization to give respondents access to the designed question on the Google Form, answer, and submit their responses. The responses were exported to Microsoft Excel after the questionnaire had been administered. The total number of collected responses exported to Excel was 739. The data were cleaned in Excel, and after cleaning, the data were exported and coded in Statistical Package for the Social Sciences (SPSS) for further analysis.

4. Results

The entire population of trainees was selected for this study, and the respondent trainees served as the chosen sample. The study focuses on trainees' knowledge of cybersecurity issues, including phishing attacks, which is based on targeting specific people through their available data or exploiting errors caused by these people through their use of systems [33]; malware, which is programming code that helps perform malicious actions used by attackers to steal information or harm others without user permission [34]; patching, which is intended to fix defects in programs; and adding features, including improving the security of programs by identifying, verifying, and installing updates [35]. The actions of trainees exposed to cybercrime were also studied. The survey also gathered information from trainees regarding cybersecurity concepts such as countermeasures, password protection, website security, and social media platforms.

4.1. Descriptive Analysis

This section focuses on data analysis, which is presented as frequency distribution tables, bar charts, percentages, and proportions using Chi-square test techniques [36]. Tests were conducted at a 95% confidence level, and the decision rule was based on the null hypothesis; if the p-value was less than 0.05 we reject the null hypothesis and conclude that the two groups are dependent on each other, and if the p-value is greater than 0.05, we do not reject the null hypothesis and conclude that the two groups are independent of each other [37].

The accuracy of the assessment of cybersecurity knowledge of trainees depends on measuring the influence of the life cycle variables of the trainees. Therefore, variables such as sex, the level of qualification, specialization, and the operating system used were selected to help the assessment. Table 1 summarizes the variable information of the sampled population in more detail.

Table 1. Shows the respondents' gender, level of qualification, operating system, and specializations.

Variables		Freq.	Percentage %
Sex	Male	281	38.02
	Female	458	61.98
Degree	BA	19	2.57
	Diploma	720	97.43
Specialization	Accounting	4	0.54
	Administrative technology	194	26.25
	Arabic language	1	0.14
	Chemical technology	4	0.54
	Civil and architectural technology	4	0.54
	Computer technology	281	38.02
	Decoration, beauty technology, and clothing design	146	19.76
	Electrical technology	2	0.27
	Electronic technology	53	7.17
	Food technology and the environment	2	0.27
	Human resources	5	0.68
	Insurance	8	1.08
	Library administration	11	1.49
	Mechanical technology	16	2.17
Operating Systems used	Linux	8	1.08
	Mac	123	16.64
	Windows	403	54.53
	Unknown	163	22.06
	Windows system, Linux system (Linux)	13	1.76
	Windows system, Mac system (Mac OS)	24	3.25
	Windows, Mac OS, Linux	5	0.68

As the table shows, most of the respondents were female (458 (61.98%)), while there were 281 male respondents (38.02%). It was recorded that the majority of the respondents, 720 (97.43%), had a diploma, while the rest of the respondents, 19 (2.57%), had bachelor degrees. The specialization area in Table 1 shows that 4 (0.54%) respondents were accounting specialists, 194 (26.25%) respondents belong to administrative technology (either marketing and innovation, human resources, or logistics), one respondent specialized in the Arabic language, 4 (0.54%) respondents specialized in both chemical technology (chemical production and chemical laboratories) and civil and architectural technology (such as surveying, civil construction, and architectural construction). At total of 281 (38.02%) respondents specialized in computer technology (such as networking, software, technical support, and multimedia). A total of 146 (19.76%) respondents specialized in decoration, beauty technology, and clothing design (e.g., cosmetology, hair care, fashion manufacturing, and fashion design). Two (0.27%) respondents specialized in both electrical technology (such as electrical machines, electric power, and renewable energy) and food technology and the environment (e.g., food safety, occupational safety, and health, and environmental protection). A total of 53 (7.17%) respondents specialized in electronic technology (such as electronics and control systems, precision instruments and machines, and medical devices). Five (0.68%) respondents specialized in human resources, 8 (1.08%) respondents specialized in insurance, 11 (1.49%) respondents were library administration specialists, 16 (1.17%) respondents specialized in mechanical technology (such as manufacturing, engines and vehicles, and refrigeration and air conditioning), and lastly, 8 (1.98%) respondents specialized in tourism and hospitality technology (e.g., travel and tourism, hotels, and event management). Regarding the type of operating system on respondents' devices, the majority of the respondents had Windows on their device (403 (54.53%) respondents), followed by 123 (16.64%) respondents who had Mac on their devices, 8 (1.08%) had Linux on their devices, about 163 (22.06%) respondents did not know the type of operating system on their device, 13 (1.76%) respondents had both Windows and Linux on their device, and 24 (3.25%) had both Windows and Mac on their system device. The respondents were not asked about a specific device type due to the various vendors, which is out of the scope of this research. In comparison, 5 (0.68%) respondents had all three types of operating systems on their system devices, as shown in Figure 1.

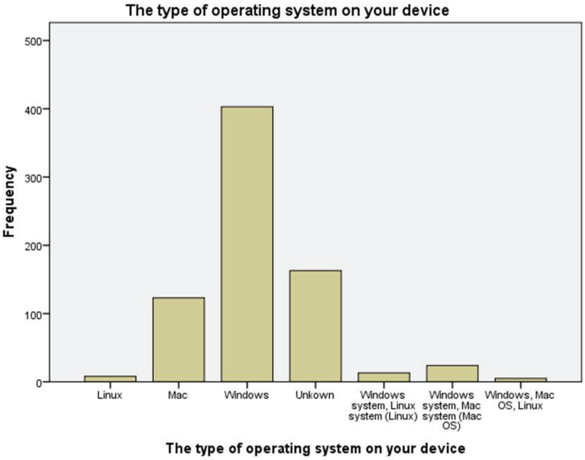


Figure 1. Respondents' Operating Systems.

4.2. Cybersecurity Concepts

In cybersecurity, the term CIA, which indicates confidentiality, integrity, and availability [38], is utilized as the main principle required to maintain the essential knowledge of cybersecurity concepts by applying specific processes to systems and services connected to the internet. Organizations, even academic institutions, protect the cyberspace by protecting weaknesses in the chain (trainees) and should take measures to educate them on how to protect their critical data and networks [38,39]. Based on the weakness in the chain (the trainees), this paper aims to assess the CIA concept among them. The questionnaire in this paper contains 40 questions, of which 26 focus on the cybersecurity aspects of the CIA (Table 2). It includes 14 questions about confidentiality, passwords, and revealing personal information on social networking sites. Twelve integrity, firewall, email policy, browser, and antivirus software questions were included in the evaluation. In addition, all 26 questions were related to measuring availability.

A small percentage of respondents (0.41%) spent the most time on Facebook [40], 27 (3.65%) respondents spent the most time on Instagram [41], 4 (0.54%) respondents spent the most time on LinkedIn [42], and a high percentage of 159 (21.52%) respondents spent the most time on Snapchat [43]. Moreover, 14 respondents spent the most time on both Instagram and Twitter [44], 11 respondents spent the most time on Instagram and YouTube [45], 2 respondents spent the most time on both Snapchat and Facebook, 78 respondents spent the most time on both Snapchat and Instagram, 27 respondents spent the most time on Snapchat and Twitter, 10 respondents spent the most time on Snapchat and YouTube, 3 respondents spent the most time on WhatsApp [46] and Facebook, 13 respondents spent the most time on WhatsApp and Instagram, a high percentage of the respondents (276, 37.35%) spent the most time on WhatsApp and Snapchat, and lastly, four respondents spent the most time on WhatsApp and YouTube.

Table 2. Time respondents spent on social media platforms.

Which Social Network Do You Spend the Most Hours on?	Freq.	Percentage %
Facebook	3	0.41
Instagram	27	3.65
LinkedIn	4	0.54
Snapchat	159	21.52
Twitter	41	5.55
WhatsApp	31	4.19
YouTube	27	3.65
Instagram, Twitter	14	1.89
Instagram, YouTube	11	1.49
Snapchat, Facebook	2	0.27
Snapchat, Instagram	78	10.55
Snapchat, Twitter	27	3.65
Snapchat, Youtube	10	1.35
WhatsApp, Facebook	3	0.41
WhatsApp, Instagram	13	1.76
WhatsApp, Snapchat	276	37.35
WhatsApp, Twitter	9	1.22
WhatsApp, YouTube	4	0.54

About 555 (75.1%) respondents have email and do use their email, while a small amount of 184 (24.9%) respondents sometimes used their email (Table 3).

Table 3. Respondents reply to email usage.

Do You Use E-Mail?	Freq.	Percentage %
Yes	555	75.1
Sometimes	184	24.9

4.2.1. System Update

Table 4 reveals that the majority of the respondents', 392 (53.04%), devices have automatic updates enables, i.e., the device updates the system if it detects a new update, which helps them keep their devices safe. A total of 258 (34.91%) respondents performed manual updates, i.e., the auto update feature is disabled and they update the device themselves when it asks for an update. A total of 59 (7.98%) respondents do not use the update feature, i.e., they use their devices without an update; this makes their devices more vulnerable to threats and hacking than others. A total of 30 (4.06%) respondents had got received device and had not updated it yet. To better understand the percentages, Figure 2 shows the responses regarding the operating system updates.

Table 4. Respondents ways of updating their OS device.

How to Update the Operating System of Your Device?	Freq.	Percentage %
Automatic update (the automatic update feature is enabled and the device updates the system if it detects a new update)	392	53.04
I do not know the update feature	59	7.98
Manual update (the auto update feature is disabled and I update the device myself when it asks for an update)	258	34.91
Never (the device is new)	30	4.06

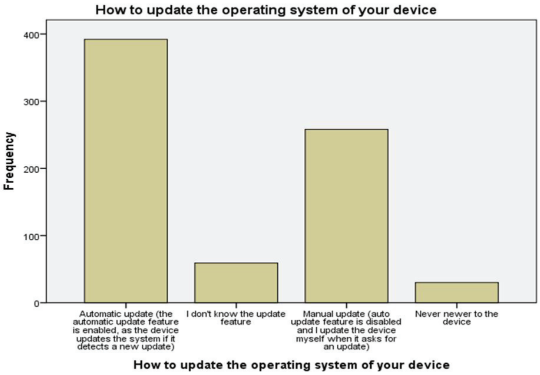


Figure 2. How respondents update their operating system.

4.2.2. Devices Attacked

The following Figure 3 shows the results of whether the trainees' devices had been attacked before. A total of 660 (89.31%) respondents' devices had not been attacked before, which means they apply proper security practices, while a virus had attacked 33 (4.47%) respondents' devices, 31 (4.19%) respondents' accounts had been hacked, and 15 (2.03%) respondents had been scammed.

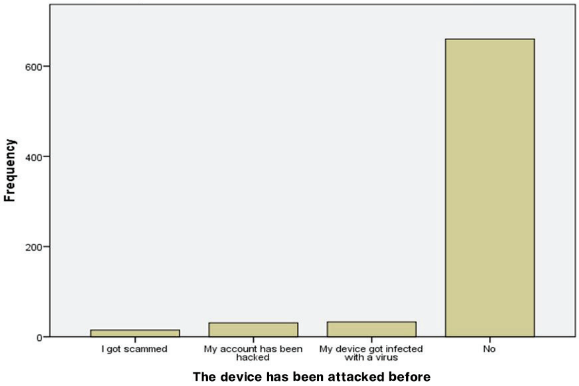


Figure 3. Previously attacked devices.

Although those who implement security measures make up the majority, this survey asked follow-up questions to the respondents whose devices had been hacked and deceived before, as Table 5 shows.

Of the respondents who had been scammed, 3 (0.4%) did nothing and 12 (1.6%) informed the concerned authorities and their card was suspended. Out of the respondents who informed us that their account was hacked, 10 (1.4%) contacted support for the hacked program, 6 (0.8%) did nothing to it, and 6 (0.8%) informed everyone that their account was hacked and contacted the support for the hack program. Eight (1.1%) only told everyone that their account was hacked. However, of respondents that said that their device was infected with a virus, 10 (1.4%) ran a device scan program (programs to detect viruses inside the device), 9 (1.2%) deleted virus-related files, 7 (0.9%) ran a device scan program (programs to detect viruses inside the device) and deleted the files associated with the virus, and 6 (0.8%) went to tech support.

Table 5. Respondents reactions to the device being attacked.

When You Were Scammed?	Freq.	Percentage %
Did not do anything	3	0.4
Informed the concerned authorities, the bank card was suspended	12	1.6
When my account was hacked		
I contacted support for the hacked program.	10	1.4
I did not do anything	6	0.8
I informed everyone that my account was hacked and I contacted support for the hacked program	6	0.8
I told everyone that my account was hacked	8	1.1
I told everyone that my account was hacked, I contacted support for the hacked program, and I did nothing	1	0.1
When my device got infected with a virus		
I did not do anything	1	1
I ran a device scan program (programs to detect viruses inside the device) and I deleted the files associated with the virus	7	0.9
I went to tech support	6	0.8
I ran a device scan program (programs to detect viruses inside the device)	10	1.4
Virus-related files were deleted	9	1.2

4.2.3. Antivirus Software

The default protection on computers enforces some countermeasures related to the security of devices, such as protection mechanisms. One of the protection mechanisms is software that detects malicious websites when visiting or downloading files containing a virus. This software, called antivirus software, detects malicious files, depending on their signature or behaviors and compares the findings with a huge related database. This type of software helps trainees protect their devices [47]. As expected, most trainees did not have antivirus software installed, as shown in Table 6. A total of 273 (36.94%) respondents had antivirus software installed on their devices, 164 (22.19%) respondents sometimes installed antivirus software on their devices, while 302 (40.87%) did not have antivirus software installed.

Table 6. Installation of antivirus software.

Have You Installed Antivirus Software (Protection Software to Detect and Protect against Viruses) on Your Devices	Freq.	Percentage %
No	302	40.87
Sometimes	164	22.19
Yes	273	36.94

Trainees need to know about cybersecurity countermeasures that help to keep their devices and information secure. Table 7 shows the rate in which respondents agree with the research questions on a Likert scale. A total of 558 (75.51%) respondents completely agree that antivirus and security software must be downloaded from licensed and trusted sources, 124 (16.78%) respondents agreed, and 49 (6.63%) respondents are neutral regarding whether antivirus and security software should be downloaded from licensed and trusted sources. A total of 3 (0.41%) respondents disagreed and 5 (0.68%) respondents strongly disagreed that antivirus and security software must be downloaded from licensed and trusted sources. The majority of the respondents (509 (68.88)) completely agreed that antivirus software must be up to date; similarly, 162 (21.92%) also agreed that antivirus software must be up to date. A total of 58 (7.85%) respondents did not know (i.e., neutral to the research question), 6 (0.81%) respondents disagreed and 4 (0.54%) respondents strongly disagreed that antivirus software must be up to date. A total of 267 respondents (36.13%) completely agreed that they were able to recognise sites that will infect their computer with viruses if they visit them and download their programs; similarly, 227 (30.72%) respondents agreed with this statement. A total of 198 (26.79%) respondents did not know (i.e., neutral), 30 (4.06%) respondents disagreed and 17 (2.30%) respondents strongly disagreed that they were able to recognise sites that will infect their computer with viruses if they visit them and download their programs. A total of 360 respondents (48.71%) completely agreed that the firewall (a program that protects the network (the internet)) must be activated in all the devices they use. Similarly, 242 (32.75%) respondents agreed with this statement. A total of 125 (16.91%) respondents did not know (i.e., neutral), 11 (1.49%) respondents disagreed, and 1 (0.14%) respondent strongly disagreed that the firewall must be activated in all the devices they use. A total of 240 respondents (32.48%) completely agreed that they felt that all the devices they used were safe. Similarly, 281 (38.02%) respondents agreed with this statement. A total of 140 (18.94%) respondents did not know (i.e., neutral), 70 (9.47%) respondents disagreed, and 8 (1.08%) respondents strongly disagreed that they felt that all the devices they used were safe. A total of 480 respondents (64.95%) totally agreed that they must use two-factor verification (for example, the method of entering Mubashir for the Al Rajhi Bank application and entering the verification code sent by text message) if it is available. Similarly, 187 (25.30%) respondents also agreed with this statement. A total of 55 (7.44%) respondents did not know (i.e., neutral), 13 (1.76%) respondents disagrees, and 4 (0.45%) respondents strongly disagrees that they must use two-factor verification if it is available. A total of 173 (23.4%) respondents completely agreed, 158 (21.4%) respondents agreed, 129 (17.5%) respondents did not know, 162 (21.9%) respondents disagreed, and 117 (15.8%) respondents strongly disagreed with the statement that public networks (internet located in airports, parks, and malls) can be used and are safe to use on personal devices. A total of 144 (19.5%) respondents totally agreed, 201 (27.2%) respondents agreed, 126 (17.1%) respondents did not know, 179 (24.2%) respondents disagreed, and 89 (12.0%) respondents strongly disagreed with the statement that attachments (sent files such as Word files or others) sent to your email or social networks may be opened without worry. Lastly, 224 (30.3%) respondents totally agreed, 209 (28.3%) respondents agreed, 110 (14.9%) respondents did not know, 171 (23.1%) respondents disagreed, and 25 (3.4%) respondents strongly disagreed with the statement that their passwords must be changed periodically.

4.2.4. Password Mechanism

Cybersecurity countermeasures include strong passwords to protect accounts and information. Passwords are one of the authentication methods which needs to be strong. Characteristics that are recommended for a strong password are a password length of at least 12 characters and a password that contains alpha (capital and small letters), numeric, and at least one special character (symbols) [48]. Therefore, in this survey, we assessed how the trainees manage their passwords and their knowledge about them, with the data summarised in Table 8.

Table 7. Respondents perception of cybersecurity countermeasures.

Questions	Totally Agree	Agree	Do Not Know	Disagree	Strongly Disagree	Total
Antivirus and security software must be downloaded from licensed and trusted sources.	558	124	49	3	5	739
%	75.51	16.78	6.63	0.41	0.68	100.00
Antivirus software must be up to date.	509	162	58	6	4	739
%	68.88	21.92	7.85	0.81	0.54	100.00
I feel that all the devices I use are safe.	240	281	140	70	8	739
%	32.48	38.02	18.94	9.47	1.08	100.00
I am familiar with sites that will infect my computer with viruses if I visit them and download their programs.	267	227	198	30	17	739
%	36.13	30.72	26.79	4.6	2.30	100.00
The firewall (a program that provides protection for the network (the internet)) must be activated in all the devices we use.	360	242	125	11	1	739
%	48.71	32.75	16.91	1.49	0.14	100.00
We must use two-factor verification (example: the method of entering Mubashir for the Al Rajhi Bank application and entering the verification code sent by text message) if it is available.	480	187	55	13	4	739
%	64.95	25.30	7.44	1.76	0.54	100.00
Public networks (internet located in airports, parks, and malls) can be used and are safe to use on personal devices.	173	158	129	162	117	739
%	23.4	21.4	17.5	21.9	15.8	100.00
You can open any attachments (sent files such as Word files or others) sent to your email or social networks without worry.	144	201	126	179	89	739
%	19.5	27.2	17.1	24.2	12.0	100.00

Table 8. Respondents perception of data protection and security regarding passwords.

Questions	Totally Agree	Agree	Do Not Know	Disagree	Strongly Disagree	Total
I can use passwords that were previously used.	118	205	94	231	91	739
%	16.0	27.7	12.7	31.3	12.3	100.00
One password can be used for multiple sites.	145	224	72	186	112	739
%	19.6	30.3	9.7	25.2	15.2	100.00
Our passwords can be shared with others.	51	46	34	145	463	739
%	6.9	6.2	4.6	19.6	62.7	100.00
What annoys me is that I have long, strong, and different passwords for several sites, and it is hard for me to remember them all.	259	221	78	120	61	739
%	35.0	29.9	10.6	16.2	8.3	100.00
We must log out of our accounts (e.g., email, university website, bank applications, etc.) when work is complete.	365	200	80	71	23	739
%	49.4	27.1	10.8	9.6	3.1	100.00
Private passwords should not be recorded on paper or in device notes.	226	173	108	162	70	739
%	30.6	23.4	14.6	21.9	9.5	100.00
We have to remember passwords without going back to the device, and we do not let the device remember our passwords.	278	238	103	96	24	739
%	37.6	32.2	13.9	13.0	3.2	100.00
We must update the internet browser (the browser we use to visit sites such as Chrome, Safari, and others) and make sure to update it constantly.	381	251	92	10	5	739
%	51.6	34.0	12.4	1.4	0.7	100.00
We must constantly check browser links (the URLs that appear at the top of the page, i.e., https: //www.google.com/ (accessed on 1 March 2023))	379	225	100	26	9	739
%	51.3	30.4	13.5	3.5	1.2	100.00

Table 8. *Cont.*

Questions	Totally Agree	Agree	Do Not Know	Disagree	Strongly Disagree	Total
Always use the incognito browser (users usually activate it when they connect to the internet from public networks such as coffee shops, airports, or public offices as it contributes to protecting privacy and your search history will not be saved after.	235	220	217	48	19	739
%	31.8	29.8	29.4	6.5	2.6	100.00
Passwords are secure if they are 12 characters long and contain lowercase and uppercase letters, numbers, special characters ($, &, ;, @, etc.), and punctuation.	430	218	45	43	3	739
%	58.19	29.50	6.09	5.82	0.41	100.00
Our passwords must be changed periodically.	224	209	110	171	25	739
%	30.3	28.3	14.9	23.1	3.4	100.00

Respondents were asked some security questions about their user password and the necessity to protect them. A total of 118 (16.0%) respondents totally agreed that they could use the passwords that have been previously used, 205 (27.7%) respondents agreed, 94 (12.7%) respondents dis not know, 231 (31.3%, the highest percentage) disagreed, and 91 (12.3%) respondents strongly disagreed. A total of 145 (19.6%) respondents agreed that one password could be used for multiple sites, 224 (30.3%, the highest percentage) respondents agreed, 72 (9.7%) respondents did not know, 186 (25.2%) respondents disagreed, and 112 (15.2%) respondents strongly disagreed. A total of 51 (6.9%) respondents totally agreed that passwords could be shared with others, 46 (6.2%) respondents agreed, 43 (4.6%) respondents did not know, 145 (19.6%) respondents disagreed, and 463 (62.7%) respondents (the highest percentage) strongly disagreed. A total of 259 (35.0%) respondents agreed that it is annoying to have long, strong, and different passwords for several sites and it was hard for them to remember them all, 221 (29.9%) respondents agreed, 78 (10.6%) respondents did not know, 120 (16.2%) respondents disagreed, and 61 (8.3%) respondents strongly disagreed. A total of 365 (49.4%) respondents (the highest percentage) totally agrees that they must log out of their accounts (e.g., email, university website, bank applications, etc.) when work is complete, 200 (27.1%) respondents agreed, 80 (10.8%) respondents did not know, 71 (9.6%) respondents disagreed, and 23 (3.1%) respondents strongly disagreed.

4.2.5. Data Protection Through Social Media Privacy

The last area of cybersecurity countermeasures this survey assesses is data protection and privacy. Table 9 shows the responses to data protection through social media privacy in detail.

Table 9. Respondents' perception of social media privacy.

Questions	Totally Agree	Agree	Do Not Know	Disagree	Strongly Disagree	Total
There is no harm in posting personal photos on social media.	131	154	149	154	151	739
%	17.7	20.8	20.2	20.8	20.4	100.00
There is no harm in accepting an extension from anyone on social media.	123	161	131	176	148	739
%	16.6	21.8	17.7	23.8	20.0	100.00
There is no harm in sharing your current location on social media.	105	121	107	174	232	739
%	14.2	16.4	14.5	23.5	31.4	100.00
There is no harm in sharing current job information on social media and updating the information continuously.	113	111	128	175	212	739
%	15.3	15.0	17.3	23.7	28.7	100.00
I know how to report any risks or threats (such as harassment or bullying) that I face when using social media.	323	238	120	36	22	739
%	43.7	32.2	16.2	4.9	3.0	100.00

Respondents were further asked some questions on data protection through social media. A total of 131 (17.7%) respondents agreed that there was no harm in posting personal photos on social media, 154 (20.8%) respondents agreed, 149 (20.2%) respondents did not know, 154 (20.4%) disagreed, and 151 (20.4%) respondents strongly disagreed. A total of 123 (16.6%) respondents totally agreed that there was no harm in accepting an extension from anyone on social media, 161 (21.8%) respondents agreed, 131 (17.7%) respondents did not know, 176 (23.8%) disagreed, and 148 (20.%) respondents strongly disagreed. A total of 105 (14.2%) respondents agreed that there was no harm in sharing your current location on social media, 121 (16.4%) respondents agreed, 107 (23.5%) respondents did not know, 174 (23.5%) respondents disagreed, and the highest percentage (31.4%, 232 respondents) strongly disagreed. About 113 (15.3%) respondents agreed that there was no harm in sharing current job information on social media and updating the data continuously., 111 (15.0%) respondents agreed, 128 (17.3%) respondents did not know, 175 (23.7%) respondents disagreed, and the highest percentage (28.7%, 212 respondents) strongly disagreed. Lastly, the highest percentage of respondents (323, 43.7%) totally agreed that they knew how to report any risks or threats (such as harassment or bullying) that they may face when using social media, 238 (32.2%) respondents agreed, 120 (16.2%) respondents did not know, 36 (4.9%) respondents disagreed, and 22 (3.0%) respondents strongly disagreed. At the end of this survey, we conducted an analysis to find out the extent to which trainees are attracted to matters related to cybersecurity and attend seminars, and the importance of raising awareness about cybersecurity, with the results shown in Tables 10–12.

Table 10. Previously attended or participated in an awareness program on cybersecurity.

Have You Previously Attended or Participated in an Awareness Program on Cybersecurity?	**Freq.**	**Percentage %**
No	507	68.6
Yes	232	31.4
Total	739	100.00
How long was the program you attended?		
1 to 3 days	40	5.4
3 to 5 days	21	2.8
Less than a day	142	19.2
More than 5 days	29	3.9
Total	232	31.4

Table 10 shows that 232 (31.4%) respondents had previously attended or participated in an awareness program on cybersecurity, while a higher percentage of respondents (507, 68.6%) had not previously attended or participated in an awareness program on cybersecurity. Out of the 232 respondents that had participated in an awareness program on cybersecurity, 40 respondents attended an awareness program that lasted for one to three days, 21 respondents attended an awareness program that lasted for three to five days, 142 respondents attended an awareness program that lasted for less than a day, and lastly, 29 respondents participated in an awareness program on cybersecurity that lasted for more than five days.

Table 11. Participant perceptions on the necessity of awareness programs.

Questions	**Totally Agree**	**Agree**	**Do Not Know**	**Disagree**	**Strongly Disagree**	**Total**
It is necessary to have an awareness program on cyber security these days to protect others from falling victim to hacking	506	164	58	8	3	739
%	68.5	22.2	7.8	1.1	0.4	100.00
Filling out this questionnaire was interesting	352	261	69	43	14	739
%	47.6	35.3	9.3	5.8	1.9	100.00

Respondents were questioned on the necessity of an awareness program on cybersecurity; 506 (68.5%) respondents totally agreed that it was necessary to have an awareness program on cybersecurity these days to protect others from falling victim to hacking, 164 (22.2%) respondents agreed, 58 (7.8%) respondents did not know, 8 respondents disagreed, and a very low proportion of respondents (3, 0.4%) strongly disagreed. However, the majority of the respondents (352, 47.6%) totally agreed that filling out this questionnaire was interesting and exciting, 261 (35.3) respondents agreed, 69 (9.3%) respondents did not know, 43 (5.8%) respondents disagreed, and very few respondents (14, 1.9%) strongly disagreed.

Table 12. Previous discussions of security aspects

This Is the First Time I Have Discussed the Security Aspects of the Devices I Have Used Regularly.	Freq.	Percentage %
No	60	8.1
Sometimes	205	27.7
Yes	474	64.1
Total	739	100.00

A total of 474 (64.1%) respondents said that this was the first time they had discussed the security aspects of the devices they use on a regular basis, 205 (27.7%) respondents said that they sometimes discuss the security aspects of the devices they use on a regular basis, while 60 (8.1%) respondents do not discuss the security aspects of the devices they use on a regular basis.

Figure 4 shows a bar graph between the type of operating system on respondents' devices and the tendency of being attacked, which was extracted from this survey. The chart shows that respondents with Windows devices are more likely to be either attacked by viruses, scammed, or hacked.

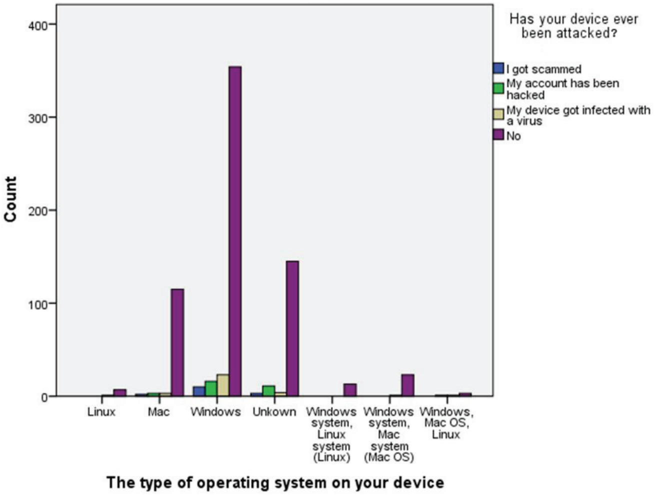


Figure 4. Relationship between type of operating systems and attacks.

4.3. Chi-Square Tests to Hypothesis Statement

This part of the study was conducted to help assess whether the likelihood of attacks on respondents' devices is dependent on the operating system they have installed on their devices. A Pearson's chi-squared test was used to evaluate the differences, where chi-square test use two categorical variables of independence: null hypothesis (0) if the variables are independent, and alternative hypothesis (a) if the variables are dependent. If the *p*-value is less than 0.05, we will reject the null hypothesis and can conclude that the two groups are dependent on each other. If the *p*-value is greater than 0.05, we will not reject the null hypothesis and can conclude that the two groups are independent of each other [36]. The *p*-value in Table 13 is greater than the 0.05 significance level and thus we do not reject the null hypothesis and conclude that the respondents' type of operating system they use, either Windows, Linux, or Mac, is not linked to the likelihood of being attacked.

That is, there is no relationship between the operating system and the whether the device will be attacked.

Table 13. Chi-Square Tests on OS and hacking.

Chi-Square Tests	Value	df	*p*-Value
Pearson Chi-Square	19.448a	18	0.365

In order to evaluate if respondents' perceptions of an awareness program on cyber security is dependent on their educational system, we used the chi-squared test of independence. chi-square test use two categorical variables of independence: null hypothesis (0): if the variables are independent, and alternative hypothesis (a): if the variables are dependent. Furthermore, this test was used to assess if respondents' perceptions on the necessity to have an awareness program on cyber security were dependent on their educational system or not. The *p*-value for both research questions in Table 14 is greater than the 0.05 significance level. We reject the null hypothesis and conclude that respondents who attended or participated in an awareness program on cybersecurity are not dependent on their educational system. Similarly, respondent perception of the necessity of having an awareness program on cybersecurity is not dependent on their educational system.

Table 14. Chi-squared test on security awareness.

Chi-Square Tests, Pearson Chi-Square Educational System	Value	df	*p*-Value
Previously attended or participated in an awareness program on cyber security?	0.348	2	0.840
It is necessary to have an awareness program on cybersecurity these days to protect others from hacking and falling victim.	10.989	8	0.202

5. Discussion and Limitations

The analyses were presented in frequency distribution tables, charts, percentages, and proportions using Chi-square test techniques. However, most of the respondents were female (61.98%), followed by males (38.02), out of which 98.78% attended the Technical and Vocational Training Corporation. The results in Table 1 report that most respondents were diploma holders (97.43%), while very few were bachelor degree holders (2.57%). Most respondents (54.53%) had a Windows operating system on their device, 16.64% had a Mac operating system, and few had a Linux operating system. In contrast, some respondents 5.69% had more than one operating system on their device. However, the majority of the respondents operating systems on their devices were updated automatically as the auto update feature was enabled, while 34.91% of respondents updated the operating systems on their devices manually, few respondents 4.06% had not updated their operating system on their device before because it was new, and 7.98% had never updated the operating system on their device. A higher percentage of the respondents used email, while few respondents only sometimes used email. The time respondents spent on social media was assessed, and the majority spent most of their time on Snapchat, WhatsApp, Instagram, and YouTube. The result reveal that the majority of respondents' devices have not been attacked before, at about 89.31%, while 4.47% had been infected by a virus, 4.19% had been hacked, and 2.03% had been scammed. A total of 0.4% of respondents who had been scammed did nothing afterwards and 1.6% informed the concerned authority and their bank card was suspended to secure their account from losing money without their authentication. A total of 1.4% of respondents who had had their account hacked also contacted the support for the hacked program, 0.8% did nothing, 0.8% informed everyone that their account had been hacked at the same time as contacting the support for the hacked program, while only 1.1% told everyone that their account was hacked. Some respondents' devices were infected with a virus, and of these respondents, 0.9% ran a device scan program and deleted the files associated with the virus as a solution and 0.8% of these respondents went to tech support this while also running a device scan program to detect the viruses in the device. In contrast, 1.2% of respondents deleted the related virus files. In order to provide and build solutions to enhance protection, 36.94% of respondents had antivirus software installed to detect and protect devices against viruses, while 22.19% had only once or sometimes installed it on their devices. Respondents were assessed on their perception of the use and importance of antivirus software. Most agreed that antivirus and security software must be downloaded from licensed and trusted sources, while very few disagreed. A higher percentage of the respondents also agreed that antivirus software must be up to date, and very few disagreed. The responses to security questions showed that the majority disagreed with reusing previously used passwords and the majority agreed that one password can be used for multiple sites. In contrast, most of the respondents strongly disagreed with sharing their passwords with others. Finally, the perceptions of social media privacy were accessed, and most of the respondents strongly disagreed with the statement that there is no harm in sharing their current location on social media. Similarly, most respondents strongly disagreed that there was no harm in sharing current job information on social media and updating the information continuously. Theses results further reveal that most respondents know how to report any risks or threats faced on social media. Finally, respondents were asked about their awareness of cyber security programs. The results revealed that only 31.4% of respondents had previously attended or participated in an awareness program on cyber security. In contrast, the rest (68.6%) have never attended or participated in any awareness program on cyber security.

The results indicate that a significant portion of the awareness and responses concerning security and data privacy hinges on individual behavior and decision making, followed by the policies and guidelines set by organizations for their members. Making informed decisions and devising strategies to protect individuals and raise awareness about privacy and security when using personal devices, or those owned by an organization, can be challenging due to factors such as commitment, cost, and suitability for the specific environment.

In response to these challenges, researchers [4] have proposed the Nudge model, an approach that focuses on gentle interventions or prompts to encourage users to make more advantageous choices, considering both individual behavior and organizational needs. Rooted in behavioral economics, the Nudge concept assists individuals by subtly guiding them toward better decisions rather than enforcing rigid rules or regulations. This approach enables users to make more informed choices about privacy and security, fostering a safer online environment for both individuals and organizations.

5.1. Reliability Test

We have addressed the quality criteria using a reliability test; the closer the coefficient is to 1.0, the greater the internal consistency of items that are variables in the scale. Table 15 provides the value for Cronbach's alpha [49], showing a value of 0.808, indicating a high internal consistency level for our scale for these data. The item for each question presents Cronbach's alpha if the item is deleted. The column would present the value of Cronbach's alpha if a particular item were deleted from the scale shown in Table 15.

Table 15. Reliability test statistics.

Cronbach's Alpha	N of Items
0.808	30

5.2. Limitations

Although there are some limitations, this survey provides help and guidance for the TVTC to increase cybersecurity awareness and enhance existing policies. Nevertheless, several limitations have been faced and should be avoided in the future, such as the data collection time and the sample size. Another limitation of this work is the number of questions, which can be optimized in the future to cover the most suitable cybersecurity awareness information instead of expanding it to more dimensions, such as the behavior on social media.

6. Conclusions and Future Work

Cybersecurity awareness is one of the most significant aspects of modern life that should be recognized and improved, particularly at educational institutions due to their direct connection to the network and the internet. Therefore, awareness of cybersecurity concepts and mechanisms should be improved, such as establishing solid passwords, upgrading systems, and employing antivirus software with the main aims of preventing data leaks and device hacking. Therefore, this quantitative study was conducted on trainees at the TVTC institution in the Kingdom of Saudi Arabia utilizing questionnaires. The results indicated that the majority needed an appropriate foundation in cyber security expertise and statistical analysis. Therefore, awareness must be raised among TVTC trainees and training on cyber security strategies that help them protect their devices and data should be implemented. Furthermore, a focus should be placed on developing plans and strategies for cybersecurity awareness among students and trainees of educational institutions to enable users to understand the threats and factors that lead to weaknesses on their devices and data, and their effectiveness should be tested continuously. Based on the survey in our paper, we suggest the following:

- A course should be included in each foundation specialization to raise awareness of cybersecurity, which can be implemented as an electronic course.
- Trainees should be offered the chance to specialize in technology under the supervision of cybersecurity specialists who conduct awareness campaigns in the institution's departments (for example, during a week, each day is devoted to a section of the institution).
- Sensitive applications such as banks or university pages should contain an awareness list regarding the application's security, so the reader is encouraged to read it before opening the application.
- During job interviews, a set of cybersecurity questions and their basic concepts should be presented to test the applicability of the candidates.
- Cybersecurity awareness should be raised by conducting educational experiments to attempt to penetrate the trainees' devices to educate them about possible vulnerabilities and the usefulness of auxiliary programs such as antivirus software.
- The Nudge model [4] is a helpful factor to assist users in making better privacy and security decisions online for particular individuals who may not have the knowledge or motivation to make optimal choices on their own. The Nudge model includes several additional dimensions such as providing a realistic view of risks by making information clear and consistent, improving the user interface, which helps in increasing cognitive awareness, and also introducing incentives to encourage users to act. By providing gentle guidance, nudges can encourage users to take actions that will improve their online safety without feeling overwhelmed or burdened by complex decision-making processes.

A more scalable questionnaire can be implemented to increase the sample size and include more than one educational institution for comparison. Furthermore, another study could be conducted after providing a cybersecurity awareness course to measure its impact of on the respondents. The questionnaire also can be expanded to include members, employees, and trainees of industrial sectors to compare the results with academic institutions.

Author Contributions: Conceptualization, S.A. (Shouq Alrobaian) and A.A.; methodology, S.A. (Shouq Alrobaian); software, S.A. (Shouq Alrobaian); validation, S.A. (Saif Alshahrani) and S.A. (Shouq Alrobaian) and investigation, S.A. (Saif Alshahrani); resources, A.A.; writing—original draft preparation, S.A. (Shouq Alrobaian); writing—review and editing, A.A.; visualization, S.A. (Shouq Alrobaian); supervision, A.A.; project administration, A.A.; funding acquisition, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work through large group Research Project under grant number RGP.2/550/44.

Data Availability Statement: data is unavailable due to privacy or ethical re-strictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jang-Jaccard, J.; Nepal, S. A survey of emerging threats in cybersecurity. *J. Comput. Syst. Sci.* **2014**, *80*, 973–993. [CrossRef]
2. Reis, J.; Amorim, M.; Melão, N.; Matos, P. Digital transformation: A literature review and guidelines for future research. In Proceedings of the World Conference on Information Systems and Technologies ; Springer: Berlin/Heidelberg, Germany, 2018; pp. 411–421.
3. Alharbi, T.; Tassaddiq, A. Assessment of cybersecurity awareness among students of Majmaah University. *Big Data Cogn. Comput.* **2021**, *5*, 23. [CrossRef]
4. Acquisti, A.; Adjerid, I.; Balebako, R.; Brandimarte, L.; Cranor, L.F.; Komanduri, S.; Leon, P.G.; Sadeh, N.; Schaub, F.; Sleeper, M.; et al. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–41. [CrossRef]
5. Guarino, A.; Malandrino, D.; Zaccagnino, R. An automatic mechanism to provide privacy awareness and control over unwittingly dissemination of online private information. *Comput. Netw.* **2022**, *202*, 108614. [CrossRef]
6. Lippi, M.; Pałka, P.; Contissa, G.; Lagioia, F.; Micklitz, H.W.; Sartor, G.; Torroni, P. CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service. *Artif. Intell. Law* **2019**, *27*, 117–139. [CrossRef]
7. Guarino, A.; Lettieri, N.; Malandrino, D.; Zaccagnino, R. A machine learning-based approach to identify unlawful practices in online terms of service: Analysis, implementation and evaluation. *Neural Comput. Appl.* **2021**, *33*, 17569–17587. [CrossRef]
8. Galinec, D.; Možnik, D.; Guberina, B. Cybersecurity and cyber defence: National level strategic approach. *Autom. Časopis Autom. Mjer. Elektron. Računarstvo Komun.* **2017**, *58*, 273–286. [CrossRef]
9. Oliver, D.; Randolph, A.B. Hacker definitions in information systems research. *J. Comput. Inf. Syst.* **2022**, *62*, 397–409. [CrossRef]
10. Craigen, D.; Diakun-Thibault, N.; Purse, R. Defining cybersecurity. *Technol. Innov. Manag. Rev.* **2014**, *4* . [CrossRef]
11. National Cybersecurity Authority, Saudi Arabia. 2017. Available online: https://nca.gov.sa/en/about (accessed on 10 January 2023) .
12. Saudi Federation for Cybersecurity, Programming & Drones. 2017. Available online: https://safcsp.org.sa/en/ (accessed on 10 February 2023).
13. Almudaires, F.; Rahman, M.H.; Almudaires, M. An Overview of Cybersecurity, Data Size and Cloud Computing in light of Saudi Arabia 2030 Vision. In Proceedings of the 2021 International Conference on Information Technology (ICIT),Amman, Jordan, 14–15 July 2021; IEEE: New York, NY, USA, 2021; pp. 268–273.
14. Alzubaidi, A. Measuring the level of cyber-security awareness for cybercrime in Saudi Arabia. *Heliyon* **2021**, *7*, e06016. [CrossRef]
15. Cyberattacks Hit 95% of Saudi Businesses Last Years, Says Study. 2020. Available online: https://www.arabnews.com/node/1718596/saudi-arabia (accessed on 22 October 2022).
16. Nurse, J.R. Cybersecurity Awareness. *arXiv* **2021**, arXiv:2103.00474.
17. Majmaah University. 2023. Available online: https://www.mu.edu.sa/en (accessed on 15 March 2023).
18. Khader, M.; Karam, M.; Fares, H. Cybersecurity Awareness Framework for Academia. *Information* **2021**, *12*, 417. [CrossRef]
19. Capital Area Finance Authority. 2023. Available online: https://thecafa.org (accessed on 30 March 2023).
20. Nidup, Y. Awareness about the Online Security Threat and Ways to Secure the Youths. *J. Cybersecur.* **2021**, *3*, 133. [CrossRef]
21. Taherdoost, H. Determining sample size; how to calculate survey sample size. *Int. J. Econ. Manag. Syst.* **2017**, *2* .
22. Taha, N.; Dahabiyeh, L. College students information security awareness: A comparison between smartphones and computers. *Educ. Inf. Technol.* **2021**, *26*, 1721–1736. [CrossRef]
23. Alqahtani, M.A. Cybersecurity Awareness Based on Software and E-mail Security with Statistical Analysis. *Comput. Intell. Neurosci.* **2022**, *2022*, 6775980 . [CrossRef]
24. Aldawood, H.; Skinner, G. Educating and raising awareness on cyber security social engineering: A literature review. In Proceedings of the 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), Wollongong, Australia, 4–7 December 2018; IEEE: New York, NY, USA, 2018; pp. 62–68.
25. Khushali, V. A Review on Fileless Malware Analysis Techniques. *Int. J. Eng. Res. Technol. (IJERT)* **2020**, *9* , 46–49. [CrossRef]

26. Mogal, M.M.; Gonsalves, F. How Two Factor Authentication Helps in Cybersecurity. *Int. Res. J. Mod. Eng. Technol. Sci.* **2022**, *4*, 2390–2395.
27. Arefin, M.T.; Uddin, M.R.; Evan, N.A.; Alam, M.R. Enterprise network: Security enhancement and policy management using next-generation firewall (NGFW). In Proceedings of the Computer Networks, Big Data and IoT: Proceedings of ICCBI 2020, Madurai, India, 19–20 December 2018; Springer: Berlin/Heidelberg, Germany, 2021; pp. 753–769.
28. Armstrong, L.; Phillips, J.G.; Saling, L.L. Potential determinants of heavier internet usage. *Int. J. Hum.-Comput. Stud.* **2000**, *53*, 537–550. [CrossRef]
29. Sosanya, O.V. Beyond Cyber Security Tools: The Increasing Roles Of Human Factors Furthermore, Cyber Insurance in the Survival of Social Media Organisations. Available online: https://www.cybsafe.com/research/beyond-cyber-security-tools-the-increasing-roles-of-human-factors-and-cyber-insurance-in-the-survival-of-social-media-organisations/ (accessed on 16 March 2023).
30. Rasool, A.; Jalil, Z. A review of web browser forensic analysis tools and techniques. *Res. J. Comput.* **2020**, *1*, 15–21.
31. Eke, H.N.; Odoh, N.J. The use of social networking sites among the undergraduate students of University of Nigeria, Nsukka. *Libr. Philos. Pract.* **2014**, 1–11.
32. Allen, I.E.; Seaman, C.A. Likert scales and data analyses. *Qual. Prog.* **2007**, *40*, 64–65.
33. Khonji, M.; Iraqi, Y.; Jones, A. Phishing detection: A literature survey. *IEEE Commun. Surv. Tutorials* **2013**, *15*, 2091–2121. [CrossRef]
34. Monnappa, K. *Learning Malware Analysis: Explore the Concepts, Tools, and Techniques to Analyze and Investigate Windows Malware*; Packt Publishing Ltd.: Birmingham, UK, 2018.
35. Souppaya, M.; Scarfone, K. *Guide to Enterprise Patch Management Technologies*; NIST Special Publication: Washington, DC, USA, 2013; Volume 800, p. 40.
36. Turney, S. Chi-Square Tests: Types, Formula & Examples. 2022. Available online: www.scribbr.com/statistics/chi-square-tests/ (accessed on 26 October 2022).
37. Frick, R.W. Accepting the null hypothesis. *Mem. Cogn.* **1995**, *23*, 132–138. [CrossRef] [PubMed]
38. Samonas, S.; Coss, D. The CIA strikes back: Redefining confidentiality, integrity and availability in security. *J. Inf. Syst. Secur.* **2014**, *10* , 21–45.
39. Shen, L. The NIST cybersecurity framework: Overview and potential impacts. *Scitech Lawyer* **2014**, *10*, 16.
40. Facebook Mainpage. Available online: https://www.facebook.com/public/Main-Page (accessed on 18 February 2023).
41. Instgram. Available online: https://www.instagram.com (accessed on 18 February 2023).
42. Linkedin. Available online: https://www.linkedin.com (accessed on18 February 2023).
43. Snapchat. Available online: https://www.snapchat.com (accessed on18 February 2023).
44. Twitter. Available online: https://twitter.com (accessed on 18 February 2023).
45. YouTube. Available online: https://www.youtube.com (accessed on 18 February 2023).
46. WhatsApp. Available online: https://www.whatsapp.com (accessed on 18 February 2023).
47. Baskerville, R.; Rowe, F.; Wolff, F.C. Integration of information systems and cybersecurity countermeasures: an exposure to risk perspective. *ACM SIGMIS Database DATABASE Adv. Inf. Syst.* **2018**, *49*, 33–52. [CrossRef]
48. Kruger, H.; Steyn, T.; Dawn Medlin, B.; Drevin, L. An empirical assessment of factors impeding effective password management. *J. Inf. Priv. Secur.* **2008**, *4*, 45–59. [CrossRef]
49. Bonett, D.G.; Wright, T.A. Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *J. Organ. Behav.* **2015**, *36*, 3–15. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

big data and cognitive computing

MDPI

Review

Ransomware Detection Using Machine Learning: A Survey

Amjad Alraizza [1,*] and Abdulmohsen Algarni [2]

[1] Department of Information Systems, King Khalid University, Alfara, Abha 61421, Saudi Arabia
[2] Department of Computer Science, King Khalid University, Alfara, Abha 61421, Saudi Arabia; a.algarni@kku.edu.sa
* Correspondence: 444800503@kku.edu.sa

Abstract: Ransomware attacks pose significant security threats to personal and corporate data and information. The owners of computer-based resources suffer from verification and privacy violations, monetary losses, and reputational damage due to successful ransomware assaults. As a result, it is critical to accurately and swiftly identify ransomware. Numerous methods have been proposed for identifying ransomware, each with its own advantages and disadvantages. The main objective of this research is to discuss current trends in and potential future debates on automated ransomware detection. This document includes an overview of ransomware, a timeline of assaults, and details on their background. It also provides comprehensive research on existing methods for identifying, avoiding, minimizing, and recovering from ransomware attacks. An analysis of studies between 2017 and 2022 is another advantage of this research. This provides readers with up-to-date knowledge of the most recent developments in ransomware detection and highlights advancements in methods for combating ransomware attacks. In conclusion, this research highlights unanswered concerns and potential research challenges in ransomware detection.

Keywords: machine learning; ransomware techniques; cybersecurity; ransomware detection; ransomware attacks

Citation: Alraizza, A.; Algarni, A. Ransomware Detection Using Machine Learning: A Survey. *Big Data Cogn. Comput.* **2023**, *7*, 143. https://doi.org/10.3390/bdcc7030143

Academic Editors: Peter R.J. Trim, Yang-Im Lee and Min Chen

Received: 18 May 2023
Revised: 7 August 2023
Accepted: 11 August 2023
Published: 16 August 2023

1. Introduction

The rapid proliferation of ransomware attacks has emerged as one of the most significant cybersecurity threats facing organizations today. In recent years, ransomware has become an increasingly popular tool with which cybercriminals extort money from victims by encrypting their data and demanding payment for a decryption key. The impact of ransomware attacks has been felt across all industries, from healthcare and finance to government and education. Given the high stakes involved, it is crucial to understand the nature of ransomware attacks, how they spread, and the potential consequences of falling victim to one [1]. The importance of research in this area cannot be overstated. With the threat of ransomware attacks continuing to grow, there is a pressing need for scholars and practitioners to delve deeper into the problem and identify effective strategies for prevention and mitigation. This paper aims to contribute to this effort by providing a comprehensive overview of the ransomware threat landscape, analyzing the factors that contribute to the spread of ransomware, and exploring potential avenues for future research. By shedding light on this critical issue, we hope to help individuals and organizations better-protect themselves against ransomware attacks and mitigate the potential damage caused by these malicious programs [1].

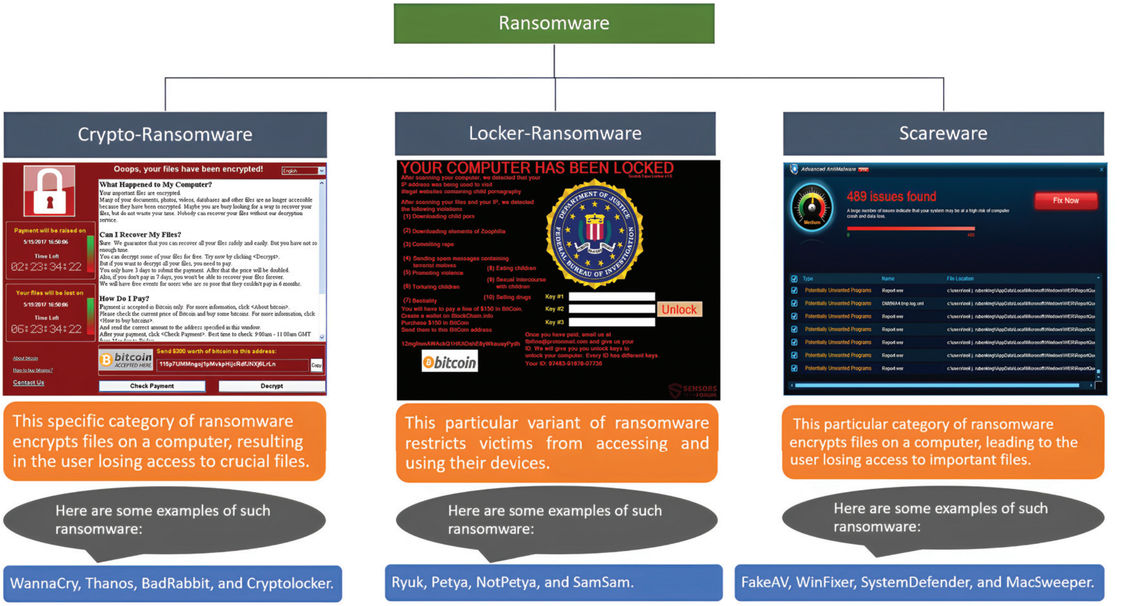
This paper is organized as follows: Section 2 introduces the concept of ransomware and how it works. It also discusses the different types of ransomware attacks, such as encrypting ransomware, locker ransomware, and scareware. Section 3 describes the methodology used for this paper. Section 4 provides studies of machine-learning-based ransomware-detection systems developed by researchers. It discusses the methodology used, the performance achieved, and the limitations of each system. It also discusses the challenges of collecting and preprocessing data for ransomware detection using machine learning. Section 5 provides an in-depth analysis of the evolution of ransomware over the last twelve years. Section 6 provides an overview of the existing ransomware detection techniques, including signature-based detection, behavior-based detection, and machine-learning-based detection. Furthermore, it discusses the different evaluation metrics used for measuring the performance of machine learning models for ransomware detection. It also focuses on the use of machine learning techniques for ransomware detection. It discusses the different machine learning algorithms used for this purpose, such as decision trees, random forests, support vector machines, and neural networks. It also addresses the different features used for ransomware detection using machine learning and covers the techniques used for feature selection. Section 7 discusses the challenges of developing effective machine-learning-based ransomware-detection systems. It also highlights future directions in this field, such as developing more robust and accurate models, incorporating real-time detection capabilities, and addressing the issue of adversarial attacks. Section 8 concludes what has been achieved in this research. This research offers a valuable resource for researchers and practitioners interested in developing effective ransomware-detection systems using machine-learning techniques.

2. Background

Ransomware encrypts information or computer systems and prevents unauthorized users from accessing them. Ransomware attacks use tactics, techniques, and procedures that can lock computers or encrypt data and are challenging for a computer professional to undo. They might also steal private information from victims' PCs and network systems. Individual PCs, commercial systems (and the data and software they contain), and industrial control systems are all potential targets for ransomware attacks. Additionally, we emphasize the variety of sensors that Internet of Things (IoT) users employ [1]. A ransomware attack employs private key encryption to prevent authorized users from accessing a system or data unless they pay a ransom (cash), typically in Bitcoin [2]. Ransomware operations may include data exfiltration techniques. Hackers steal private information from vulnerable networks and threaten to release it if the owner does not pay a ransom. The infection is disseminated through malicious advertising, email attachments, and connections to rogue websites. The attacker also sends a file (or files) with instructions for paying the ransom. Once the attacker has verified that the ransom has been paid, the victim can access the decryption key [3]. Files with encryption or ransomware infections frequently include extensions, such as Locky, Cryptolocker, Vault, Micro, Encrypted, TTTT, XYZ, ZZZ, Petya, etc. Each file's extension indicates the type of ransomware that affected it. Examples of ransomware include WannaCry, WannaCry.F, Fusob, TorrentLocker, CryptoWall, CryptoTear, and Reveton [4]. Figure 1 illustrates the classification of ransomware into three categories: scareware, locker ransomware, and crypto-ransomware [2,4].

Crypto is the most prevalent ransomware that targets computer systems and networks. Ransomware encrypts files and data using symmetric and asymmetric encryption algorithms. Even if the malicious software is removed from an infected computer or a compromised storage device is introduced into another system, crypto-ransomware renders the encrypted data unusable. Because the malware frequently does not corrupt imported essential data, the compromised device can still be used to pay the ransom [4]. Figure 2 provides a visual representation of crypto-ransomware, a form of malicious software that is becoming increasingly prevalent in cyberattacks [4].

However, by locking a computer or other device and demanding money, locker ransomware prevents its owner from using it. The workstation is affected by the locker ransomware, but saved data are not rendered inaccessible. Once the malicious program has been eliminated, the data are not altered. The data are often recoverable by connecting the infected storage device, such as a hard drive, to another machine. Individuals wanting to extort money from assault victims will not be drawn to locker ransomware. Figure 3 provides a visual representation of locker ransomware, a form of malicious software that is becoming increasingly prevalent in cyberattacks [4].



Ransomware

Crypto-Ransomware

Ooops, your files have been encrypted!

What Happened to My Computer?
Your important files are encrypted.
Many of your documents, photos, videos, databases and other files are no longer accessible because they have been encrypted. Maybe you are busy looking for a way to recover your files, but do not waste your time. Nobody can recover your files without our decryption service.

Can I Recover My Files?
Sure. We guarantee that you can recover all your files safely and easily. But you have not so enough time.
You can decrypt some of your files for free. Try now by clicking <Decrypt>.
But if you want to decrypt all your files, you need to pay.
You only have 3 days to submit the payment. After that the price will be doubled.
Also, if you don't pay in 7 days, you won't be able to recover your files forever.
We will have free events for users who are so poor that they couldn't pay in 6 months.

How Do I Pay?
Payment is accepted in Bitcoin only. For more information, click <About bitcoin>.
Please check the current price of Bitcoin and buy some bitcoins. For more information, click <How to buy bitcoins>.
And send the correct amount to the address specified in this window.
After your payment, click <Check Payment>. Best time to check: 9:00am - 11:00am GMT

Payment will be raised on
5/15/2017 16:50:06
Time Left
02:23:34:22

Your files will be lost on
5/19/2017 16:50:06
Time Left
06:23:34:22

Send $300 worth of bitcoin to this address:
115p7UMMngoj1pMvkpHijcRdfJNXj6LrLn

This specific category of ransomware encrypts files on a computer, resulting in the user losing access to crucial files.

Here are some examples of such ransomware:

WannaCry, Thanos, BadRabbit, and Cryptolocker.

Locker-Ransomware

YOUR COMPUTER HAS BEEN LOCKED

After scanning your computer, we detected that your IP address was being used to visit illegal websites containing child pornography

After scanning your files and your IP, we detected the following violations
(1) Downloading child porn
(2) Downloading elements of Zoophilia
(3) Commiting rape
(4) Sending spam messages containing terrorist motives
(5) Promoting violence
(6) Torturing children
(7) Bestiality
(8) Eating children
(9) Sexual intercourse with children
(10) Selling drugs

You will have to pay a fine of $150 in BitCoin.
Create a wallet on BlockChain.info
Purchase $150 in BitCoin
Send them to this BitCoin address

12mgfmmAWAckQ1HfUiDahE8yWkauayPydh

Once you have paid, email us at fbifine@protonmail.com and give us your ID. We will give you you unlock keys to unlock your computer. Every ID has different keys.
Your ID: 97483-91676-07736

This particular variant of ransomware restricts victims from accessing and using their devices.

Here are some examples of such ransomware:

Ryuk, Petya, NotPetya, and SamSam.

Scareware

Advanced AntiMalware

489 issues found

A large number of issues indicate that your system may be at a high risk of computer crash and data loss.

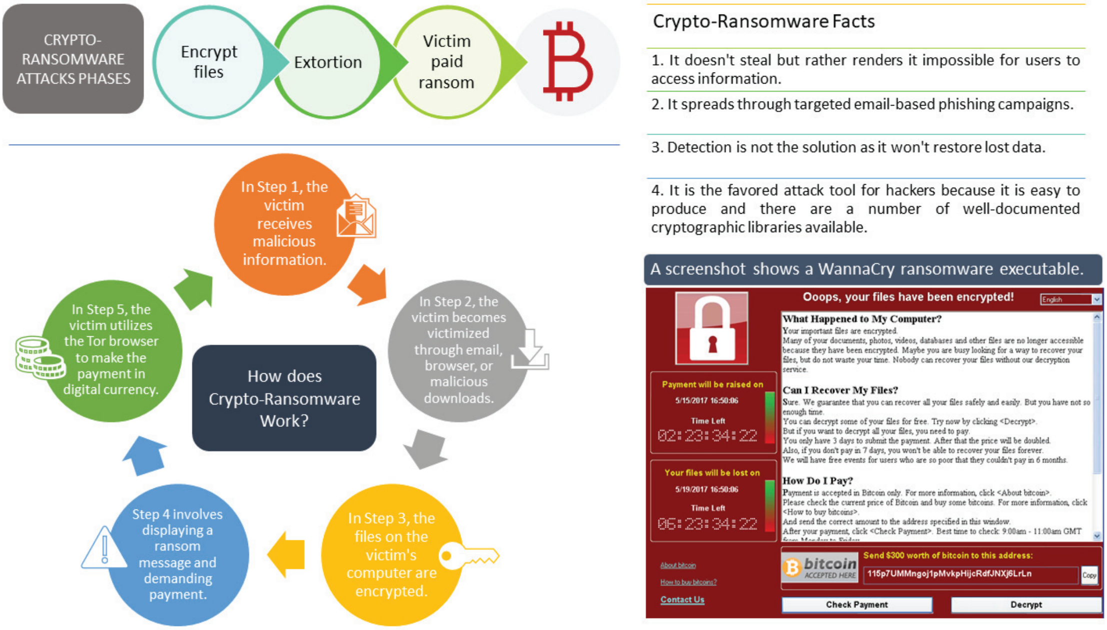
☑	Type	Name	File Location
☑	Potentially Unwanted Programs	Report.wer	c:\users\[illegible]\AppData\Local\Microsoft\Windows\WER\ReportQu
☑	Potentially Unwanted Programs	DMS9A4.tmp.log.xml	c:\users\[illegible]\AppData\Local\Microsoft\Windows\WER\ReportQu
☑	Potentially Unwanted Programs	Report.wer	c:\users\[illegible]\AppData\Local\Microsoft\Windows\WER\ReportQu
☑	Potentially Unwanted Programs	Report.wer	c:\users\[illegible]\AppData\Local\Microsoft\Windows\WER\ReportQu
☑	Potentially Unwanted Programs	Report.wer	c:\users\[illegible]\AppData\Local\Microsoft\Windows\WER\ReportQu
☑	Potentially Unwanted Programs	Report.wer	c:\users\[illegible]\AppData\Local\Microsoft\Windows\WER\ReportQu
☑	Potentially Unwanted Programs	Report.wer	c:\users\[illegible]\AppData\Local\Microsoft\Windows\WER\ReportQu

This particular category of ransomware encrypts files on a computer, leading to the user losing access to important files.

Here are some examples of such ransomware:

FakeAV, WinFixer, SystemDefender, and MacSweeper.

Figure 1. Types of ransomware [2,4].



CRYPTO-RANSOMWARE ATTACKS PHASES

Encrypt files

Extortion

Victim paid ransom

Crypto-Ransomware Facts

1. It doesn't steal but rather renders it impossible for users to access information.

2. It spreads through targeted email-based phishing campaigns.

3. Detection is not the solution as it won't restore lost data.

4. It is the favored attack tool for hackers because it is easy to produce and there are a number of well-documented cryptographic libraries available.

How does Crypto-Ransomware Work?

In Step 1, the victim receives malicious information.

In Step 2, the victim becomes victimized through email, browser, or malicious downloads.

In Step 3, the files on the victim's computer are encrypted.

Step 4 involves displaying a ransom message and demanding payment.

In Step 5, the victim utilizes the Tor browser to make the payment in digital currency.

A screenshot shows a WannaCry ransomware executable.

Ooops, your files have been encrypted!

What Happened to My Computer?
Your important files are encrypted.
Many of your documents, photos, videos, databases and other files are no longer accessible because they have been encrypted. Maybe you are busy looking for a way to recover your files, but do not waste your time. Nobody can recover your files without our decryption service.

Can I Recover My Files?
Sure. We guarantee that you can recover all your files safely and easily. But you have not so enough time.
You can decrypt some of your files for free. Try now by clicking <Decrypt>.
But if you want to decrypt all your files, you need to pay.
You only have 3 days to submit the payment. After that the price will be doubled.
Also, if you don't pay in 7 days, you won't be able to recover your files forever.
We will have free events for users who are so poor that they couldn't pay in 6 months.

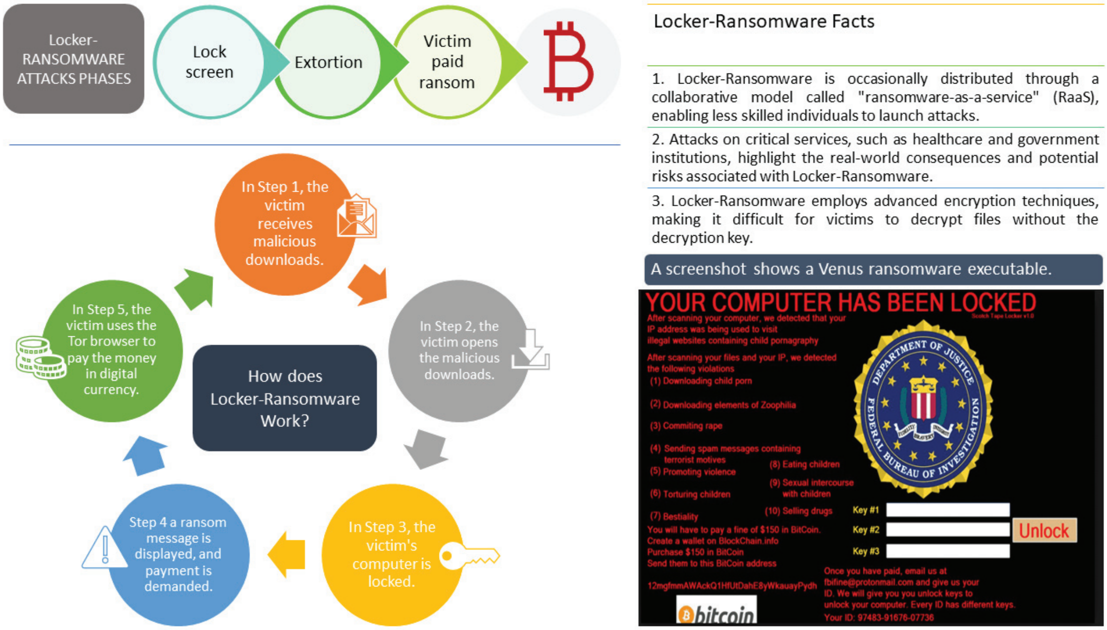
How Do I Pay?
Payment is accepted in Bitcoin only. For more information, click <About bitcoin>.
Please check the current price of Bitcoin and buy some bitcoins. For more information, click <How to buy bitcoins>.
And send the correct amount to the address specified in this window.
After your payment, click <Check Payment>. Best time to check: 9:00am - 11:00am GMT

Payment will be raised on
5/15/2017 16:50:06
Time Left
02:23:34:22

Your files will be lost on
5/19/2017 16:50:06
Time Left
06:23:34:22

Send $300 worth of bitcoin to this address:
115p7UMMngoj1pMvkpHijcRdfJNXj6LrLn

Figure 2. Crypto-ransomware [4].


Locker-RANSOMWARE ATTACKS PHASES
Lock screen
Extortion
Victim paid ransom
In Step 1, the victim receives malicious downloads.
In Step 2, the victim opens the malicious downloads.
In Step 3, the victim's computer is locked.
Step 4 a ransom message is displayed, and payment is demanded.
In Step 5, the victim uses the Tor browser to pay the money in digital currency.
How does Locker-Ransomware Work?
DEPARTMENT OF JUSTICE
FEDERAL BUREAU OF INVESTIGATION
FIDELITY
BRAVERY
INTEGRITY
Key #1
Key #2
Key #3
Unlock
bitcoin

Locker-Ransomware Facts

1. Locker-Ransomware is occasionally distributed through a collaborative model called "ransomware-as-a-service" (RaaS), enabling less skilled individuals to launch attacks.

2. Attacks on critical services, such as healthcare and government institutions, highlight the real-world consequences and potential risks associated with Locker-Ransomware.

3. Locker-Ransomware employs advanced encryption techniques, making it difficult for victims to decrypt files without the decryption key.

A screenshot shows a Venus ransomware executable.

YOUR COMPUTER HAS BEEN LOCKED

Scotch Tape Locker v1.0

After scanning your computer, we detected that your
IP address was being used to visit
illegal websites containing child pornagraphy

After scanning your files and your IP, we detected
the following violations
(1) Downloading child porn
(2) Downloading elements of Zoophilia
(3) Commiting rape
(4) Sending spam messages containing
terrorist motives
(5) Promoting violence
(6) Torturing children
(7) Bestiality
(8) Eating children
(9) Sexual intercourse
with children
(10) Selling drugs

You will have to pay a fine of $150 in BitCoin.
Create a wallet on BlockChain.info
Purchase $150 in BitCoin
Send them to this BitCoin address

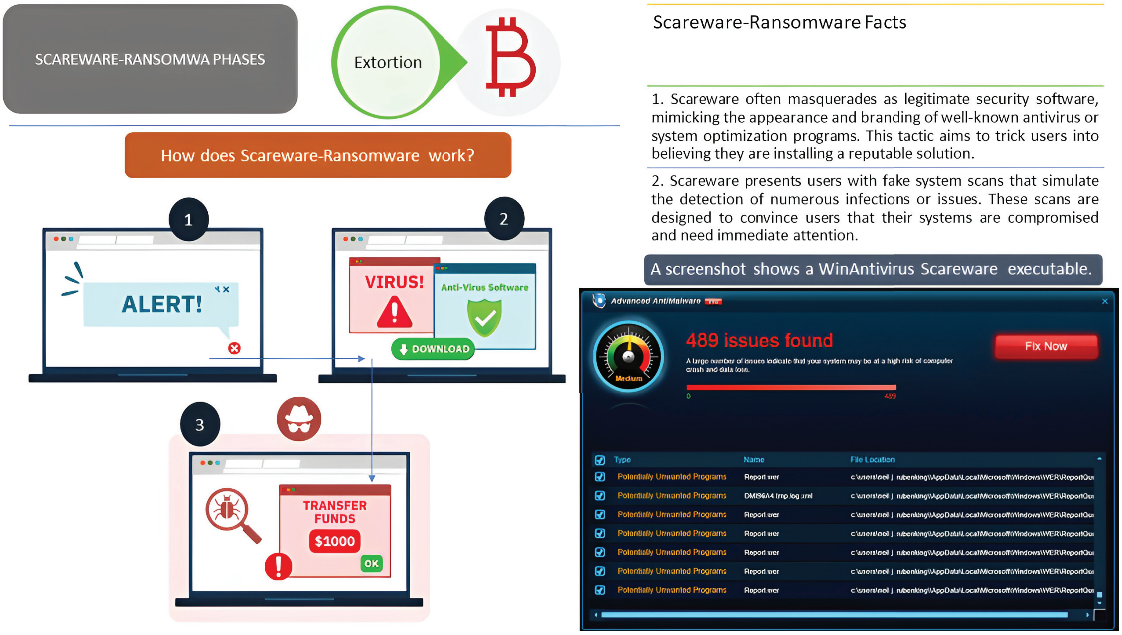
12mgfmmAWAckQ1HfUtDahE8yWkauayPydh

Once you have paid, email us at
fbifine@protonmail.com and give us your
ID. We will give you you unlock keys to
unlock your computer. Every ID has different keys.
Your ID: 97483-91676-07736

Figure 3. Locker ransomware [4].

Scareware preys on its victims by informing them that their machines have been hijacked and promising to eradicate the ransomware using a false antivirus program backed by the attacker. Numerous innocent consumers buy and install fake antivirus software due to scareware alerts' frequent appearance [5]. Human-operated malware and ransomware without data are different from ransomware. Cybercriminals also employ human-operated ransomware to break into networks or cloud infrastructure, carry out privilege escalation, and launch attacks on sensitive data. Instead of simply one system, the attack actively targets an entire organization. Attackers typically access a whole IT system, move laterally, and exploit flaws via improper security configurations. Ultimately, unauthorized access to privileged user credentials leads to ransomware assaults on IT systems that enable crucial corporate activities [3,4]. Figure 4 provides a visual representation of scareware, a form of malicious software that is becoming increasingly prevalent in cyberattacks [4].

However, ransomware without files uses a native and reliable system to launch attacks. It is difficult to identify the attack because no code needs to be placed on the victim's machine for it to work. As a result, anti-ransomware technologies do not find any suspicious files to trace during an attack. Depending on the attacker's intentions, file-based and human-operated ransomware can encrypt, lock, or leak data from files [2]. Ransomware poses a danger to businesses' technology and files. Until the ransom is paid, typically with Bitcoin, infected files or compromised devices are locked out of reach. The decryption key is frequently withheld even after a victim pays the ransom the hackers want. They periodically try to use the attacker's key to decrypt the data, which damages the system's stored files. Technology advancements such as ransomware development kits, ransomware-as-a-service, and bitcoins are to blame for the ongoing rise in ransomware attacks on desktop PCs, networks, and mobile devices [2]. Attacks using ransomware cost businesses and individuals hundreds of millions yearly [3]. New types of malware are continually being created thanks to the enormous cash benefits that hackers gain from ransomware assaults. Since 2013, numerous ransomware variants have appeared. Therefore, new, effective, and reliable techniques are needed to detect, prevent, and mitigate ransomware attacks. Different ransomware strains cannot be created using conventional antivirus software or other intrusion-detection systems. People and companies experience significant financial losses as a result of ransomware attacks. The encryption of files or devices until a ransom is paid can result in the permanent loss of important data, which can have severe consequences for individuals and businesses alike. Even after the ransom is paid, the decryption key is often withheld, causing additional damage to the system's stored files when attackers attempt to decrypt the data [1,6].



Scareware-Ransomware Facts

1. Scareware often masquerades as legitimate security software, mimicking the appearance and branding of well-known antivirus or system optimization programs. This tactic aims to trick users into believing they are installing a reputable solution.
2. Scareware presents users with fake system scans that simulate the detection of numerous infections or issues. These scans are designed to convince users that their systems are compromised and need immediate attention.

A screenshot shows a WinAntivirus Scareware executable.

489 issues found

A large number of issues indicate that your system may be at a high risk of computer crash and data loss.

☑ Type	Name	File Location
☑ Potentially Unwanted Programs	Report.wer	c:\users\neil j rubenking\AppData\Local\Microsoft\Windows\WER\ReportQu
☑ Potentially Unwanted Programs	DMI96A4.tmp.log.xml	c:\users\neil j rubenking\AppData\Local\Microsoft\Windows\WER\ReportQu
☑ Potentially Unwanted Programs	Report.wer	c:\users\neil j rubenking\AppData\Local\Microsoft\Windows\WER\ReportQu
☑ Potentially Unwanted Programs	Report.wer	c:\users\neil j rubenking\AppData\Local\Microsoft\Windows\WER\ReportQu
☑ Potentially Unwanted Programs	Report.wer	c:\users\neil j rubenking\AppData\Local\Microsoft\Windows\WER\ReportQu
☑ Potentially Unwanted Programs	Report.wer	c:\users\neil j rubenking\AppData\Local\Microsoft\Windows\WER\ReportQu
☑ Potentially Unwanted Programs	Report.wer	c:\users\neil j rubenking\AppData\Local\Microsoft\Windows\WER\ReportQu

Figure 4. Scareware ransomware [4].

3. Survey Planning

The present research involved several phases to achieve its overall objectives, including data collection and information gathering, data extraction and analysis, information synthesis, and reporting. A visual representation of the research process flow is presented in Figure 5, which depicts the activities involved in each phase and their interrelation.

The data collection process was carried out by selecting relevant and up-to-date journal and conference papers from reputable databases such as IEEE, Springer, MDPI, Elsevier, IET, and Archive.org, as well as other sources including university-based journals, theses/dissertations, and blogs published by reputable organizations such as Microsoft, Crowdstrike, Symantec, and Techspot. The collected materials were then categorized into two main groups: non-technical sources and technical sources. Non-technical sources contained general information on ransomware and were used to provide reliable information while writing the introduction and detailing the history of ransomware/chronology of attacks. Technical papers proposing solutions for ransomware attacks were divided into detection groups based on the nature and purpose of the proposed solution. Papers focusing on detection were further sub-categorized into artificial-intelligence-based methods and non-AI-based approaches. AI-based approaches were classified into machine learning methods, deep learning approaches, and artificial neural network approaches, while non-AI-based papers were grouped into packet and traffic analysis categories. The data extraction phase involved a detailed analysis and summary of each technical paper by identifying the problem it addressed, its objectives, the method/technique used, the achievements of the paper in terms of results obtained, and the research's limitations. Information synthesis was applied to identify similarities or relationships among papers in each group and to determine if and how the research improved upon or addressed the limitations of another work. The reporting phase placed papers that addressed similar problems or used similar techniques in the same group and presented their reviews in the same paragraph. This approach provided a good flow of communication and enhanced the readability of the paper, while also providing readers with a clear understanding of the concepts discussed in the research.

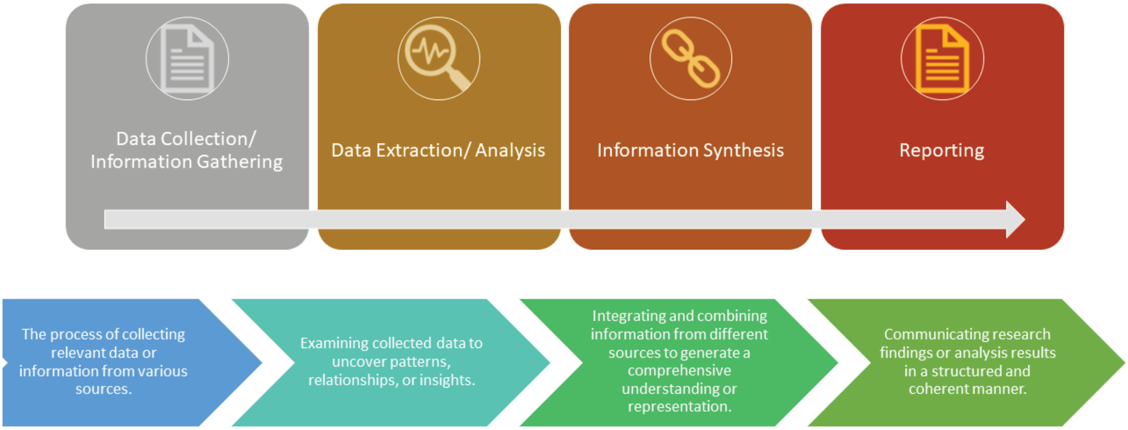

Data Collection/ Information Gathering
Data Extraction/ Analysis
Information Synthesis
Reporting
The process of collecting relevant data or information from various sources.
Examining collected data to uncover patterns, relationships, or insights.
Integrating and combining information from different sources to generate a comprehensive understanding or representation.
Communicating research findings or analysis results in a structured and coherent manner.

Figure 5. Research process flow.

4. Literature Review

Preventing ransomware is challenging for several reasons. The way ransomware functions is the same as benign software, which acts covertly. Ransomware detection in zero-day assaults is, therefore, crucial at this time. The primary objectives are to avoid ransomware-caused system damage, identify zero-day (previously unidentified) malware, and minimize detection, which means reducing the number of false positives while still detecting all instances of ransomware. False positives are instances where the system flags a harmless program or file as ransomware, leading to unnecessary alerts and actions. Ransomware can be found using a variety of tools and methodologies. Methods based on static analysis decompose source code without running it. They generate many false positives and cannot find ransomware that is disguised. Attackers frequently create new variations and modify their codes using various packaging techniques. To solve these issues, researchers use dynamic behavior analysis methods that monitor interactions between the executed code and a virtual environment. However, these detection methods are cumbersome and memory-intensive. Machine learning is ideal for analyzing any process or application's behavior.

Machine learning is considered ideal for analyzing the behavior of processes or applications because it can effectively learn patterns and anomalies in large datasets, which can be difficult for humans to detect. In the context of ransomware detection, machine learning algorithms can be trained on large datasets of both benign and malicious software to learn the behavioral characteristics that distinguish ransomware from legitimate software. This training can be used to identify new and previously unseen variants of ransomware, including zero-day attacks, based on their behavioral patterns.

Moreover, machine learning can be used to continuously learn and adapt to new threats, making it an effective approach to keep up with the constantly evolving tactics of ransomware attackers. Machine learning can also reduce false positives by accurately distinguishing between benign software and ransomware based on their behavioral patterns.

Compared with traditional signature-based detection and static analysis methods, machine learning is considered ideal because it can provide a more comprehensive and accurate analysis of the behavior of software, making it a powerful tool for ransomware detection. However, it is important to note that machine learning models need to be properly trained and validated to ensure their effectiveness and avoid biases or errors. The following are some machine-learning-based detection systems that follow highly traditional methodologies.

Table 1 summarizes previous studies on machine learning techniques (behavioral techniques) for ransomware detection from 2017 to 2022.

Table 1. Studies on machine learning techniques (behavioral techniques) for ransomware detection from 2017 to 2022.

Reference	Year	Author	Resolved the Issue	Utilized Technique	Result	Limitation
[7]	2017	Zahra and Sha	Detecting a ransomware attack using Cryptowall.	Blocklisting of command-and-control (C&C) servers.	The web proxy server, which acts as the TCP/IP traffic gateway, extracts the TCP/IP header.	The model's efficacy and precision in identifying ransomware and its attack techniques against various operating system environments were not demonstrated through implementation.
[8]	2018	Shaukat and Ribeiro	Detection of ransomware.	RansomWall, a layered and hybrid mechanism.	Effective at identifying zero-day attacks.	N/A
[9]	2019	Makinde et al.	To determine whether an actual network system is vulnerable to a ransomware assault.	Learning machines.	Correlation greater than 0.8.	It imitated the behavior of a small group of users.
[10]	2019	Ahmad et al.	Differentiating Locky ransomware users.	Utilizing parallel classifiers, a behavioral approach to ransomware detection.	Highly reliable detection with a low proportion of false positives.	N/A
[11]	2022	Singh et al.	Discovery of new ransomware families and classification of newly discovered ransomware assaults.	Checks process memory access privileges to enable rapid and accurate malware detection.	Between 81.38% and 96.28% accuracy.	N/A

An application's normal behavior is assessed from a user and resource perspective. A baseline for normal behavior is established based on what is thought to be the typical or routine operation of a computer system or network. Indicators of usual activity include logins, file access, user and file behaviors, resource utilization, and other significant indicators [1].

The length of the learning process is determined by the amount of data needed to build a baseline to represent typical system behavior. The tool investigates behavioral outliers from the baseline's depiction of the typical behavioral pattern. A ransomware-detection and -prevention model was created for unstructured datasets derived from Ecuadorian Control and Regulatory Institution (EcuCERT) logs [12].

The methodology uses musing to spot peculiar behavioral patterns connected to Windows malware. Feature selection is applied to the Log data to extract the most beneficial and discriminating information that indicate a ransomware attack. The extracted data represent that autonomous learning algorithms in ransomware are swiftly and precisely identified using the input feature set and algorithms that mimic abnormal behavioral patterns. Code obfuscation tools and new polymorphic variants have been developed as signature additions in identifying ransomware attacks, which are constantly evolving [8].

Since generic malware attack vectors cannot effectively capture the particular behavioral traits of cryptographic ransomware, they are insufficient or inaccurate for ransomware detection. The suggested approach, RansomWall, is a hybrid system that uses static and dynamic analytics to present a research set of properties that mimic ransomware activity. The technique allows for early ransomware detection while utilizing a strong trap layer to detect zero-day attacks. RansomWall with the Gradient Tree Boosting Algorithm demonstrated a detection rate of 98.25% and an incredibly low (almost nil) false-positive rate when tested against 574 samples of 12 cryptographic ransomware running on the Microsoft Windows operating system. It also had a detection rate of less than 10% for 30 zero-day attack samples compared with 60 VirusTotal security engines. One version of behavioral detection methodologies uses a machine learning baseline model for simulating and forecasting the specific network user behavior pattern at the micro level to identify potential scenarios that could indicate a vulnerability or a true ransomware assault [9].

The goal was to find a simple network system's vulnerability to a ransomware attack. Comparing the outcomes from the simulated network and the log data from the server in the existing network system revealed a realistic model with a correlation above 0.8. This method's drawback was that it only adequately captured the activity of a small percentage of users. Future studies should focus on mimicking user behavior over a large user base using big data analytics tools. A more recent method of behavioral ransomware detection used two parallel classifiers [10].

To distinguish between the several Locky ransomware variants, one technique focused on early detection based on the behavioral analysis of ransomware network traffic to prevent ransomware from connecting to command-and-control servers and carrying out damaging payloads. The study employed a dedicated network to collect information and extract important details from network traffic. Using data at the packet and datagram levels, two different (parallel) classifiers were used to analyze the extracted properties of the Locky ransomware family. The results of the studies show that the technology has a high level of success in detecting ransomware activities on the network. Furthermore, it permits an extreme lexicon with a low percentage of false positives. Using command-and-control (C&C), the server blocklists ransomware attacks as the means of communication and conducts behavioral analysis of the ransomware in an IoT environment [7].

A domain-specific strategy for identifying Cryptowall ransomware attacks is provided. The operation obtains the TCP/IP header from the web proxy server, which serves as the TCP/IP traffic gateway. Furthermore, it retrieves source and destination IPs and compares them to the IPs of forbidden command-and-control servers. Ransomware is identified if the source or destination IPs match an attack targeting Internet of Things devices. However, the model was not used to demonstrate how well it could spot ransomware and its attack vectors against different operating system environments. Using a very recent technique of behavioral-based detection that uses access privileges in process memory, ransomware may now be quickly and accurately detected [11,13].

It is possible to categorize new ransomware attacks and find malware families that have not yet been recognized by looking at a file or application's access privileges and the area of memory it intends to access. Examining the behavior and ascertaining the purposes of lawful files and applications before executing them is beneficial. The experimental results employing these several approaches show good detection accuracy, ranging from 81.38% to 96.28%.

Table 2 summarizes previous studies on machine learning techniques (static and dynamic analysis) for ransomware detection from 2017 to 2022.

Table 2. Studies on machine learning techniques (static and dynamic analysis) for ransomware detection from 2017 to 2022.

Reference	Year	Author	Problem Addressed	Method Used	Result
[14]	2017	Rahman and Hasan	Enhanced ransomware-detection method.	Using support vector machines as an analysis tool.	Better ransomware detection is achieved with an integrated approach than static or dynamic analysis used separately.
[13]	2018	Dehghantanha et al.	Windows ransomware detection that is quick and accurate.	Netconverse (classifier using j48 decision tree).	97.1% actual-positive detection rate.
[15]	2019	Jasmin	Separating ransomware traffic and regular traffic.	Algorithms used in logistic regression include random forest and support vector machine.	The best detection rate is 99.9% for the random forest, with 0% false positives.
[16]	2019	Ameer	Detection of ransomware.	Analyses that are static and dynamic.	100% detection and classification precision.
[17]	2020	Khammas	Detection of ransomware.	Random forest method.	97.74% of samples are detected.
[18]	2020	Hwang et al.	An improved method of detecting ransomware.	Random forest and Markov models.	97.3% overall accuracy, 4.8% for false positives, and 1.5% for false negatives.
[19]	2022	Talabani and Abdulhadi	Tools for detecting ransomware that involve data mining and machine learning approaches have poor accuracy.	Decision Table and PARTially Decided Decision Tree.	Recall (96%), accuracy (96.01%), F-measure (95.6%), and precision (95.9%).

Several improved machine learning approaches have been applied for accurate and efficient ransomware detection. These methods are meant to address the drawbacks of the current ML-based ransomware-detection tools. One of these advancements regards the challenges detection systems (such as sandbox analysis and pipelines) face in isolating a sample and handling the wait time for isolated ransomware samples to be evaluated [20].

The approach predicts ransomware using a dataset containing 30,000 attributes as independent variables. Five qualities that were obtained through feature selection were used in the support vector machine technique. The approach provides a respectable 88.2% accuracy rate in ransomware detection. To reduce the number of false positives, this hybrid technique combines the “guilt by association” hypothesis with content-, metadata-, and behavior-based analysis. Giving the user control over recovery is necessary, and file versioning in cloud storage is used to halt the process. The only duty of the end user is to keep track of the recovery. Users are given classification information so they may make educated decisions and prevent false positives. The method results in more-accurate detection and reliable recovery. An innovative method for detecting network-level ransomware uses machine learning, certificate information, and network connection information [21].

This technique can be used with system-level monitoring to detect ransomware outbreaks early. This method uses connection-, encryption-, and certificate-based network traffic characteristics to extract and model ransomware features. It is a feature model that uses support vector machines, logistic regression, and random forest to distinguish ransomware traffic. According to experimental findings on various datasets, random forest has the best detection rate of 99.9% and the lowest rate of false positives. Another more-effective detection method is a decision tree model based on big data technology that uses Argus for packet preprocessing, combining, and malware file identification [21].

The flow replaced the packet data, resulting in a 1000-fold (1000:1) reduction in data size. Feature selection and concatenation were used to extract and aggregate the attributes of the actual network traffic. In order to improve classification accuracy, the technique made use of six feature selection techniques. Machine learning has recently been creatively applied to monitor Android device power usage as a ransomware-detection technique [13].

The suggested method measures how much energy particular Android processes use to distinguish ransomware from valuable programs. Data on the ransomware's unique local energy fingerprint are gathered and analyzed to accomplish this. According to experimental findings, the approach offers high detection and precision rates of 95.6% and 89%, respectively. Additionally, it outperforms k-nearest neighbor, neural network, support vector machine, and random forest regarding the accuracy, recall rate, precision rate, and F-measure.

Another superior option is the cutting-edge, portable RanDroid approach for automatically detecting polymorphic ransomware [22]. The RanDroid approach uses both static and dynamic analyses to detect polymorphic ransomware. The method compares the structural similarity of pieces obtained from an application with a collection of threat information from well-known ransomware variants to detect new ransomware variants on Android devices. Image similarity measurements (ISMs) and string similarity measurements (SSMs) are the two similarity measures used. Using language analysis, the app's behavioral attributes and picture textural strings are mined for additional information. The strategy reduces ransomware threats without changing the Android OS or its underlying security module while addressing the constraints of static analysis. The methodology can detect ransomware using evasive tactics such as complex codes or dynamic payloads, according to an analysis of the method based on 950 malware samples. According to a related study, a strategy combining static and dynamic analysis can help identify and separate Android ransomware from other malware [16].

We looked at network-based features, text, and permissions using static analysis. Furthermore, dynamic analysis was performed on the system call, CPU, and memory logs. The strategy's effectiveness in reducing evasive ransomware assaults is demonstrated by experiments using traits from malicious and benign samples. Additionally, it is 100 percent accurate at classifying and identifying unknown ransomware.

5. Evolution of Ransomware

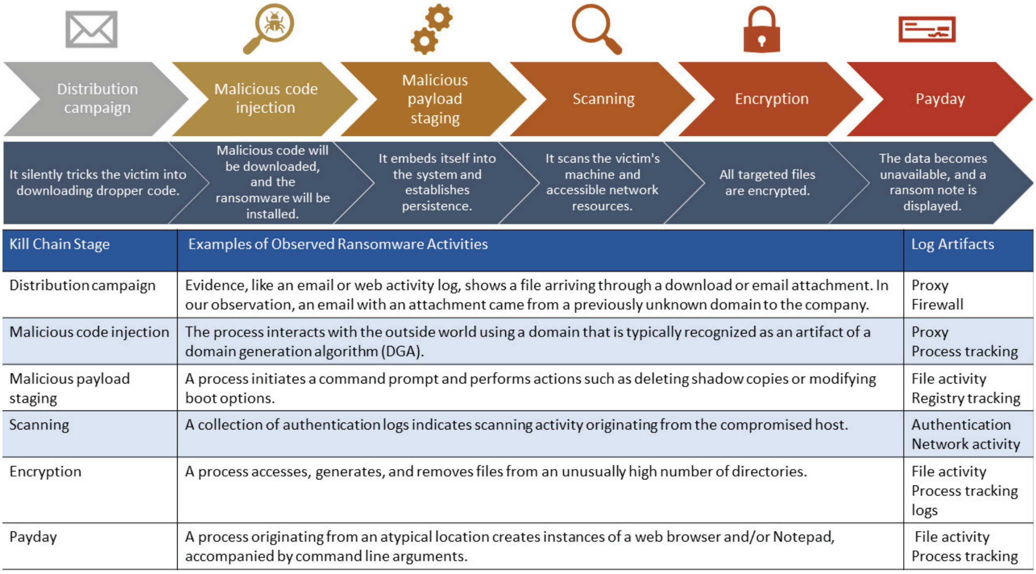
Ransomware attacks have been around since the late 1980s; Joseph Popp showcased the first instance of ransomware. This attack utilized symmetric-key encryption to take control of victims' hard drives and request a ransom. The flaw in this system was that the same key was used for encryption and decryption, making it vulnerable. As a result, it was possible to research the AIDS ransomware (also known as PC Cyborg) to find the decryption key and create a solution for the malware's encryption. Ransomware attacks have continued evolving and have become more sophisticated in recent years, making them a significant threat to individuals and organizations [23]. A brief timeline of various potent ransomware attacks is shown in Table 3. The table, an excerpt from a timeline of the most significant ransomware attacks from 2012 to 2023, contains essential information on the evolution of ransomware based on the year the ransomware first appeared, its name, and its primary description [2,3,23].

Table 3. Brief chronology of major ransomware attacks from 2012 to 2022.

Reference	Year	Name of the Ransomware	Description
[4]	1989	AIDS Trojan	The first known ransomware attack, the AIDS Trojan, was distributed on floppy disks and demanded a payment of USD 189 to unlock infected files.
[5]	2012	Reveton	Ransomware that posed as law enforcement and demanded payment for supposed illegal activities.
[23]	2013	CryptoLocker	One of the first widespread ransomware attacks that used encryption to lock victims' files.
[24]	2014	CryptoWall	A variant of CryptoLocker that caused millions of dollars in damages.
[3]	2015	TeslaCrypt	A ransomware strain that targeted gamers and encrypted game-related files.
[25]	2016	Locky	Ransomware that was spread through malicious email attachments.
[3]	2017	WannaCry	A ransomware attack affecting over 200,000 systems across 150 different countries.
[26]	2018	SamSam	A ransomware attack that targeted hospitals, municipalities, and other organizations.
[3]	2019	Ryuk	A ransomware attack that caused significant damage to several companies and organizations.
[27]	2020	Maze	A ransomware attack that encrypted victims' files and threatened to leak sensitive data if the ransom was not paid.
[3]	2021	REvil/Sodinokibi	A ransomware attack that targeted Kaseya, a software company, and affected over 1500 businesses worldwide.
[28]	2022	Royal Ransomware	A ransomware attack that encrypted victims and demanded a ransom payment in order to decrypt them, targeting businesses, governments, and healthcare organizations, with victims mostly from the United States.
[28]	2023	LockBit Ransomware	A ransomware attack that encrypts the files and demands payment in exchange for the decryption key, often in conjunction with phishing emails or other social engineering techniques.

Ransomware has become a popular tool for cybercriminals to extort money from individuals and organizations. As technology advances, preventing such attacks is more challenging. It is essential to remain vigilant and take appropriate measures to protect against these threats, such as keeping software up-to-date and regularly backing up important data [5]. There are six levels, which can be summarized as follows, as adapted from [29] and shown in Figure 6.

1. Distribution campaign: The attacker silently induces the victim to download the infection-starting dropper code. The attacker uses methods including email phishing, social engineering, and others.
2. Malicious code injection: During this phase, the target's computer is infected with ransomware, and malicious code is downloaded.
3. Malicious payload staging: Ransomware sets up persistence by inserting the system.
4. Scan checks for encryption on the target computer and any network-accessible resources.
5. Encryption: The process of encrypting all of the selected documents begins.
6. Payday: Victims cannot access their data, and a notification seeking payment is visible on the screen of the targeted device.


Distribution campaign
Malicious code injection
Malicious payload staging
Scanning
Encryption
Payday
It silently tricks the victim into downloading dropper code.
Malicious code will be downloaded, and the ransomware will be installed.
It embeds itself into the system and establishes persistence.
It scans the victim's machine and accessible network resources.
All targeted files are encrypted.
The data becomes unavailable, and a ransom note is displayed.

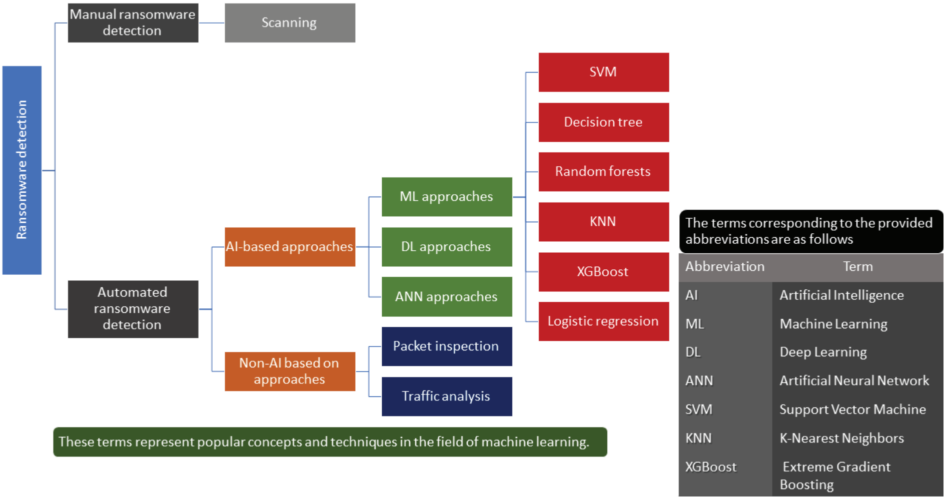
Kill Chain Stage	Examples of Observed Ransomware Activities	Log Artifacts
Distribution campaign	Evidence, like an email or web activity log, shows a file arriving through a download or email attachment. In our observation, an email with an attachment came from a previously unknown domain to the company.	Proxy Firewall
Malicious code injection	The process interacts with the outside world using a domain that is typically recognized as an artifact of a domain generation algorithm (DGA).	Proxy Process tracking
Malicious payload staging	A process initiates a command prompt and performs actions such as deleting shadow copies or modifying boot options.	File activity Registry tracking
Scanning	A collection of authentication logs indicates scanning activity originating from the compromised host.	Authentication Network activity
Encryption	A process accesses, generates, and removes files from an unusually high number of directories.	File activity Process tracking logs
Payday	A process originating from an atypical location creates instances of a web browser and/or Notepad, accompanied by command line arguments.	File activity Process tracking

Figure 6. Six levels of ransomware attacks [29].

6. Ransomware Detection

6.1. Ransomware-Detection Methods

The two main types of ransomware-detection methods are automated and manual. Employing technologies to identify and report ransomware attacks is a prerequisite for automated methods. These tools are typically software programs that have the potential to be able to stop attacks. Techniques for manual detection focus on routinely scanning data and devices for indicators of attacks. Checking to see if a malware attack has not modified data or stopped authorized users from accessing their devices or files includes looking at any changes to file extensions, the accessibility of devices and files by authorized users, and any changes to file extensions. The flow of the presentation in this section is illustrated in Figure 7.


Ransomware detection
Manual ransomware detection
Scanning
Automated ransomware detection
AI-based approaches
Non-AI based on approaches
ML approaches
DL approaches
ANN approaches
Packet inspection
Traffic analysis
SVM
Decision tree
Random forests
KNN
XGBoost
Logistic regression
These terms represent popular concepts and techniques in the field of machine learning.
The terms corresponding to the provided abbreviations are as follows

Abbreviation	Term
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
ANN	Artificial Neural Network
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
XGBoost	Extreme Gradient Boosting

Figure 7. Ransomware detection taxonomy.

6.1.1. Manual Ransomware Detection

Manual ransomware detection refers to the process of detecting ransomware through human analysis and intervention rather than automated systems. This approach involves analyzing system logs, network traffic, and other indicators of compromise to identify patterns and behaviors associated with ransomware attacks. While manual detection can be time-consuming and resource-intensive, it can be an effective complement to automated detection methods, as it can help identify new or unknown types of ransomware that may not be detected by automated systems [30].

Despite its effectiveness, manual ransomware detection has some limitations. It can be labor-intensive and requires highly trained personnel to analyze system logs and network traffic. Additionally, manual detection may not scale well in large organizations or networks, where automated detection methods may be more efficient [30].

Scanning

Manual ransomware-detection scanning is a technique used to detect ransomware through the manual analysis of files and systems. This approach involves scanning individual files or systems for signs of ransomware activity, such as encrypted files or abnormal network traffic. Manual scanning can be a complementary approach to automated scanning methods, as it can help detect new or unknown types of ransomware that may not be detected by automated systems [30].

While manual ransomware-detection scanning can be effective, it has some limitations. It can be time-consuming and labor-intensive, especially when scanning large networks or systems. Additionally, manual scanning may generate false positives, which can be disruptive to normal system operations [30].

6.1.2. Automated Ransomware Detection

The current methods for detecting ransomware primarily involve monitoring the system at the file system level. Automated approaches to detecting ransomware can be categorized into two main groups: those based on artificial intelligence (AI) and those that are not based on AI. AI-based methods typically employ machine learning (ML), deep learning (DL), and artificial neural network (ANN) techniques to detect ransomware. Some tools utilize variations of these techniques or a hybrid approach that combines two or more techniques to combat the threat of ransomware attacks. Non-AI methods rely on packet inspection and traffic analysis to detect ransomware. One of the major advantages of automated approaches is their ability to detect, block, and recover from ransomware attacks without human intervention. Additionally, these tools are highly accurate and reliable in terms of detecting, preventing, and recovering from ransomware attacks [31].

Artificial-Intelligence-Based Approaches

Artificial intelligence (AI) techniques, including machine learning, deep learning, and artificial neural networks, have been utilized for automated ransomware detection. These techniques involve the use of behavioral techniques, as well as static and dynamic analysis, to identify and prevent ransomware attacks. Machine learning algorithms can learn from previous ransomware attacks and detect new variants by analyzing patterns and behaviors. On the other hand, deep learning methods can leverage neural networks to detect ransomware attacks by analyzing large amounts of data. Artificial neural networks can also be used to identify ransomware by processing and analyzing multiple data sources. These AI-based approaches offer a more efficient and reliable way to detect and prevent ransomware attacks, reducing the potential impact on businesses and individuals [31]. AI-based approaches include the following:

1. Machine Learning Approaches

Machine-learning-based detection is a more advanced approach that relies on training a machine learning model to detect ransomware based on its behavior patterns or features. This approach is based on collecting a large dataset of benign and malicious samples, extracting relevant features from them, and then training a machine learning model to classify new samples as peaceful or hostile based on their characteristics [32,33].

Machine-learning-based detection has several benefits, including its ability to detect new or unknown ransomware variants that do not match existing signatures or patterns and to adapt to changing ransomware behavior patterns over time. Moreover, this approach is less prone to false positives than signature-based and heuristic-based detection, as it relies on detecting actual behavior patterns rather than static code signatures or predefined rules. However, machine-learning-based detection is limited by its reliance on a large and representative dataset of training samples and by its susceptibility to adversarial attacks that can manipulate the features or behavior of the ransomware to evade detection [31].

a. Machine Learning Algorithms for Ransomware Detection

A particular kind of artificial intelligence known as machine learning enables computer systems to improve their performance on a given job without being explicitly taught. Malicious ransomware malware encrypts a victim's files and demands payment for the decryption key. Due to their rising prevalence and severity, machine learning techniques are increasingly needed to identify and stop ransomware attacks. Table 4 lists the machine learning algorithms that are employed. Support vector machines, decision trees, random forests, k-nearest neighbors, XGBoost, and logistic regression are just a few machine learning approaches that can detect ransomware. Each method has advantages and disadvantages, and the best approach depends on the situation and the data [1,6].

Table 4. Machine learning algorithms.

References	Algorithm	Characteristics
[17,34]	Decision tree	Decision trees can be trained on features such as file modifications, network traffic, and system calls to distinguish between ransomware and benign software behavior. The resulting decision tree can then be used to determine whether new data contain ransomware.
[17,34]	Random forest	In order to guarantee that each tree in the forest has the same distribution and is dependent on the values of a randomly selected random vector, this strategy uses an ensemble method that combines tree predictors. Performance may be enhanced in comparison to standalone decision trees. Using a network of decision trees, the random forest approach is used to select and forecast the input data type.
[14,35]	Support vector machine	Support vector machines can be trained on features such as system calls, network traffic, and file behavior to distinguish between ransomware and benign software behavior. After that, it is possible to determine whether new data constitute ransomware using the resultant support vector machines. Support vector machines are handy when the data are high-dimensional and non-linearly separable, as is often the case in ransomware detection.
[36,37]	k-nearest neighbor	k-nearest neighbor is a popular machine learning algorithm used in various research fields. It is a non-parametric approach that can be used for both classification and regression tasks. KNN is known for its simplicity, but is also computationally expensive, with simplified and concise hyperparameters.
[38]	XGBoost	Extreme gradient boosting is a powerful machine learning algorithm that has gained widespread popularity in research. It is an ensemble method that combines multiple decision trees to improve the accuracy of the model. XGBoost is known for its scalability, speed, and ability to handle complex datasets.
[39]	Logistic regression	Logistic regression is a widely used machine learning algorithm in various research fields. It is a linear model that can be used for binary classification tasks. Logistic regression is known for its simplicity, interpretability, and ability to handle small datasets.

Decision trees are a simple and intuitive machine learning algorithm that can be used for classification tasks, including ransomware detection. Decision trees work by recursively partitioning the data into subsets based on the values of the features and creating a tree-like structure representing the decision-making process. Both categorical and continuous components can be handled by decision trees, which are simple to interpret but susceptible to overfitting and sensitive to minute changes in the data [13,31,34].

Random forests are an extension of decision trees that improve performance and reduce overfitting. By randomly selecting features and data, random forests create multiple decision trees and combine their predictions. They are better-equipped to handle high-dimensional data and are less likely to overfit. However, they can be computationally demanding and difficult to interpret [17].

Support vector machines are reliable machine learning techniques that can be utilized for ransomware detection and classification and regression applications. Support vector machines operate by identifying the hyperplane that divides the data into distinct classes according to the values of the features as thoroughly as possible. Support vector machines can effectively handle high-dimensional data. They can accept both linear and nonlinear borders, but the choice of the kernel function and its parameters may impact them [14].

k-NN is a non-parametric algorithm used for classification and regression tasks. It works by finding the k closest data points in the training set to a given input, and then predicting the label of the input based on the most common label among those k neighbors. It is a simple but effective algorithm that can be used in a wide range of applications [36,37]

XGBoost (short for "Extreme Gradient Boosting") is a powerful machine learning algorithm that is especially popular for gradient boosting tasks. It uses a combination of decision trees and gradient boosting to create a highly accurate model that can handle large datasets and complex feature interactions. XGBoost has become widely used in the industry [38].

Logistic regression is a parametric algorithm used for binary classification tasks (i.e., where the output is one of two possible classes). It works by modeling the probability of the output class as a function of the input features. The algorithm is trained to find the optimal parameters that maximize the likelihood of the training data and can be regularized to prevent overfitting [39].

The choice of a machine learning algorithm for ransomware detection depends on the specific problem and data available. Decision trees, random forests, support vector machines, and neural networks are all effective options, and researchers have successfully used each of these algorithms for ransomware detection in different contexts [5,31].

2. Deep Learning Approaches

Deep learning techniques have been proposed as a solution to address the limitations of traditional supervised ransomware-detection tools to enhance the accuracy and reliability of ransomware detection. These algorithms utilize automatic feature generation and are well-suited to handle unstructured datasets, requiring minimal or no human intervention due to their self-learning capabilities. Their effectiveness in classifying audio, text, and image data makes them particularly useful in detecting textual and image-based ransomware data. However, training deep learning algorithms demand a considerable amount of data, which may render them unsuitable for general-purpose applications, particularly those involving small datasets or sizes. Other challenges associated with deep learning include the need for high processing power and difficulty with adapting to real-world datasets [6,40].

3. Artificial Neural Network Approaches

Artificial neural network approaches are well-suited for detecting various types and variants of ransomware data, including text and image ransomware variants, due to their wide range of applications. Neural networks are an excellent choice for adapting to new ransomware data and identifying zero-day attacks because of their ability to continuously learn. The versatility of neural networks makes them highly effective in detecting different forms of ransomware data and adapting to new threats. However, these techniques are dependent on hardware and can be vulnerable to data dependencies, as well as the black-box nature of the technology, which limits the ability of human analysts to monitor data processing and identify deviations in the process [5,6,41].

Non-Artificial-Intelligence-Based Methods

Non-AI techniques such as packet inspection and traffic analysis can be utilized to detect ransomware. Anomaly detection is one effective algorithm used to detect ransomware. These algorithms analyze network traffic and identify patterns that deviate from normal behavior. Unusual patterns of network traffic, such as a sudden increase in file encryption activity or a large number of outbound network connections to suspicious IP addresses, are indications of ransomware activity. By comparing network traffic to a baseline of normal behavior, anomaly-detection algorithms can quickly identify and alert security teams to potential ransomware attacks [2].

Other non-AI techniques include signature-based detection, which involves comparing network traffic to known ransomware signatures, and behavior-based detection, which looks for patterns of behavior consistent with known ransomware attacks [2].

Another approach involves the use of honeypots to monitor network activity and detect the presence of ransomware. This method entails the establishment of a honeypot folder and observing any changes that may indicate the presence of ransomware. The early detection of ransomware is critical in mitigating its impact and preventing further damage [2].

It is important to note that these detection techniques are not foolproof and should be used in conjunction with other security measures such as user education, regular backups, and security patches [2].

Antivirus software is an example of a non-AI-based approach for detecting and preventing malware, including ransomware. It typically uses a combination of signature-based detection and behavior-based detection to identify and block malicious software. Signature-based detection involves comparing files against a database of known malware signatures, while behavior-based detection looks for patterns of behavior that are indicative of malware activity. While antivirus software has been an effective tool for detecting and preventing malware, it has some limitations. For example, signature-based detection is only effective against known malware signatures, meaning that new or unknown forms of malware can bypass this detection method. Additionally, some types of malware can be designed to evade behavior-based detection methods [42].

In recent years, AI-based approaches, such as machine learning and deep learning, have been introduced to enhance the accuracy and effectiveness of malware detection. However, antivirus software continues to be a critical component of cybersecurity, particularly for organizations with limited resources or expertise in AI-based techniques. By using a combination of signature-based and behavior-based detection, antivirus software can provide an effective defense against known and unknown forms of malware, including ransomware [42].

1. Packet Inspection

Packet inspection refers to examining individual data packets' contents as they move through a network. This technique can be used to detect the presence of malware by identifying packets that contain suspicious data or have characteristics that are inconsistent with normal network traffic. For example, packets containing large amounts of encrypted data or sent from suspicious IP addresses may indicate ransomware activity [43,44].

2. Traffic Analysis

Traffic analysis, on the other hand, involves the examination of patterns of network traffic over a period of time. This technique can be used to detect ransomware by identifying patterns of behavior that are consistent with known ransomware attacks. For example, traffic analysis may reveal a sudden increase in network traffic during off-hours or a large number of outbound network connections to suspicious IP addresses. Packet inspection and traffic analysis are two important techniques used in detecting malicious software, including ransomware. These techniques involve the examination of network traffic to identify potentially harmful data packets and patterns of behavior that may indicate the presence of malware. By examining network traffic and identifying patterns of behavior indicative of malicious activity, these techniques can help organizations detect ransomware attacks and protect their critical data and systems [45,46].

Packet inspection and traffic analysis are two essential techniques for detecting ransomware and other forms of malware. By examining network traffic and identifying behavior indicative of malicious activity, these techniques can help organizations detect ransomware attacks and protect their critical data and systems. They should be used alongside other security measures, such as regular backups and security patches, as they are not completely infallible. Furthermore, these techniques necessitate specialized tools and expertise, which can pose a challenge for organizations without dedicated cybersecurity resources [43–46].

6.2. Ransomware-Detection Techniques

Ransomware detection is a critical component of cybersecurity, and various techniques have been developed to detect ransomware attacks. This section will discuss different ransomware-detection techniques proposed in the literature and their strengths, weaknesses, and limitations.

6.2.1. Signature-Based Detection

Signature-based detection is a traditional approach that relies on identifying known ransomware signatures or patterns in the code or behavior of the malware. This approach is based on creating a database of known ransomware signatures or marks and scanning the system or network for matching signatures or patterns. If a match is found, the ransomware is flagged as malicious and appropriate actions are taken [32,33].

One benefit of signature-based detection is its simplicity and effectiveness in detecting known ransomware variants. However, this approach is limited by its inability to detect new or unknown ransomware variants that do not match existing signatures or patterns. Moreover, attackers can easily evade signature-based detection by modifying the code or behavior of the ransomware to avoid detection [31].

6.2.2. Heuristic-Based Detection

Heuristic-based detection is a more advanced approach that identifies ransomware behavior patterns or anomalies indicative of malicious activity. This approach is based on creating rules or heuristics that describe typical ransomware behavior and then monitoring the system or network for any deviations or anomalies from these rules. If such variations or abnormalities are detected, the ransomware is flagged as suspicious or malicious, and appropriate actions are taken [32,33].

One of the advantages of heuristic-based detection is its ability to detect new or unknown ransomware variants that do not match any existing signatures or patterns. Moreover, this approach is less prone to false positives than signature-based detection, as it relies on detecting actual behavior patterns rather than static code signatures. However, heuristic-based detection is limited by its reliance on predefined rules or heuristics, which may only capture some possible ransomware behavior patterns or anomalies. Moreover, attackers can easily evade heuristic-based detection by modifying the behavior of the ransomware to avoid detection [31].

6.2.3. Network-Based Detection

Network-based detection is an approach that relies on monitoring the network traffic for suspicious or malicious activity that may be indicative of a ransomware attack. This approach is based on analyzing the network traffic for anomalies or patterns characteristic of ransomware, such as large volumes of outbound traffic, unusual network connections, or network traffic encryption [32,33].

One of the advantages of network-based detection is its ability to detect ransomware activity even if the malware has not yet infected the system or if the ransomware is using non-standard encryption methods. Moreover, this approach is less prone to false positives than other detection approaches, as it relies on detecting actual network traffic patterns rather than static code signatures or predefined rules. However, network-based detection is limited by its reliance on network traffic analysis tools that may not be available or may not capture all ransomware activity. Moreover, attackers can easily evade network-based detection by encrypting their network traffic or using stealthy communication channels [31].

6.2.4. Hybrid Detection

Hybrid detection is an approach that combines different ransomware-detection techniques to improve the overall detection accuracy and speed. This approach combines the strengths of other detection techniques, such as signature-based, heuristic-based, machine-learning-based, and network-based detection, to create a more robust and effective detection system [32,33].

One of the advantages of hybrid detection is its ability to overcome the limitations of individual detection approaches and to improve the overall detection accuracy and speed. Moreover, this approach is less prone to false positives and negatives than unique detection approaches, as it combines different sources of information and analysis. However, hybrid detection is limited by its complexity and resource requirements, as it requires integrating and coordinating other detection systems and tools [31].

6.3. Feature Extraction and Selection

Machine learning techniques have been increasingly used to detect ransomware due to their ability to learn behavior patterns and detect anomalies. In this section, we will discuss different features used for ransomware detection using machine learning and the techniques used for feature selection, such as principal component analysis and correlation analysis [18,47].

6.3.1. Features Used for Ransomware Detection

There are several features that can be used for ransomware detection, with the most common ones including the following:

1. File access patterns are a common feature used to detect ransomware. Ransomware often accesses and encrypts files in a specific pattern, such as alphabetical order, extension type, or creation date. This behavior can be detected using file access patterns as features. For example, analysis of file access patterns may reveal that a large number of files are being accessed and modified in a short period of time, indicating a potential ransomware attack [48].
2. System calls are another feature commonly used for ransomware detection. Ransomware frequently uses system calls to perform malicious activities, such as reading and writing files, creating processes, and network communication. System-call traces can be extracted and used as features for detection. For example, analysis of system-call traces may reveal that a process is making an unusually high number of system calls, which could indicate ransomware activity [34].
3. Network traffic analysis is a valuable feature for detecting ransomware. Typically, ransomware uses a command-and-control (C&C) server to deliver and receive orders. Analysis of network traffic can provide valuable features for detecting ransomware. For example, analysis of network traffic may reveal that a large amount of data are being sent to an unusual IP address, which could indicate that the system is infected with ransomware [49].
4. Behavioral analysis is another approach to ransomware detection. This involves monitoring the behavior of running processes and identifying anomalies that indicate malicious activity. Features such as process creation, termination, and file access can be used for this type of analysis. For example, the analysis of process creation and termination events may reveal that a process is spawning multiple child processes, which could indicate ransomware activity [1].

5. Static analysis is the examination of the executable file's source code to spot malicious activity. Features such as code size, entropy, and string patterns can be used for this purpose. For example, analysis of code size and entropy may reveal that a file contains obfuscated code, which could indicate ransomware activity [32]. Behavioral analysis and dynamic analysis are similar in that they both involve the monitoring of running processes to identify malicious activity. However, there are some key differences between the two approaches.

Behavioral analysis involves monitoring the behavior of running processes on a system to identify anomalies that indicate malicious activity. This is typically carried out in real-time, allowing the detection of ransomware as it is executed on a system. In contrast, dynamic analysis involves running an executable file in a controlled environment, such as a sandbox, to observe its behavior and identify any malicious activity. This is typically conducted prior to deploying the executable file on a production system.

The confusion between static and dynamic analysis may arise from the fact that both approaches involve the analysis of executable files, but they do so in different ways. Static analysis involves looking at the executable file's source code to spot malicious activity, while dynamic analysis involves running the executable file in a controlled environment to observe its behavior.

Dynamic analysis can be performed in real-time, but it can also be conducted in a sandbox environment before deploying the executable file on a production system. In a sandbox environment, the executable file is executed in a controlled environment, allowing its behavior to be monitored and analyzed without affecting the production system. Once the analysis is complete, the results can be used to determine whether the executable file is malicious or benign.

In the case of ransomware, real-time behavioral analysis is typically the preferred approach for detecting and responding to attacks. However, dynamic analysis can also be useful for identifying new and previously unseen variants of ransomware, which can then be used to improve the effectiveness of real-time behavioral analysis.

By using these features, machine-learning-based ransomware-detection methods can achieve high detection rates and low false-positive rates.

6.3.2. Feature Selection Techniques

- Principal component analysis: This technique is used to reduce the dimensionality of a dataset by identifying the most critical features that explain the majority of the variance in the data. Principal component analysis can help identify redundant or irrelevant features and select the most informative ones for ransomware detection [50].
- Correlation analysis: Correlation analysis is a technique used to identify the correlation between features in a dataset. Highly correlated features may be redundant and can be removed to simplify the model and improve performance [27].

6.4. Performance Evaluation of Machine Learning Models for Ransomware Detection

Evaluating the performance of machine learning models for ransomware detection is crucial to determine their effectiveness in detecting and preventing its spread. In this section, we will discuss different evaluation metrics used for measuring the performance of machine learning models for ransomware detection, including accuracy, precision, recall, F1-score, and ROC curve.

1. Accuracy: Accuracy is the most straightforward evaluation metric, representing the percentage of correct predictions made by the model. It is calculated as the ratio of accurate predictions to the total number of predictions. However, accuracy can be misleading when dealing with imbalanced datasets, where negative samples greatly outweigh the positive models [51,52].
2. Precision: Out of all samples predicted to be positive (recognized as ransomware by the algorithm), precision is the percentage of true positives (samples of successfully identified malware). The ratio of true positives to the total of true and false positives is

known as precision. A model with a high precision score will have a low false-positive rate, making it less likely to mistakenly label innocent files as ransomware [52].

3. Recall: Recall counts the number of positive samples in the collection that are true positives. The ratio of true positives to true and false negatives is computed. A high recall score suggests that the model has a low incidence of false negatives, which makes it less likely to fail to detect actual ransomware samples [13,52].
4. ROC curve: The performance of a binary classifier as the discrimination threshold is changed is graphically represented by a receiver operating characteristic (ROC) curve. At various threshold values, it plots the actual-positive rate (TPR) versus the false-positive rate (FPR). The model's overall performance is assessed using the area under the ROC curve (AUC), with higher AUC values indicating better performance [53].

7. Challenges and Future Directions

Developing effective machine-learning-based ransomware-detection systems is challenging due to several factors. This section will discuss the challenges of developing such systems and highlight the future directions in this field.

7.1. Challenges in Developing Effective Machine-Learning-Based Ransomware-Detection Systems

Developing effective machine-learning-based ransomware-detection systems presents several challenges, with the most common ones being:

1. Data quality and quantity—A vast amount of high-quality data are needed to train machine learning models effectively. However, obtaining high-quality data for ransomware detection is challenging due to the limited availability of labeled ransomware samples [54,55].
2. Rapidly evolving ransomware—Ransomware is a constantly changing threat, with new variants and attack techniques being developed regularly. This makes it challenging to build machine learning models that can detect all ransomware accurately and quickly [56].
3. Adversarial attacks involve modifying the input data to bypass the machine learning model's detection capabilities. Malicious attacks can be used to evade ransomware-detection systems, making the systems less effective [56].
4. Real-time detection requirements—Ransomware can spread rapidly and cause significant damage within a short time-frame. Therefore, ransomware-detection systems must be able to detect ransomware in real-time to prevent further spread and damage [57].
5. One of the main challenges in collecting data for ransomware detection is the need for publicly available datasets that include real-world ransomware samples. This is due to the sensitive nature of the data and the fact that many victims are reluctant to report ransomware attacks. As a result, researchers often rely on synthetic datasets or datasets generated from sandbox environments, which may not accurately reflect the complexity and variability of real-world ransomware attacks [3].
6. Another challenge is the diversity of ransomware families and variants, which require a large and diverse dataset to ensure adequate coverage. Ransomware behavior can also vary depending on the victim's system and network environment, making generalizing detection models across different contexts challenging [2,54].
7. Preprocessing data for ransomware detection also presents several challenges. Ransomware often employs obfuscation techniques to evade detection, such as encrypting the payload or using anti-analysis mechanisms. This can make extracting relevant data features and identifying patterns that distinguish ransomware from benign software difficult. In addition, ransomware may use legitimate system functions that are difficult to distinguish from malicious behavior, requiring advanced feature engineering and modeling techniques [54].
8. Despite these challenges, several datasets have been used to train and evaluate ransomware-detection models.

9. Collecting and preprocessing data for ransomware detection using machine learning presents several challenges, including the lack of real-world datasets, the diversity of ransomware families and variants, and the obfuscation techniques used by ransomware. However, several datasets have been developed to address these challenges, providing valuable resources for training and evaluating ransomware-detection models [54].

7.2. Future Work

Future work in machine-learning-based ransomware detection could include the following:

1. Developing more robust and accurate models—Researchers must build more substantial and precise machine learning models that detect a wide range of ransomware variants and attack techniques. This can be achieved through advanced techniques such as deep learning and ensemble learning [4,54,58].
2. Incorporating real-time detection capabilities—Ransomware-detection systems must incorporate real-time detection capabilities to quickly identify and prevent ransomware attacks. This can be achieved through the use of real-time monitoring and analysis techniques [55].
3. Addressing the issue of adversarial attacks—Researchers need to develop machine learning models that are robust to malicious attacks. This can be achieved through techniques such as negative training and defensive distillation [54,56].
4. Collaboration and sharing of data—Collaboration and sharing of data among researchers and organizations can help develop more effective ransomware-detection systems. This can help build more comprehensive datasets for training and testing machine learning models [56].
5. Developing effective machine-learning-based ransomware-detection systems is challenging for several reasons. However, with advanced techniques and collaboration among researchers and organizations, it is possible to develop more robust and accurate ransomware-detection systems [54].

8. Conclusions

Ransomware attacks have caused significant harm to computer systems and the data they manage, resulting in unauthorized access, disclosure, and the destruction of important and sensitive information. These attacks have led to substantial financial losses and reputational damage for both individuals and businesses. In response, various methods have been suggested to detect ransomware accurately, quickly, and dependably. This research provides readers with a historical background and timeline of ransomware attacks, as well as a discussion of the issue's context. The review of the recent literature offers an up-to-date understanding of automated ransomware-detection approaches. This knowledge will help readers stay current on the latest advances in automated ransomware detection, prevention, mitigation, and recovery. Additionally, this research discusses future research directions, highlighting open issues and potential research problems for those interested in researching ransomware detection, prevention, mitigation, and recovery.

Author Contributions: Author Contributions: collecting the papers, A.A. (Amjad Alraizza); Formal analysis, A.A. (Amjad Alraizza); Resources, A.A. (Abdulmohsen Algarni); Writing—original draft, A.A. (Amjad Alraizza); Writing review and editing, A.A. (Abdulmohsen Algarni); Supervision, A.A. (Abdulmohsen Algarni); Funding acquisition, A.A. (Abdulmohsen Algarni). All authors have read and agreed to the published version of the manuscript.

Funding: This research was financially supported by the Deanship of Scientific Research at King Khalid University under research grant number (R.G.P.2/549/44).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Celdrán, A.H.; Sánchez, P.M.S.; Castillo, M.A.; Bovet, G.; Pérez, G.M.; Stiller, B. Intelligent and behavioral-based detection of malware in IoT spectrum sensors. *Int. J. Inf. Secur.* **2022**, *22*, 541–561. [CrossRef]
2. Chesti, I.A.; Humayun, M.; Sama, N.U.; Jhanjhi, N. Evolution, mitigation, and prevention of ransomware. In Proceedings of the 2020 2nd International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 13–15 October 2020; pp. 1–6.
3. Philip, K.; Sakir, S.; Domhnall, C. Evolution of ransomware. *IET Netw.* **2018**, *7*, 321–327.
4. Jegede, A.; Fadele, A.; Onoja, M.; Aimufua, G.; Mazadu, I.J. Trends and Future Directions in Automated Ransomware Detection. *J. Comput. Soc. Inform.* **2022**, *1*, 17–41. [CrossRef]
5. Brewer, R. Ransomware attacks: Detection, prevention and cure. *Netw. Secur.* **2016**, *2016*, 5–9. [CrossRef]
6. Bello, I.; Chiroma, H.; Abdullahi, U.A.; Gital, A.Y.; Jauro, F.; Khan, A.; Okesola, J.O.; Abdulhamid, S.M. Detecting ransomware attacks using intelligent algorithms: Recent development and next direction from deep learning and big data perspectives. *J. Ambient Intell. Humaniz. Comput.* **2021**, *12*, 8699–8717. [CrossRef]
7. Zahra, A.; Shah, M.A. IoT based ransomware growth rate evaluation and detection using command and control blacklisting. In Proceedings of the 2017 23rd International Conference on Automation and Computing (ICAC), Huddersfield, UK, 7–8 September 2017; pp. 1–6.
8. Shaukat, S.K.; Ribeiro, V.J. RansomWall: A layered defense system against cryptographic ransomware attacks using machine learning. In Proceedings of the 2018 10th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, India, 3–7 January 2018; pp. 356–363.
9. Makinde, O.; Sangodoyin, A.; Mohammed, B.; Neagu, D.; Adamu, U. Distributed network behaviour prediction using machine learning and agent-based micro simulation. In Proceedings of the 2019 7th International Conference on Future Internet of Things and Cloud (FiCloud), Istanbul, Turkey, 26–28 August 2019; pp. 182–188.
10. Almashhadani, A.O.; Kaiiali, M.; Sezer, S.; O'Kane, P. A multi-classifier network-based crypto ransomware detection system: A case study of locky ransomware. *IEEE Access* **2019**, *7*, 47053–47067. [CrossRef]
11. Singh, A.; Ikuesan, R.A.; Venter, H. Ransomware detection using process memory. *arXiv* **2022**, arXiv:2203.16871.
12. Silva, J.A.H.; Hernández-Alvarez, M. Large scale ransomware detection by cognitive security. In Proceedings of the 2017 IEEE Second Ecuador Technical Chapters Meeting (ETCM), Salinas, Ecuador, 16–20 October 2017; pp. 1–4.
13. Azmoodeh, A.; Dehghantanha, A.; Conti, M.; Choo, K.K.R. Detecting crypto-ransomware in IoT networks based on energy consumption footprint. *J. Ambient Intell. Humaniz. Comput.* **2018**, *9*, 1141–1152. [CrossRef]
14. Ghouti, L.; Imam, M. Malware classification using compact image features and multiclass support vector machines. *IET Inf. Secur.* **2020**, *14*, 419–429. [CrossRef]
15. Modi, J. Detecting Ransomware in Encrypted Network Traffic Using Machine Learning. Ph.D. Thesis, University of Victoria, Saanich, BC, Canada, 2019.
16. Ameer, M. Android Ransomware Detection Using Machine Learning Techniques to Mitigate Adversarial Evasion Attacks. Master's Thesis, Capital University of Science and Technology, Islamabad, Pakistan, 2019.
17. Khammas, B.M. Ransomware detection using random forest technique. *ICT Express* **2020**, *6*, 325–331. [CrossRef]
18. Hwang, J.; Kim, J.; Lee, S.; Kim, K. Two-stage ransomware detection using dynamic analysis and machine learning techniques. *Wirel. Pers. Commun.* **2020**, *112*, 2597–2609. [CrossRef]
19. Talabani, H.S.; Abdulhadi, H.M.T. Bitcoin ransomware detection employing rule-based algorithms. *Sci. J. Univ. Zakho* **2022**, *10*, 5–10. [CrossRef]
20. Adamu, U.; Awan, I. Ransomware prediction using supervised learning algorithms. In Proceedings of the 2019 7th International Conference on Future Internet of Things and Cloud (FiCloud), Istanbul, Turkey, 26–28 August 2019; pp. 57–63.
21. Wan, Y.L.; Chang, J.C.; Chen, R.J.; Wang, S.J. Feature-selection-based ransomware detection with machine learning of data analysis. In Proceedings of the 2018 3rd International Conference on Computer and Communication Systems (ICCCS), Nagoya, Japan, 27–30 April 2018; pp. 85–88.
22. Alzahrani, A.; Alshehri, A.; Alshahrani, H.; Alharthi, R.; Fu, H.; Liu, A.; Zhu, Y. Randroid: Structural similarity approach for detecting ransomware applications in android platform. In Proceedings of the 2018 IEEE International Conference on Electro/Information Technology (EIT), Rochester, MI, USA, 3–5 May 2018; pp. 0892–0897.
23. Scaife, N.; Carter, H.; Traynor, P.; Butler, K.R. Cryptolock (and drop it): Stopping ransomware attacks on user data. In Proceedings of the 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS), Nara, Japan, 27–30 June 2016; pp. 303–312.
24. Sgandurra, D.; Muñoz-González, L.; Mohsen, R.; Lupu, E.C. Automated dynamic analysis of ransomware: Benefits, limitations and use for detection. *arXiv* **2016**, arXiv:1609.03020.
25. Prakash, K.P.; Nafis, T.; Biswas, S.S. Preventive Measures and Incident Response for Locky Ransomware. *Int. J. Adv. Res. Comput. Sci.* **2017**, *8*, 392–395.
26. Paquet-Clouston, M.; Haslhofer, B.; Dupont, B. Ransomware payments in the bitcoin ecosystem. *J. Cybersecur.* **2019**, *5*, tyz003. [CrossRef]
27. Kok, S.; Abdullah, A.; Jhanjhi, N.; Supramaniam, M. Ransomware, threat and detection techniques: A review. *Int. J. Comput. Sci. Netw. Secur* **2019**, *19*, 136.

28. Thakran, E.; Kumari, A. Impact of "Ransomware" on Critical Infrastructure Due to Pandemic. 2023; p. 5. Available online: https://ssrn.com/abstract=4361110 (accessed on 3 July 2023).
29. Ahmed, Y.A.; Huda, S.; Al-rimy, B.A.S.; Alharbi, N.; Saeed, F.; Ghaleb, F.A.; Ali, I.M. A weighted minimum redundancy maximum relevance technique for ransomware early detection in industrial IoT. *Sustainability* **2022**, *14*, 1231. [CrossRef]
30. Aslan, Ö.A.; Samet, R. A comprehensive review on malware detection approaches. *IEEE Access* **2020**, *8*, 6249–6271. [CrossRef]
31. Akhtar, M.S.; Feng, T. Malware Analysis and Detection Using Machine Learning Algorithms. *Symmetry* **2022**, *14*, 2304. [CrossRef]
32. Yamany, B.; Elsayed, M.S.; Jurcut, A.D.; Abdelbaki, N.; Azer, M.A. A New Scheme for Ransomware Classification and Clustering Using Static Features. *Electronics* **2022**, *11*, 3307. [CrossRef]
33. Yamany, B.; Azer, M.A.; Abdelbaki, N. Ransomware Clustering and Classification using Similarity Matrix. In Proceedings of the 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 8–9 May 2022; pp. 41–46.
34. Ullah, F.; Javaid, Q.; Salam, A.; Ahmad, M.; Sarwar, N.; Shah, D.; Abrar, M. Modified decision tree technique for ransomware detection at runtime through API Calls. *Sci. Program.* **2020**, *2020*, 8845833. [CrossRef]
35. Arunkumar, M.; Kumar, K.A. GOSVM: Gannet optimization based support vector machine for malicious attack detection in cloud environment. *Int. J. Inf. Technol.* **2023**, *15*, 1653–1660. [CrossRef]
36. Selamat, N.; Ali, F. Comparison of malware detection techniques using machine learning algorithm. *Indones. J. Electr. Eng. Comput. Sci.* **2019**, *16*, 435. [CrossRef]
37. Mezquita, Y.; Alonso, R.S.; Casado-Vara, R.; Prieto, J.; Corchado, J.M. A review of k-nn algorithm based on classical and quantum machine learning. In *Distributed Computing and Artificial Intelligence, Special Sessions, 17th International Conference*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 189–198.
38. Saadat, S.; Joseph Raymond, V. Malware classification using CNN-XGBoost model. In *Artificial Intelligence Techniques for Advanced Computing Applications: Proceedings of ICACT 2020*; Springer, Berlin/Heidelberg, Germany, 2021; pp. 191–202.
39. Noorbehbahani, F.; Rasouli, F.; Saberi, M. Analysis of machine learning techniques for ransomware detection. In Proceedings of the 2019 16th International ISC (Iranian Society of Cryptology) Conference on Information Security and Cryptology (ISCISC), Mashhad, Iran, 28–29 August 2019; pp. 128–133.
40. Sharmeen, S.; Ahmed, Y.A.; Huda, S.; Koçer, B.Ş.; Hassan, M.M. Avoiding future digital extortion through robust protection against ransomware threats using deep learning based adaptive approaches. *IEEE Access* **2020**, *8*, 24522–24534. [CrossRef]
41. Swami, S.; Swami, M.; Nidhi, N. Ransomware Detection System and Analysis Using Latest Tool. *Int. J. Adv. Res. Sci. Commun. Technol.* **2021**, *7*, 2581–9429. [CrossRef]
42. Wang, X.b.; Yang, G.y.; Li, Y.c.; Liu, D. Review on the application of artificial intelligence in antivirus detection system i. In Proceedings of the 2008 IEEE Conference on Cybernetics and Intelligent Systems, Chengdu, China, 21–24 September 2008; pp. 506–509.
43. Yang, B.; Liu, D. Research on Network Traffic Identification based on Machine Learning and Deep Packet Inspection. In Proceedings of the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 15–17 March 2019; pp. 1887–1891. [CrossRef]
44. Pimenta Rodrigues, G.A.; de Oliveira Albuquerque, R.; Gomes de Deus, F.E.; de Sousa Jr, R.T.; de Oliveira Júnior, G.A.; Garcia Villalba, L.J.; Kim, T.H. Cybersecurity and network forensics: Analysis of malicious traffic towards a honeynet with deep packet inspection. *Appl. Sci.* **2017**, *7*, 1082. [CrossRef]
45. Song, W.; Beshley, M.; Przystupa, K.; Beshley, H.; Kochan, O.; Pryslupskyi, A.; Pieniak, D.; Su, J. A software deep packet inspection system for network traffic analysis and anomaly detection. *Sensors* **2020**, *20*, 1637. [CrossRef]
46. Cascarano, N.; Ciminiera, L.; Risso, F. Optimizing deep packet inspection for high-speed traffic analysis. *J. Netw. Syst. Manag.* **2011**, *19*, 7–31. [CrossRef]
47. Dargahi, T.; Dehghantanha, A.; Bahrami, P.N.; Conti, M.; Bianchi, G.; Benedetto, L. A Cyber-Kill-Chain based taxonomy of crypto-ransomware features. *J. Comput. Virol. Hacking Tech.* **2019**, *15*, 277–305. [CrossRef]
48. Sheen, S.; Asmitha, K.; Venkatesan, S. R-Sentry: Deception based ransomware detection using file access patterns. *Comput. Electr. Eng.* **2022**, *103*, 108346. [CrossRef]
49. Madani, H.; Ouerdi, N.; Boumesaoud, A.; Azizi, A. Classification of ransomware using different types of neural networks. *Sci. Rep.* **2022**, *12*, 4770. [CrossRef] [PubMed]
50. Arivudainambi, D.; Varun Kumar, K.A.; Visu, P.; Sibi Chakkaravarthy, S. Malware traffic classification using principal component analysis and artificial neural network for extreme surveillance. *Comput. Commun.* **2019**, *147*, 50–57.
51. Kok, S.; Azween, A.; Jhanjhi, N. Evaluation metric for crypto-ransomware detection using machine learning. *J. Inf. Secur. Appl.* **2020**, *55*, 102646. [CrossRef]
52. Masum, M.; Faruk, M.J.H.; Shahriar, H.; Qian, K.; Lo, D.; Adnan, M.I. Ransomware classification and detection with machine learning algorithms. In Proceedings of the 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 26–29 January 2022; pp. 0316–0322.
53. Edis, D.; Hayman, T.; Vatsa, A. Understanding Complex Malware. In Proceedings of the 2021 IEEE Integrated STEM Education Conference (ISEC), Princeton, NJ, USA, 13 March 2021; pp. 1–2.
54. Beaman, C.; Barkworth, A.; Akande, T.D.; Hakak, S.; Khan, M.K. Ransomware: Recent advances, analysis, challenges and future research directions. *Comput. Secur.* **2021**, *111*, 102490. [CrossRef] [PubMed]

55. McIntosh, T.; Kayes, A.; Chen, Y.P.P.; Ng, A.; Watters, P. Ransomware mitigation in the modern era: A comprehensive review, research challenges, and future directions. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–36. [CrossRef]
56. Aboaoja, F.A.; Zainal, A.; Ghaleb, F.A.; Al-rimy, B.A.S.; Eisa, T.A.E.; Elnour, A.A.H. Malware detection issues, challenges, and future directions: A survey. *Appl. Sci.* **2022**, *12*, 8482. [CrossRef]
57. Gorment, N.Z.; Selamat, A.; Cheng, L.K.; Krejcar, O. Machine Learning Algorithm for Malware Detection: Taxonomy, Current Challenges and Future Directions. *IEEE Access* **2023**, 1. [CrossRef]
58. Kapoor, A.; Gupta, A.; Gupta, R.; Tanwar, S.; Sharma, G.; Davidson, I.E. Ransomware detection, avoidance, and mitigation scheme: A review and future directions. *Sustainability* **2021**, *14*, 8. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Security and Privacy Threats and Requirements for the Centralized Contact Tracing System in Korea

Sungchae Park and Heung-Youl Youm *

Department of Information Security Engineering, Soonchunhyang University, Asan-si 31538, Republic of Korea
* Correspondence: hyyoum@sch.ac.kr

Abstract: As COVID-19 became a pandemic worldwide, contact tracing technologies and information systems were developed for quick control of infectious diseases in both the private and public sectors. This study aims to strengthen the data subject's security, privacy, and rights in a centralized contact tracing system adopted for a quick response to the spread of infectious diseases due to climate change, increasing cross-border movement, etc. There are several types of contact tracing systems: centralized, decentralized, and hybrid models. This study demonstrates the privacy model for a centralized contact tracing system, focusing on the case in Korea. Hence, we define security and privacy threats to the centralized contact tracing system. The threat analysis involved mapping the threats in ITU-T X.1121; in order to validate the defined threats, we used LIDDUN and STRIDE to map the threats. In addition, this study provides security requirements for each threat defined for more secure utilization of the centralized contact tracing system.

Keywords: centralized contact tracing system; Korea COVID-19 smart management system (SMS); privacy model; security threats; privacy threats; security and privacy requirements

Citation: Park, S.; Youm, H.-Y. Security and Privacy Threats and Requirements for the Centralized Contact Tracing System in Korea. *Big Data Cogn. Comput.* **2022**, *6*, 143. https://doi.org/10.3390/bdcc6040143

Academic Editors: Peter R.J. Trim and Yang-Im Lee

Received: 5 August 2022
Accepted: 7 October 2022
Published: 28 November 2022

1. Introduction

As COVID-19 became a pandemic worldwide, contact tracing technologies and information systems were developed for the quick control of infectious diseases in both the private and public sectors. The systems were implemented to collect and process various data to monitor the COVID-19 pandemic, and the pandemic was managed through contact tracing systems that can identify contacts and prevent the spread of infectious diseases.

There are several types of contact tracing systems; apart from centralized and decentralized models, a hybrid way has been approached. In centralized contact-tracing, mobiles share their anonymous IDs to a central server maintaining a centralized database, and the server uses this database to perform contact tracing, risk analysis, and alert notifications to the users [1]. The ROBERT (ROBust and privacy-presERving proximity Tracing) protocol is an example of the centralized contact tracing system adopted by France and Europe. ROBERT is a joint contribution in the framework of the PEPP-PT (Pan European Privacy-Preserving Proximity Tracing) initiative, which aims to enable the development of interoperable contact tracing solutions that comply with European data protection, privacy, and security standards as part of a more comprehensive response to the pandemic [2]. De-centralized contact tracing, on the other hand, does not send any PII data to the centralized server but stores all PII data in the user's mobile phone and notifies them when they come into contact with a confirmed case. In addition, each user's mobile phone acts as a local server that shares only the infected individual's data to the centralized server, and then mobile phones will fetch this data periodically from the server and do contact matching locally [1]. An example of a decentralized contact tracing system is DP-3T: a decentralized, privacy-preserving proximity tracing system. DP-3T aims to minimize privacy and security risks for individuals and communities and guarantees the highest level of data protection [3]. A hybrid architecture may have a component of both approaches, with some information handled on individual devices with a central server analyzing data and sending notifications [4]. PIVOT (Private and Effective Contact Tracing) and DESIRE(a novel exposure notification system that leverages the best of centralized and decentralized systems) have been known as representative examples of the hybrid approach for contact tracing systems [5,6].

Each system has different priorities in terms of a quick response to confirmed cases and privacy. A major advantage of a centralized contact tracing system is that health authorities enable an infectious diseases situation to be controlled more effectively, such as COVID-19. However, a centralized system requires the extensive collection of personal data within the centralized server or systems. In addition, it may cause higher risks of security and privacy issues compared to decentralized and hybrid systems.

This study analyzes a privacy model and the security and privacy threats of a centralized contact tracing system, based on Korea's COVID-19 smart management system. We also identify relevant security and–privacy requirements, which should be taken into account at each processing of data.

2. Related Works

Contact tracing is an effective method to control emerging infectious diseases. Since the 1980s, modelers have been developing a consistent theory for contact tracing, with the aims to find effective and efficient implementations and to assess the effects of contact tracing on the spread of an infectious disease. Contact tracing is a more focused method: once an infected individual is diagnosed and isolated, contact persons are identified, who potentially had infectious interactions with that index case [7].

This section summarizes the previous literature with regards to contact tracing technology, which is an effective method to control emerging infectious diseases and compares the features of this paper with the related literature.

2.1. Case Study of Contact Tracing for COVID-19 in Korea

In April 2021, the *Journal of the American Medical Association* (JAMA) published research regarding the information-technology-based tracing strategy in response to COVID-19 in South Korea and the related privacy controversies, as studied by Seoul National University (SNU)'s Haksoo Ko. This research covers legal and policy responses of contact tracing related to COVID-19 in South Korea. It explains that South Korea extensively utilized the country's advanced information technology (IT) system for tracing individuals suspected to be infected or who had been in contact with an infected person. In addition, this research emphasizes that there is a need for a balance between privacy issues and the effects of epidemiological investigations brought about by the extensive tracing of infected people and disclosure of collected information [8].

2.2. Case Study on COVID-19 Contact Tracing in Taiwan

Inspired by the lessons learned from the Ebola outbreak in West Africa, the Taiwan Center for Disease Control (TCDC) developed a national contact tracing platform named TRACE in 2017, to link other data systems, monitor the health status of contacts, and support the management of contacts by compiling the daily descriptive analysis and relevant performance indicators. The modules in TRACE were applicable for all notifiable diseases in Taiwan, and they have been implemented for contact tracing in diseases such as measles and rubella and for health monitoring of individuals exposed to animals with avian influenza. For the COVID-19 outbreak response, Taiwan's government developed a COVID-19 module in mid-January 2020 to support contact tracing. To ensure confidentiality, the database that contained contacts' personally identifiable information(PII) would be deleted in six months and could not be used for other purposes [9].

There is another study on contact tracing in Taiwan. The purpose of this study was to measure the high national acceptance for COVID-19 contact tracing technologies in Taiwan. The study is regarding that the effectiveness of government policies in the control of the spread of COVID-19 and the acceptance of such government policies among people are different. In addition, it shows acceptance increased with the perceived technology benefits; trust in the providers' intent, data security, and privacy measures; the level of ongoing control; and one's level of education. Acceptance decreased with data sensitivity perceptions and perceived low policy compliance by others in the general public [10].

2.3. Analysis and Comparison of Privacy in Contact Tracing Apps

When people first started using contact tracing applications, privacy issues for the people who are infected with COVID-19 occurred. In addition, there was resistance to the use of contact tracing apps and discrimination against patients with coronavirus disease. A study based on this situation modeled specific privacy threats to explain the detailed analysis results of COVID-19 tracing apps and the main differences between privacy protection and security performance among various contact tracing apps. This study described different national cultures that tend to select centralized and decentralized contact tracking applications. Furthermore, this study emphasized that it is undeniable that privacy has been violated to some extent no matter what application is used in the context of prevention and control. In order to protect personal data, privacy threat analysis of various contact tracing technologies and a comparison of the results of contact tracing apps for COVID-19 suggest that infectious disease prevention, control, and privacy can be effectively protected [11].

2.4. Case Study of COVID-19 Contact Tracing Mobile Application in Singapore

The Singaporean government released a mobile phone app, TraceTogether, which is designed to assist health officials in tracking down exposures after an infected individual is identified. However, there are important privacy implications of the existence of such tracking apps. A related study analyzes some of those implications and proposes ways of ameliorating privacy concerns without decreasing the usefulness to public health [12].

2.5. Differences and Contributions of This Paper

In addition to the studies described above, there are various meaningful studies on infectious disease contact tracing techniques or systems such as those for COVID-19. There have been a lot of papers, research, studies, etc., in terms of infectious disease control and security, including a comparison of centralized and decentralized contact tracing systems. These published studies revealed the positive effects of the IT technologies reflecting each government's policy, enabling a rapid response to global infectious diseases such as the COVID-19 pandemic. A comparison of this paper and the abovementioned related works is displayed in Table 1. In this Table 1, ○ means that relevance issue in column 1 is addressed and ✕ is not addressed.

Table 1. A comparison of this paper and the related works.

Contents of This Paper	This Paper	2.1	2.2	2.3	2.4
Privacy modeling of contact tracing system	○	✕	✕	✕	✕
Security and privacy threats analysis	○	✕	✕	○	○
Security and privacy requirements mapping	○	✕	✕	○	○
Contact tracing technology based method	QR code Credit card	Not mentioned	Bluetooth	Bluetooth GPS	Bluetooth
Comparison of a centralized and decentralized model	✕	✕	○	○	✕

○ means that relevance issue in column 1 is addressed and ✕ is not addressed.

Table 2 provides some contact tracing systems or applications developed by many countries including Korea as well as companies.

Table 2. The examples of contact tracing systems/applications by countries/authors.

Country or Authors	Examples of Contact Tracing Systems or Applications	Approach
Korea	Korea COVID-19 smart management system	Centralized
UK	NHS contact tracing app [13]	Centralized
China	Health Code [14]	Centralized
Singapore	TraceTogether (OpenTrace/BlueTrace) [12,15]	Centralized
EU	PEPP-PP [16]	Centralized
EU	DP-3T [17]	Decentralized
TCN Coalition	TCN [18]	Decentralized
Google/Apple	Google–Apple Exposure Notification application programming interface (API) [19]	Decentralized
Norway	Smittestopp [20]	Centralized Decentralized
Mahabir Prasad Jhanwar, Sumanta Sarkar	PHyCT (Privacy preserving Hybrid Contact Tracing) [21]	Hybrid
Giuseppe Garofalo, Tim Van hamme, et al.	PIVOT (PrIVate and effective cOntact Tracing) [6]	Hybrid
Claude Castelluccia, Nataliia Bielova, et al.	DESIRE (a novel exposure notification system that leverages the best of centralized and decentralized systems) [5]	Hybrid

The NHS COVID-19 app uses Bluetooth Low Energy (BLE) to understand the distance, over time, between app users and send an exposure notification to someone who has had close contact [13]. The Chinese government relies on Health Code, developed by Alipay and WeChat, for identifying people potentially exposed to COVID-19 [14]. TraceTogether is the first national deployment of a Bluetooth-based contact tracing system in the world. It was developed by the Singaporean government's Technology Agency and Ministry of Health to help the country better respond to epidemics [15]. The purpose of the Pan-European Privacy-Preserving Proximity Tracing (PEPP-PT) approach is to provide a common basis for management systems that can be integrated into national public health responses to the COVID-19 pandemic. The PEPP-PT approach has been created by a multinational European team [16]. DP-3T determines who has been in close physical proximity to a COVID-19-positive person without revealing that person's identity or where the contact occurred, requiring a centralized database or server [17]. TCN is a protocol developed by the TCN Coalition, which has jointly developed a common protocol between their apps [18]. The Google–Apple Exposure Notification application programming interface (API) is the most representative example of a decentralized contact tracing system based on Bluetooth. This exposure notification app generates a random ID for a mobile phone without tracking a person's location [19]. Norway released two types of contact tracing applications, based on the centralized approach for the first version and the decentralized approach for the second version. The decentralized approach is based on the protocol for exposure notification by Apple and Google [20]. In addition, a hybrid model may have a component of both approaches, with some information handled on individual devices with a central server analyzing data and sending notifications [4].

In Europe and North America, a decentralized contact tracing system has been mainly preferred, but, in Asia, a centralized contact tracing system has been used more. Hybrid contact tracing systems have been introduced in journals and some technical reports. Not all the contact tracing systems or applications in the table above have been used or adopted successfully. Table 2 lists some representative examples of contact tracing systems or applications used during the COVID-19 pandemic. This study describes a privacy perspective model for a centralized infectious disease contact tracking system and a life cycle for processing collection information, focusing on Korean cases. Moreover, our work analyzes the security and privacy threats affecting the privacy model of a centralized contact tracing system on a QR code basis and aims to validate it; we present the mapping result of the security and–privacy requirements against each threat.

3. Data Processing Model for a Centralized Contact Tracing System

3.1. Korea's COVID-19 Smart Management System (SMS) [22]

The Korean government defines basic activities that need to be accomplished to prevent the spread of COVID-19, per 'Infectious Disease Control and Prevention Act' as an 'epidemiological investigation' [23], and has developed a centralized system to control the spread of COVID-19, which is called the COVID-19 smart management system. This system enables the automation of the epidemiological investigation, as specified in 'Infectious Disease Control and Prevention Act', and it has developed the application of smart city technologies to collect, process, and analyze a huge volume of urban data.

Through this system, it is possible to secure epidemiological investigation results within 10 min by the real-time analysis of the movements of confirmed patients and large-scale outbreak areas, by using big data linked to 28 institutions through the cooperation of government agencies. The use of the personal information from confirmed cases in this system is based on the regulations of the 'Infectious Disease Control and Prevention Act' that allow for the public to use some personal information that would be sensitive for accurate epidemiological investigations in infectious disease crisis situations [23]. This policy was put in place to conduct accurate epidemiological investigations, with the legal change occurring after the Middle East Respiratory Syndrome (MERS) outbreak in 2015. This law allows for the use of personal information in exceptional cases for the prevention of infectious diseases such as COVID-19, through the cooperation and approval of relevant agencies. Korea's COVID-19 smart management system collects minimal data and applies a strict data collection process for the use and safe management of PII [24].

Korea's government has developed a centralized infectious disease contact tracing system, as shown in Figure 1. When a confirmed case of COVID-19 occurs, the data collection method for tracking the movement of the confirmed person is as follows:

- QR-code-based electronic access list used when entering specific facilities;
- Handwritten list;
- Collected mobile phone numbers recorded by people calling with phone numbers issued by local governments when entering specific facilities.

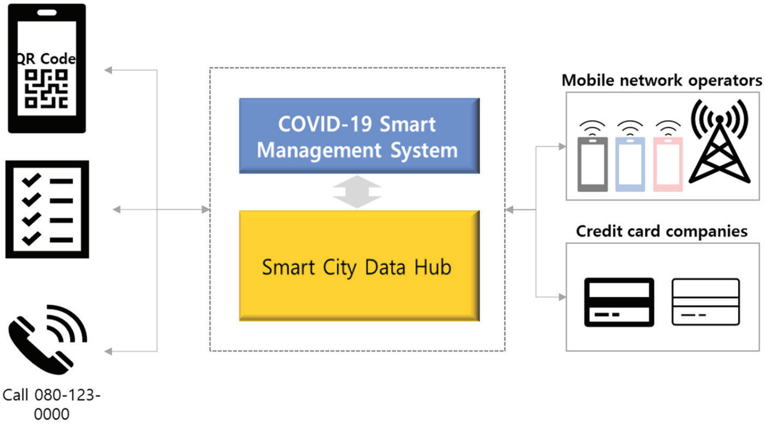


Figure 1. Korea's COVID-19 smart management system (SMS).

The records of subjects' facility visits are stored on the management server of the Korea Social Security Information Service (SSIS), and the PII of the QR code is encrypted and stored on the server of each company that issued the QR code. In addition, when data are required, in the case of the occurrence and tracing of confirmed COVID-19 cases, the distributed data that were stored on differently located servers are called by the COVID-19 smart management system, which combines the required data for tracing.

In the case of Korea's contact tracing system, it is not simply limited to the mobile app as a system. Korea's contact tracing system, the COVID-19 Smart Management System (SMS), operates by leveraging the central data hub platform of the Korean government, taking into account both the pre-confirmation status and the confirmed status [22].

The KCDC shares data and cooperates with central, municipal, or local governments; national health insurance agencies; and health care professionals and their associations, as depicted in Figure 2. This system enabled the prompt delivery of data pertaining to the confirmed cases to relevant agencies. Furthermore, the MOHW must release information such as the path and means of transportation of infected persons, etc., on the Internet or through a press release [8].

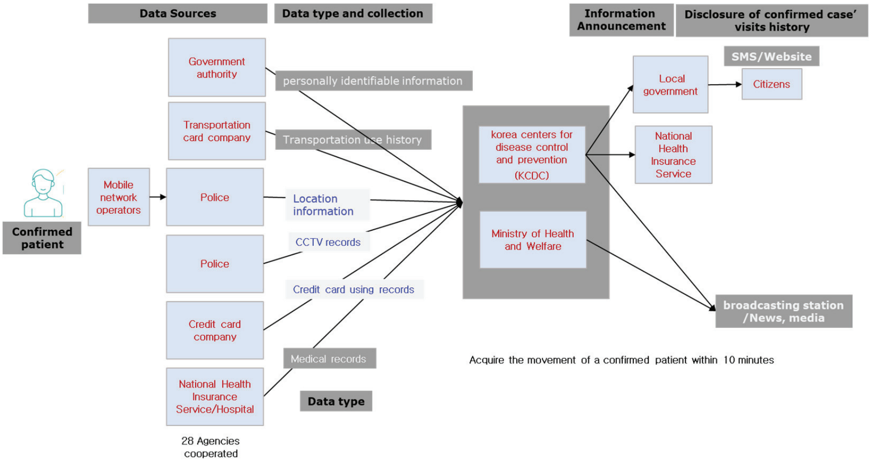

Data Sources
Data type and collection
Information Announcement
Disclosure of confirmed case' visits history
Government authority
personally identifiable information
Transportation card company
Transportation use history
Confirmed patient
Mobile network operators
Police
Location information
Police
CCTV records
Credit card using records
Credit card company
Medical records
National Health Insurance Service/Hospital
Data type
28 Agencies cooperated
korea centers for disease control and prevention (KCDC)
Ministry of Health and Welfare
Local government
SMS/Website
Citizens
National Health Insurance Service
broadcasting station /News, media
Acquire the movement of a confirmed patient within 10 minutes

Figure 2. An example of the overall structure of the Korean contact tracing system [25].

An example of the overall structure of the Korean contact tracing system is shown in Figure 2.

3.2. Data Processing Model

This paper suggests a privacy model for a centralized contact tracing system based on the case study of the Korean system for the prevention and control of infectious diseases. In the Korean system, a third party, the Korea Centers for Disease Control and Prevention (KCDC), works as a centralized server that mainly processes the health information and infection status, analyzing that to identifying patients and contactors and collaborating with other third parties to quarantine patients and publish infection information to the public [24]. In a centralized privacy model, third parties takes a significant role in the response and control of infectious diseases.

The privacy model of a centralized contact tracing system is depicted in Figure 3 [1].

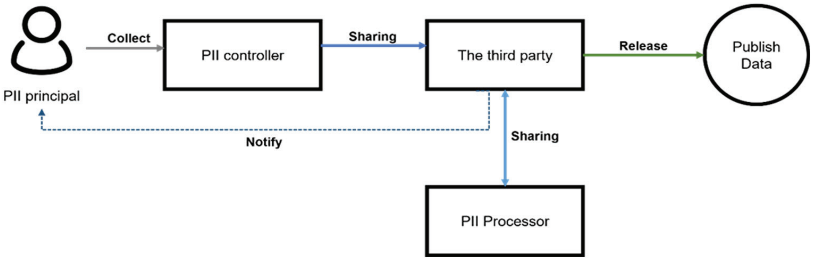


Figure 3. Privacy model of a centralized contact tracing system.

In this model, each party performs the following roles:

- PII: PII is data such as the phone number or credit card number of a data subject, the place where a credit card is used, and the location of the mobile base station.
- PII principal: A stakeholder who provides PII to classify whether they are a contact of an epidemic patient or is diagnosed as positive for an infectious disease. PII may include epidemic information and other information that could help to identify a recent contact including geological information. A PII principal could receive notification of exposure to an epidemic patient from the third party.
- PII controller: A stakeholder who collects PII and shares it to a third party. The collected information can be directly related to infection information such as diagnosis or other information to track a patient/contact path to identify other contacts [26], though a PII controller may need additional consent from a PII principal to use such data in this system. Medical institutions or private service providers related to geological/financial services can be PII controllers in this model.
- Third Party: A stakeholder receiving information from the PII controllers, who takes measures to prevent and manage epidemics and has an obligation to disclose information to share the status of outbreak and spread. Organizations could be a third party such as the Korea Centers for Disease Control and Prevention (KCDC), which oversees all health data processing, as well as local governments that carry out quarantine measures for infected persons/contacts [24].
- PII Processor: A stakeholder who processes data on behalf of a third party and can process data analysis, integration, and de-identification on behalf of a third party. If a third party operates a health information system by their own, a PII processor could help to establish or maintain such information systems [24]. PII processors include data service providers [24].

3.3. Data Processing Life Cycle

The data processing of the centralized contact tracing system may have six steps from collection to retention, as shown in Figure 4. The cycle includes the collection, sharing, processing, notification, release, and retention of data.

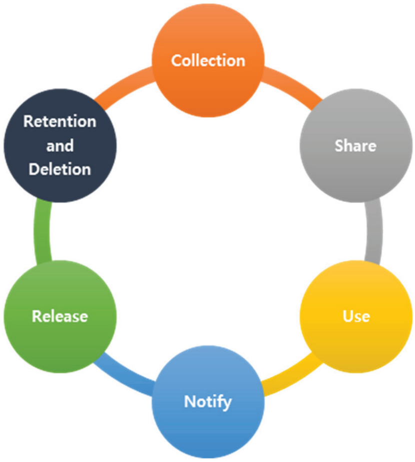


Figure 4. Data processing steps of the centralized contact tracing system.

- Data Collection

The data from a PII principal are collected by a PII controller [27]. The PII controller collects infection information and other contact data from infected persons to trace the path and identify contact [24]. Collection from private PII controllers can only be practiced when it has permission from an authority to legally permit collection [24].

- Data Sharing

The data collected from PII controllers are shared to a third party to respond against epidemic diseases. If a PII principal is diagnosed as positive, the PII controller, such as epidemiological investigators or the medical institution that performed tests, reports the infection information to the third party and requests the third party to take quarantine measures [24]. The third party additionally receives data from PII controllers after permission is given by an authority such as a national police agency or credit association to further identify the contact, if necessary [24].

The third party should use the data provided by the PII controller only for the purpose of preventing, controlling, and treating infectious diseases. Such a process could include the aggregation of geological data to identify contact with an infected person or anonymization or de-identification to create the disease statistics to be used in a data release to the public. A third party may request a PII processor complete data analysis, aggregation, and anonymization on behalf of the third party, in accordance with instructions [24]. When a third party builds, maintains, and manages its own information system to process infection information, a PII processor could support the maintenance of such systems [24]. The PII processor processes the data and sends them to a third party after processing them for such a purpose. This information must only be used for the purposes of epidemic responses [23].

- Data Notification

When a PII principal is tested as positive or classified as a contact, the third party quickly notifies the PII principal of their status and quarantines them [28]. The third party receives the data, notifies PII principals that one has been in contact with a disease patient, and performs measures to prevent infectious diseases, such as quarantining a PII principal under their jurisdiction [28].

- Data Release

The third party releases the statistics of an infectious disease to inform the public about the outbreak and spread of the disease [23]. In order to do so, the third party could request the de-identification of PII processors that then only receive statistical information. Next, the third party conducts a data release and provides the media and the public with information about the status of medical institutions and contacts and the occurrence and testing of infectious diseases by region and age group [24].

- Data Retention and Deletion

All institutions must destroy all data when the purpose of an epidemic response is achieved, which must be destroyed without delay. For example, the data are destroyed after 4 weeks in regard to the impact of COVID-19 in Korea [29,30].

4. Security and Privacy Threats and Requirements

4.1. Security Threats and Requirements in ITU-T X.1121

In this study, we analyze the security and privacy threats and security requirements for the contact tracing system in Korea, as mentioned in Sections 4.2 and 4.4. However, there are security threats and requirements in ITU-T X.1121, which is the framework for security technologies for mobile end-to-end data communications. To analyze and identify more specific threats focusing on the centralized contact tracing system in Korea, we refer to ITU-T X.1121 and compare it with the security threats and the security–privacy threats in this study.

Table 3 shows the security threats and requirements in ITU-T X.1121 [31].

Table 3. Security threats and requirements in ITU-T X.1121.

Threat Requirement	Eavesdropping	Communication Jamming	Shoulder Surfing	Lost/Stolen Terminal	Unprepared Shutdown	Misreading/ Input Error
Identity management	X					
Communication data confidentiality	X					
Stored data confidentiality				X		
Communication data integrity						
Stored data integrity				X		
Entity authentication				X		
Message authentication						
Access control				X		
Non-repudiation						
Anonymity				X		
Privacy	X		X	X		
Usability						X
Availability		X			X	

We compared the security threats of ITU-T X.1121 with the threats derived by this paper. In addition, we could find additional threats related to the loss of terminals. Moreover, the security requirements for this additional threat were also addressed.

- Additional threat: Lost/stolen terminal
- Corresponding requirement: If a terminal is lost, people will not be able to receive the information related to it when they become a close contact. Therefore, various notification methods, such as e-mail notification, etc., for the recipient, that is, the closer contact, should be improved.

4.2. Security and Privacy Threats for the Contact Tracing System in Korea

This section lists the affectable security and privacy threats in the centralized privacy model of the mentioned system and maps each threat to the related entity, which are the affectable privacy threats to the privacy model of the health information system for epidemic alert and response:

ST1. Compromise of data confidentiality: threats to data being disclosed or available to the unintended entity;

ST2. Compromise of data integrity: threats to data being changed or destructed improperly;

ST3. Compromise of data availability: threats to data being accessed or used by unauthorized third parties;

ST4. Data recovery due to insufficient data deletion: threats to data being recovered from data storage, due to insufficient data deletion;

ST5. Degradation in data quality when processing: threats of data being corrupted or redundant due to processing on data such as de-identification and a failure to identify past or present physical proximity;

ST6. Malicious activities by internal attackers: threats of data being maliciously leaked by internal attackers from inside;

ST7. Use of unsecured tunneling protocol: protocol attacks caused by using versions to be vulnerable to communication protocols;

ST8. Lost/stolen terminal: threats to lost or stolen terminal such as mobiles;

PT1. Data use for purposes other than infectious disease responses: threats to data being used for purposes other than the prevention, management, and treatment of infectious diseases;

PT2. Unauthorized data transfer to third party: threats of data being acquired or provided to unauthorized third party by false or other fraudulent means or methods;

PT3. Insufficient legal and regulation grounds for PII processing: threats of insufficient legal grounds for collection of PII;

PT4. Excessive data processing beyond the intended purposes: threats of data being collected unreasonably because of too many attributes for the original purpose;
PT5. Data collection without the consent of a PII principal: threats in a process of collection when prior consent of a PII principal is not being obtained;
PT6. De-identification risk of re-identified data: the potential that some supposedly anonymous or pseudonymous data sets could be being de-anonymized to recover the identities of users [32];
PT7. Identification of a specific data from publicly announced data: threats to leak specific PII by using and combining publicly announced information or data such as the movement routes of people with infectious diseases;
PT8. Leakage of PII on a handwritten list: threat to leakage of a specific PII being written on a handwritten list when an individual enters various facilities.

Here, ST and PT mean security threat and privacy threat, in order to assign numbers to use in the mapping tables shown in Table 4.

Table 4. Security and privacy threats by stakeholders.

Stakeholders	**Security and Privacy Threats**															
	ST1	**ST2**	**ST3**	**ST4**	**ST5**	**ST6**	**ST7**	**ST8**	**PT1**	**PT2**	**PT3**	**PT4**	**PT5**	**PT6**	**PT7**	**PT8**
PII Controller	○	○	○	○		○	○	○	○	○			○			○
Third Party	○	○	○	○	○		○				○	○		○	○	
PII Processor	○	○	○	○	○		○				○	○		○		

○ means that relevance issue is addressed.

As can be seen from Table 4, two types of threats can occur for each stakeholder. Since the contact tracing system should collect and handle personal sensitive data, especially if the system is based on the centralized model, the threats that can occur are classified as general security or privacy. When analyzing threats such as the above, each stakeholder has two types of threats all, though a PII controller could have more privacy threats.

4.3. Mapping Security and Privacy Threats to LINDDUN and STRIDE Threat Models

In Section 4.2, we map the security and privacy threats derived in Section 4.1 to LINDDUN and STRIDE, which are security-threat modeling techniques. This would mean that the threats derived in this study are complete. LINDDUN is a privacy-threat modeling methodology that supports analysts in systematically eliciting and mitigating privacy threats in software architectures [33]. The STRIDE models were developed by Microsoft for categorizing threats. The classification of threats in this model is accomplished by categorizing the kind of exploit done by an attacker or intruder [34].

The LINDDUN model has seven threat categories, and each category is as follows:

- **L (Linkability):** An adversary is able to link two items of interest without knowing the identity of the data subjects involved [33];
- **I (Identifiability):** An adversary is able to identify a data subject from a set of data subjects through an item of interest [33];
- **N (Non-repudiation):** The data subject is unable to deny a claim [33];
- **D (Detectability):** An adversary is able to distinguish whether an item of interest about a data subject exists or not, regardless of being able to read the contents itself [33];
- **D (information Disclosure):** An adversary is able to learn the content of an item of interest about a data subject [33];
- **U (content Unawareness):** The data subject is unaware of the collection, processing, storage, or sharing activities and the corresponding purposes of their personal data [33];
- **N (policy and consent Non-compliance):** The processing, storage, or handling of personal data is not compliant with legislation, regulation, and/or policy [33].

The mapping table for each category of the derived threat and LINDDUN is shown in Table 5.

Table 5. Security–privacy threats and LINDDUN mapping lists.

No.	Threats	L	I	N	D	D	U	N
ST1	Compromised data confidentiality					○		
ST2	Compromised data integrity			○				
ST3	Compromised data availability					○		
ST4	Data recovery due to insufficient data deletion					○		
ST5	Degradation in data quality when processing	○	○					
ST6	Malicious actions by internal attackers		○		○	○		
ST7	Risk of using unsecure tunneling protocol							○
ST8	Lost/stolen terminal						○	
PT1	Data use for purposes other than infectious-disease responses							○
PT2	Data transfer to unauthorized third party						○	
PT3	Insufficient legal basis for PII collection							○
PT4	Excessive data collection and use beyond purpose					○		○
PT5	Data collection without consent of PII principal	○	○					○
PT6	Risk of re-identification due to data combination	○	○					
PT7	Threats to know a specific subject of information using publicly announced information	○	○					
PT8	Leakage of PII on the handwritten list	○	○			○		

○ means that relevance issue is addressed.

The STRIDE model has 7 threat categories, and each category is as follows:

- **S (Spoofing):** Spoofing or "identity spoofing" is a scenario in which a user X pretends to be a user Y by changing their identity or data and, thus, gains illegal access to data [35];
- **T (Tampering):** Tampering refers to the change of data by an illegal person who is not authorized to modify them [35];
- **R (Repudiation):** Repudiation relies on the fact that a security system must always be able to trace the entity responsible for any illegitimate modification and illegal access of resource or account [35];
- **I (Information disclosure):** Information disclosure assists an attacker or malicious user in accessing confidential information that they are not permitted to view [35];
- **D (Denial of service):** A denial-of-service (DoS) attack is an attempt to disturb a resource, network, or system in such a way that an intended and valid user would not be able to use it [35];
- **E (Elevation of privilege):** Elevation of privilege is the category of attacks in which an intruder gains authorization to access more than what hasbeengrantedoriginally [35].

The mapping table for each category of the derived threat and STRIDE is shown in Table 6.

Table 6. Security threats and STRIDE mapping list.

No.	Threats	S	T	R	I	D	E
ST1	Compromised data confidentiality				○		
ST2	Compromised data integrity		○	○			
ST3	Compromised data availability						○
ST4	Data recovery due to insufficient data deletion				○		
ST5	Degradation in data quality when processing					○	
ST6	Malicious actions by internal attackers				○		
ST7	Risk of using unsecured tunneling protocol	○		○			
ST8	Lost/stolen terminal	○	○		○		

○ means that relevance issue is addressed.

4.4. Security and Privacy Requirements for the Contact Tracing System in Korea

This section provides the security requirements to mitigate the listed privacy and security threats identified in Section 4.1 and enhance the privacy and security for the centralized privacy model of a contact tracing system. The security requirements for responding to the security and privacy threats mentioned above are as follows.

- **SR1. Processing based on legal and regulation grounds:** There are seven types of PII data to collect (location data, personally identifiable information, medical and prescription records, immigration records, credit/debit and prepaid card transaction data, public transportation use records, and CCTV images); however, only necessary information should be collected, and the consent of the data subject should be checked. In this case, if it is required or permitted by law, the above may not be considered. When providing to a third party, it is necessary to identify whether there is any personal information to be provided and to review what kind of disadvantage there is to the information subject if the information subject does not agree.
- **SR2. Minimizing data collection:** Obtaining the consent of the data subject is of the highest priority. PII should be collected and used only within the scope of the agreed purpose (conclusion and implementation), and multiple pieces of PII with similar characteristics should not be collected for the same purpose. Information automatically generated in the process of using a website, such as cookies, should be collected minimally.
- **SR3. Ensuring individual rights of PII:** When the PII controller collects PII with the consent of the data subject, the following needs to be ensured: (1) the contents of the consent, (2) the fact that the data subject has the right to refuse consent, and (3) the contents of the disadvantage if there is a disadvantage due to the refusal of consent should be specified specifically. In addition, the consent of the data subject is premised on a substantive right of choice. Even if the information subject does not agree to the optional items, a service provider cannot refuse to provide the service [35].
- **SR4. Strong access control:** Data access rights for each component of the PII processing model should be set, and a system should be established so that the data can be accessed according to the level of authority.
- **SR5. Use of a strong encryption mechanism:** Access control and restriction on PII, encryption technology, or equivalent measures that can safely store and transmit PII should be applied.
- **SR6. Providing data integrity:** In the process of sending PII (data sharing), passwords, bio information, and unique identification information must be encrypted before transmission. They must be encrypted and stored. The encryption technique used when data transmission is transmitted must be using a symmetric key encryption algorithm or a public key encryption algorithm; when data are stored in the system, they must be stored using a one-way encryption algorithm such as a hash function.
- **SR7. Data backup for availability:** Due to the characteristics of PII, media such as a tape or external USB are judged to be inappropriate, so it is considered appropriate to store data on media such as a disk or in the cloud. Even when backing up PII, it is necessary to store encrypted data rather than plain text; in the case of data storage, data should be located on the internal network rather than on an external network or DMZ.
- **SR8. Use of a complete data-deletion mechanism:** After the PII controller achieves the purpose for the user's PII, when the retention and use period ends, a PII controller should destroy the PII without delay [22,36]. When PII is destroyed, it must be destroyed in a way that cannot be restored or reproduced.
- **SR9. Data processing only for the intended purposes:** In the case of establishing an internal management plan to block the use of data for anything other than the intended purpose and requesting an external party to process PII, the purpose for which the PII processor can process PII must be determined in advance.

- **SR10. Prevention from inside attacks:** Since it is impossible to apply security policies to internal attackers with a firewall, security procedures should be clarified and checked regarding whether they are being continuously implemented.
- **SR11. Use of de-identification techniques:** The appropriateness of data de-identification measures should be evaluated to ensure that necessary identification information is used after de-identification measures. Measures to monitor the possibility of re-identification of de-identified information should be taken, and, when outsourcing the processing of pseudonymous information, the contract should include notification of the prohibition of re-identification, restrictions on re-supply/re-entrustment, and notification of the risk of re-identification.
- **SR12. Use of a strong end-to-end encryption protocol with authentication such as SSH (Secure Shell):** The latest version of the secure and secure tunneling protocol should be made sure to provide encrypted communication sessions. SR12 can counter ST2 and ST7 threats.
- **SR13. Use of data anonymization:** All information collected for tracking is converted into anonymous information and announced. SR13 can counter PT7 threats.
- **SR14. Providing various notification methods:** If a terminal is lost, people will not be able to receive the information related to it when they become a close contact. Therefore, various notification methods, such as e-mail notification, etc., for the recipient, that is, the closer contact, should be improved.

Table 7 shows the 1:1 or 1:N mapping data of the model's security and privacy threats for the corresponding security requirements.

Table 7. Security–privacy threats and security requirements mapping list.

Security and Privacy Threats	Security Requirements													
	SR1	SR2	SR3	SR4	SR5	SR6	SR7	SR8	SR9	SR10	SR11	SR12	SR13	SR14
ST1				○	○									
ST2						○						○		
ST3				○			○							
ST4								○						
ST5											○			
ST6										○				
ST7												○		
ST8														○
PT1									○					
PT2				○										
PT3	○		○											
PT4	○	○							○					
PT5	○		○											
PT6											○			
PT7													○	

ST: security threats, PT: privacy threats. ○ means that relevance issue is addressed.

5. Conclusions

This study demonstrates a centralized contact tracing system for infectious diseases focusing on a case study of the Republic of Korea, and it derives the security and privacy threats to that system. In addition, we identify corresponding security requirements for each threat one by one. Thirteen security requirements are provided to mitigate the threats for the system.

The centralized contact tracing system identifies subjects who have close contact with confirmed cases in specific partitioned spaces such as restaurants, offices, theatres, etc., based upon the substantial data control of PII. Hence, there need to be considerations such as scanning QR codes, including the PII of a subject, calling a designated official phone number provided by government offices to record subjects' visits, and obligatorily writing down the names and phone numbers on provided, formatted papers when subjects visit specific places.

It means that centralized models can undermine PII sovereignty over data. Since the privacy model of a centralized contact tracing system can have specific security requirements against threats that occur when the legal basis for PII collection is insufficient, the consent of the PII subject is not obtained in the process of data collection. Therefore, a strengthened collection process should be established to secure a legal basis for collecting data from PII subjects and prevent the invasion of the privacy of PII subjects, to utilize the centralized contact tracing system securely. In addition, more secure use of the centralized contact tracing system can be promoted by considering the threats identified in this paper and the corresponding security and privacy requirements.

As a future work, in-depth and intensive comparison studies regarding various types of contact tracing systems, such as centralized, decentralized, and hybrid-based contact tracing systems, will be carried out in terms of their security and privacy aspects.

Author Contributions: Conceptualization, S.P. and H.-Y.Y.; methodology, S.P. and H.-Y.Y.; validation, S.P. and H.-Y.Y.; formal analysis, S.P. and H.-Y.Y.; investigation, S.P. and H.-Y.Y.; resources, S.P. and H.-Y.Y.; writing—original draft preparation, S.P. and H.-Y.Y.; writing—review and editing, S.P. and H.-Y.Y.; visualization, S.P. and H.-Y.Y.; supervision, S.P. and H.-Y.Y.; project administration, S.P. and H.-Y.Y.; funding acquisition, H.-Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by an Institute of Information and Communications Technology Planning and Evaluation (IITP) of Korea grant, funded by the Ministry of Science and ICT of Korea under grant number 2021-0-00112.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shahroz, M.; Ahmad, F.; Younis, M.S.; Ahmad, N.; Kamel Boulos, M.N.; Vinuesa, R.; Qadir, J. COVID-19 digital contact tracing applications and techniques: A review post initial deployments. *Transp. Eng.* **2021**, *5*, 100072. [CrossRef]
2. World Health Organization. Available online: https://innov.afro.who.int/global-innovation/robert-robust-and-privacy-preserving-proximity-tracing-protocol-1827 (accessed on 25 September 2022).
3. Github. Available online: https://github.com/DP-3T/documents (accessed on 25 September 2022).
4. Hogan, K.; Macedo., B.; Macha., V.; Barman., A.; Jiang., X. Contact Tracing Apps: Lessons Learned on Privacy, Autonomy, and the Need for Detailed and Thoughtful Implementation. *JMIR Med. Inform.* **2021**, *9*, 27449. [CrossRef] [PubMed]
5. Boutet., A.; Castelluccia., C.; Cunche., M.; Lauradou., C.; Roca., V.; Baud., A.; Raverdy., P. Desire: Leveraging the Best of Centralized and Decentralized Contact Tracing Systems. *Digit. Threat. Res. Pract.* **2022**, *3*, 1–20. [CrossRef]
6. Giuseppe, G.; Tim, H.; Davy, P.; Wouter, J.; Aysajan, A.; Mustafa, A.M. PIVOT: PrIVate and effective cOntact Tracing. *IEEE Internet Things J.* **2021**, *9*, 22466–22489. [CrossRef]
7. Johannes, M.; Kretzschmar, M. Contact tracing—Old models and new challenges. *Infect. Dis. Model.* **2021**, *6*, 222–231. [CrossRef]
8. Park, S.; Choi, G.J.; Ko, H. Information Technology–Based Tracing Strategy in Response to COVID-19 in South Korea—Privacy Controversies. *JAMA Netw. Open* **2020**, *323*, 2129–2130. [CrossRef] [PubMed]
9. Jian, S.-H.; Cheng, H.-Y.; Huang, X.-T.; Liu, D.-P. Contact tracing with digital assistance in Taiwan's COVID-19 outbreak response. *Intern. J. Infect. Dis.* **2020**, *101*, 348–352. [CrossRef] [PubMed]
10. Garrett, P.M.; Wang, Y.-W.; White, J.P.; Kashima, Y.; Dennis, S.; Yang, C.-T. High acceptance of COVID-19 Tracing Technologies in Taiwan: A nationally representative survey analysis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3323. [CrossRef] [PubMed]
11. Yanji, P.; Dongyue, C. Privacy Analysis and Comparison of Pandemic Contact Tracing Apps. *KSII Trans. Internet Inf. Syst.* **2021**, *15*, 4145–4162. [CrossRef]
12. Cho, H.; Ippolito, D.; Yu, Y.W. Contact Tracing Mobile Apps for COVID-19: Privacy Considerations and Related Trade-offs. *arXiv* **2020**, arXiv:2003.11511. [CrossRef]
13. UK Health Security Agency. NHS COVID-19 App. 13 May 2022. Available online: https://www.gov.uk/government/collections/nhs-covid-19-app (accessed on 23 September 2022).
14. Liang, F. COVID-19 and Health Code: How Digital Platforms Tackle the Pandemic in China. *Soc. Media Soc.* **2020**, *6*, 2056305120947657. [CrossRef] [PubMed]

15. Bay, J.; Kek, J.; Tan, A.; Hau, C.S.; Yongquan, L.; Tan, J.; Quy, T.A. *BlueTrace: A Privacy-Preserving Protocol for Community-Driven Contact Tracing across Borders*; Government Technology Agency: Singapore, 2020.
16. PEPP-PP. PEPP-PT Documentation. 2020. Available online: https://github.com/pepp-pt/pepp-pt-documentation (accessed on 23 September 2022).
17. Troncoso, C.; Payer, M.; Hubaux, J.P.; Salathé, M.; Larus, J.; Bugnion, E.; Lueks, W.; Stadler, T.; Pyrgelis, A.; Antonioli, D.; et al. Decentralized Privacy-Preserving Proximity Tracing. *arXiv* **2020**, arXiv:2005.12273. [CrossRef]
18. Small, L.S.; John, H.; Matt, H.; Nathaniel, L. Summary of Bluetooth Contact Tracing Options. 2020. Available online: https://www.dta.mil.nz/assets/Publications/Bluetooth-Contact-Tracing-Options.pdf (accessed on 23 September 2022).
19. Google. Exposure Notifications: Help Slow the Spread of COVID-19, with One Step on Your Phone. 2020. Available online: https://www.google.com/covid19/exposurenotifications/ (accessed on 23 September 2022).
20. Kintvedt, M.N. COVID-19 Tracing Apps as a Legal Problem: An Investigation of the Norwegian 'Smittestopp' App. *Oslo Law Rev.* **2021**, *8*, 69–87. [CrossRef]
21. Jhanwar, M.P.; Sarkar, S. Phyct: Privacy Preserving Hybrid Contact Tracing. *IACR Cryptol. ePrint Arch.* **2020**, *2020*, 793.
22. Development Asia. COVID-19 Smart Management System (SMS) in Korea. Available online: https://events.development.asia/system/files/materials/2020/04/202004-covid-19-smart-management-system-sms-republic-korea.pdf (accessed on 25 September 2022).
23. Reliable Ministry of Government legislation Korean Law Information Center. Infectious Disease Control and Prevention Act. Available online: https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EA%B0%90%EC%97%BC%EB%B3%91%EC%9D%98%EC%98%88%EB%B0%A9%EB%B0%8F%EA%B4%80%EB%A6%AC%EC%97%90%EA%B4%80%ED%95%9C%EB%B2%95%EB%A5%A0 (accessed on 25 September 2022).
24. ICT Standardization Committee. TTAK.KO-12.0376:Privacy Protection Guidelines for Infectious Diseases Control and Prevention. Available online: https://committee.tta.or.kr/data/standard_view.jsp?order=t.publish_date&by=desc&nowPage=1&pk_num=TTAK.KO-12.0376&commit_code=TC5 (accessed on 25 September 2022).
25. Jeon, H. Official Operation of the 'COVID-19 Epidemiological Investigation System' on the 26th and Identify the Movement of Confirmed Patients. 2020. Available online: https://www.news1.kr/articles/?3884765 (accessed on 24 September 2022).
26. LX Spatial Information Research Institute. Available online: https://lxsiri.re.kr/frt/biz/bbs/selectBoardArticle.do?bbsId=BBSMSTR_000000000221&nttId=7323 (accessed on 27 April 2022).
27. International Organization for Standardization (ISO). ISO/IEC 29100:2011; Information Technology—Security Techniques—Privacy Framework. Available online: https://www.iso.org/standard/45123.html (accessed on 25 September 2022).
28. Korea Disease Control and Prevention Agency. Available online: https://www.kdca.go.kr/contents.es?mid=a20301110100 (accessed on 28 April 2022).
29. Korea Policy Briefings. Available online: https://www.korea.kr/news/policyNewsView.do?newsId=148895400#sitemap-layer (accessed on 25 September 2022).
30. Ministry of Land, Infrastructure and Transport (MOLIT). Available online: http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&CONT_SEQ=359845 (accessed on 25 September 2022).
31. International Telecommunication Union(ITU-T). ITU-T X.1121: Framework of Security Technologies for Mobile End-To-End Data Communications. Available online: https://www.itu.int/rec/T-REC-X.1121/en (accessed on 24 September 2022).
32. Google Cloud. Available online: https://cloud.google.com/blog/products/identity-security/taking-charge-of-your-data-understanding-re-identification-risk-and-quasi-identifiers-with-cloud-dlp (accessed on 1 May 2022).
33. LIDDUN. Available online: https://www.linddun.org/linddun (accessed on 12 February 2022).
34. Khan, S.A. A STRIDE Model based Threat Modelling using Unified and-Or Fuzzy Operator for Computer Network Security. *Int. J. Comput. Netw. Technol.* **2017**, *5*, 13–20. [CrossRef] [PubMed]
35. Lee, I.; Keh., J.S. Cross-Border Transfers of Personal Data and Practical Implications. *J. Korean L.* **2017**, *17*, 33–52.
36. Korea Legislation Research Institute. Personal Information Protection Act. Available online: https://elaw.klri.re.kr/eng_service/lawView.do?hseq=53044&lang=ENG (accessed on 15 July 2022).

big data and cognitive computing

Article

Argumentation-Based Query Answering under Uncertainty with Application to Cybersecurity

Mario A. Leiva [1,2], Alejandro J. García [1,2], Paulo Shakarian [3] and Gerardo I. Simari [1,2,3,*]

[1] Department of Computer Science and Engineering, Universidad Nacional del Sur (UNS), Bahia Blanca 8000, Argentina
[2] Institute for Computer Science and Engineering (UNS–CONICET), Bahia Blanca 8000, Argentina
[3] School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ 85281, USA
[*] Correspondence: gis@cs.uns.edu.ar

Abstract: Decision support tools are key components of intelligent sociotechnical systems, and their successful implementation faces a variety of challenges, including the multiplicity of information sources, heterogeneous format, and constant changes. Handling such challenges requires the ability to analyze and process inconsistent and incomplete information with varying degrees of associated uncertainty. Moreover, some domains require the system's outputs to be explainable and interpretable; an example of this is cyberthreat analysis (CTA) in cybersecurity domains. In this paper, we first present the P-DAQAP system, an extension of a recently developed query-answering platform based on defeasible logic programming (DeLP) that incorporates a probabilistic model and focuses on delivering these capabilities. After discussing the details of its design and implementation, and describing how it can be applied in a CTA use case, we report on the results of an empirical evaluation designed to explore the effectiveness and efficiency of a possible world sampling-based approximate query answering approach that addresses the intractability of exact computations.

Keywords: intelligent sociotechnical systems; human-in-the-loop computing; structured probabilistic argumentation; cybersecurity

Citation: Leiva, M.A.; García, A.J.; Shakarian, P.; Simari, G.I. Argumentation-Based Query Answering under Uncertainty with Application to Cybersecurity. *Big Data Cogn. Comput.* **2022**, *6*, 91. https://doi.org/10.3390/bdcc6030091

Academic Editors: Peter R.J. Trim and Yang-Im Lee

Received: 26 July 2022
Accepted: 22 August 2022
Published: 26 August 2022

1. Introduction

Sociotechnical systems [1] are an important class of applications of artificial intelligence (AI) tools, since many deployments of technology built on their foundations are at the core of decision processes at the individual and the organizational levels. An inherent problem in this area is that of explainability and interpretability, topics that were not central in earlier "AI booms" characterized by expert systems and rule-based models. The issues underlying this problem are within the domain of explainable AI (XAI) [2], which is now widely recognized as a crucial feature for the practical deployment of AI models [3]. The importance of this aspect can be appreciated by pointing to the Explainable Artificial Intelligence (XAI) program launched by the Defence Advanced Research Projects Agency (DARPA) [4], which aims to create a set of new artificial intelligence techniques that allow for end users to understand, properly trust, and effectively manage the emerging generation of artificial intelligence systems [5]. The danger is that complex black-box models (some of which can comprise hundreds of layers and millions of parameters) [6] are increasingly used for important predictions in critical contexts, and these models generate outputs that may not be justified or simply do not allow for detailed explanations of their behavior [4]. In this direction, recent work focused on addressing these problems from different points of view [7–9]. In this paper, we focus on cybersecurity as a salient example of a sociotechnical domain [10] in which the availability of explanations that support the output of a model are crucial. Transparency, together with a human-in-the-loop (HITL) scheme, leads to more robust decision-making processes whose results can be trusted by users [8]. Achieving this is challenging, since many domains involve information arriving from multiple heterogeneous sources with different levels of uncertainty due to gaps in knowledge (incompleteness), overspecification (inconsistency), or inherent uncertainty.

In cybersecurity domains, a clear example is the task of real-time security analysis, a complex process in which many uncertain factors are involved, given that analysts must deal with the behavior of different actors and entities, the dynamic nature of exploits, and the fact that the observations of potentially malicious activities are limited. Cyberthreat analysis (CTA) [11] is a highly technical intelligence problem in which (human) analysts take into consideration multiple sources of information, with possibly varying degrees of confidence or uncertainty, with the goal of gaining insight into events of interest that may represent a threat to a system. When building AI tools to assist such a process, knowledge engineers face the challenge of leveraging uncertain knowledge in the best possible way [12]. Due to the nature of these analytical processes, an automated reasoning system with human-in-the-loop capabilities would be best suited for the task. Such a system must be able to accomplish several goals, among which we distinguish the following main capabilities [13]: (i) reason about evidence in a formal, principled manner; (ii) consider evidence associated with probabilistic uncertainty; (iii) consider logical rules that allow for the system to draw conclusions on the basis of certain pieces of evidence and iteratively apply such rules; (iv) consider pieces of information that may not be compatible with each other, deciding which the most relevant are; and (v) show the actual status of the system on the basis of the above-described features, and provide the analyst with the ability to understand why an answer is correct, and how the system arrives at that conclusion (i.e., *explainability and interpretability*). In this context, there is a specific literature to the study of techniques and methodologies for providing explanations in cybersecurity domains [14–17]. The model that we develop in this work is based on *argumentation-based reasoning*, an approach that is designed to mimic the way humans with which rationally arrive at conclusions by analyzing arguments for and against them, and is especially well-suited for accommodating desirable features, such as reasoning about possibly uncertain evidence in a principled manner, handling pieces of information that may not be compatible with each other, and showing the actual status of the system to analysts along with the ability to understand why an output is produced.

Contributions. We contribute to the area of intelligent systems applied to cybersecurity in the following ways:

- A use case for the application of a structured probabilistic argumentation model (DeLP3E) [18] based on publicly available cybersecurity datasets.
- Design of the P-DAQAP framework, an extension of DAQAP [19], to work with DeLP3E, and the proposal of different classes of queries in the context of applications related to CTA.
- A preliminary empirical evaluation of an approximation algorithm for probabilistic query answering in P-DAQAP, showing the potential for the system to scale to nontrivial problem sizes, arriving at solutions efficiently and effectively.

To the best of our knowledge, this is the first system of its kind. In particular, being able to consider the internal structure of arguments allows for the platform to be extended to work with other defeasible argumentation formalisms, and offers greater transparency to adapt classical approaches that do not consider probabilistic information.

2. Preliminaries

Tools developed in the area of argumentation-based reasoning offer the possibility of analyzing complex and dynamic domains by studying the arguments for and against a conclusion. Specifically, *defeasible argumentation* leverages models that contain inconsistency, evaluating arguments that support contradictory conclusions and deciding which ones to keep [20]. An argument supports a conclusion from a set of premises [20]; a conclusion `C` constitutes a piece of tentative information that an agent is willing to accept. If the agent then acquires new information, conclusion `C`, along with the arguments that support it, could be invalidated. The validity of a conclusion `C` is guaranteed when there is an argument that provides justification for `C` that is undefeated. This process involves the construction of an argument $\mathcal{A}$ for C, and the analysis of counterarguments that are possible defeaters of $\mathcal{A}$; as these defeaters are arguments, it must be verified that they are not themselves defeated. There are several formalisms that are based on this idea, such as ABA [21], ASPIC+ [22], *defeasible logic programming* (DeLP) [23], and *deductive argumentation* [24], which consider the structure of the arguments that model a discussion. The DAQAP platform [19] on which the presented system is based uses DeLP as its central formalism. We now briefly present the necessary background, starting with DeLP and its probabilistic extension.

2.1. Defeasible Logic Programming (DeLP)

DeLP combines logic programming and defeasible argumentation. A DeLP program $\mathcal{P}$, also denoted as (Π, Δ), is a set of facts and strict rules (Π), and defeasible rules (Δ). *Facts* are ground literals representing atomic information (or its negation using strong negation "$\sim$"), *strict rules* represent nondefeasible information, and *defeasible rules* represent tentative information. Here, we consider the extension that incorporates *presumptions* to set Δ, which can be thought of as a kind of defeasible fact [25].

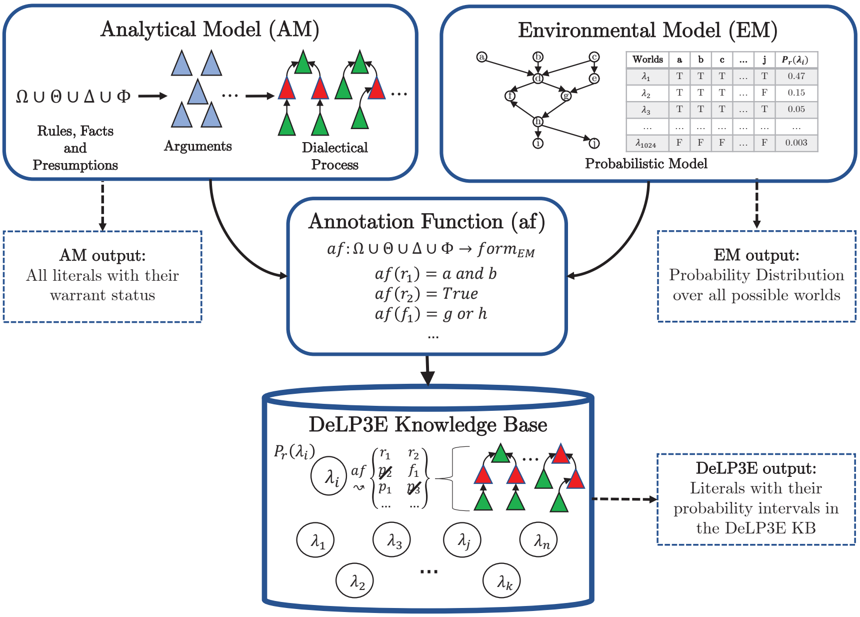
The dialectical process used in deciding which information prevails as *warranted* involves the construction and evaluation of arguments that either support or interfere with the query under analysis. An *argument* $\mathcal{A}$ is a minimal set of defeasible rules that, along with the set of strict rules and facts, are not contradictory and derive a certain conclusion α, denoted as $\langle \mathcal{A}, \alpha \rangle$. Arguments supporting the answer for a query can be organized using *dialectical trees*. A query is issued to a program (Π, Δ) in the form of a ground literal α.

A literal α is *warranted* if there exists a nondefeated argument $\mathcal{A}$ supporting α. To establish if $\langle \mathcal{A}, \alpha \rangle$ is a nondefeated argument, *defeaters* for $\langle \mathcal{A}, \alpha \rangle$ are considered, *i.e.*, *counterarguments* that by some criteria are preferred to $\langle \mathcal{A}, \alpha \rangle$. An argument $\mathcal{A}_1$ is a counterargument for $\mathcal{A}_2$ *iff* $\mathcal{A}_1 \cup \mathcal{A}_2 \cup \Pi$ is contradictory. Given a preference criterion, and an argument $\mathcal{A}_1$ that is a *defeater* for $\mathcal{A}_2$, $\mathcal{A}_1$ is called a *proper defeater* if it is preferred to $\mathcal{A}_2$, or a *blocking defeater* if it is equally preferred or is incomparable with $\mathcal{A}_2$. Since there may be more than one defeater for a particular argument, many acceptable argumentation lines could arise from one argument, leading to a tree structure. This is called a *dialectical tree* because it represents an exhaustive dialectical analysis for the argument in its root; every node (except the root) represents a defeater of its parent, and leaves correspond to nondefeated arguments. Each path from the root to a leaf corresponds to a different acceptable argumentation line. A dialectical tree provides a structure for considering all possible acceptable argumentation lines that can be generated for deciding whether an argument is defeated.

Given a literal α and an argument $\langle \mathcal{A}, \alpha \rangle$ from a program $\mathcal{P}$, to decide whether α is warranted, every node in the tree is recursively marked as *D* (*defeated*) or U (*undefeated*), obtaining a marked dialectical tree $\mathcal{T}_{\mathcal{P}}(\mathcal{A})$: (1) all leaves in $\mathcal{T}_{\mathcal{P}}(\mathcal{A})$ are marked as "U"s; and (2) let $\mathcal{B}$ be an inner node of $\mathcal{T}_{\mathcal{P}}(\mathcal{A})$; then, $\mathcal{B}$ is marked as U *iff* every child of $\mathcal{B}$ is marked as D. Thus, node $\mathcal{B}$ is marked as D *iff* it has at least one child marked as U. Given an argument $\langle \mathcal{A}, \alpha \rangle$ obtained from $\mathcal{P}$, if the root of $\mathcal{T}_{\mathcal{P}}(\mathcal{A})$: is marked as U, then $\mathcal{T}_{\mathcal{P}}(\mathcal{A})$ *warrants* α, and α is *warranted* from $\mathcal{P}$. The DeLP interpreter takes a program $\mathcal{P}$ and a DeLP query L, and returns one of the following four possible answers: YES if L is warranted from $\mathcal{P}$, NO if the complement of L regarding strong negation is warranted from $\mathcal{P}$, UNDECIDED if neither L nor its complement are warranted from $\mathcal{P}$, or UNKNOWN if L is not in the language of the program $\mathcal{P}$.

2.2. Probabilistic DeLP: DeLP3E Framework

We now provide a brief introduction to DeLP3E; for full details, we refer the reader to [18]. A DeLP3E KB $P = (AM, EM, af)$ consists of three parts that correspond to *two separate models of the world*, and a function linking the two; these components are illustrated in Figure 1.


Analytical Model (AM)
$\Omega \cup \Theta \cup \Delta \cup \Phi$
Rules, Facts and Presumptions
Arguments
Dialectical Process
Environmental Model (EM)
Probabilistic Model
Annotation Function (af)
$af: \Omega \cup \Theta \cup \Delta \cup \Phi \rightarrow form_{EM}$
$af(r_1) = a \text{ and } b$
$af(r_2) = True$
$af(f_1) = g \text{ or } h$
...
AM output:
All literals with their warrant status
EM output:
Probability Distribution over all possible worlds
DeLP3E Knowledge Base
$P_r(\lambda_i)$
λ_i
af
r_1 r_2 p_2 f_1 p_1 p_3 ... ...
λ_1 λ_3 λ_j λ_n λ_2 ... λ_k
DeLP3E output:
Literals with their probability intervals in the DeLP3E KB
a b c d e f g h i j

Worlds	a	b	c	...	j	$P_r(\lambda_i)$
λ_1	T	T	T	...	T	0.47
λ_2	T	T	T	...	F	0.15
λ_3	T	T	T	...	T	0.05
...	...	...	...	...	...	...
λ_{1024}	F	F	F	...	F	0.003

Figure 1. Overview of the DeLP3E framework.

The *environmental model* (EM) is used to describe background knowledge that is probabilistic in nature, while the *analytical model* (AM) is used to analyze competing hypotheses that can account for a given phenomenon. The EM *must be consistent*, while the AM allows for contradictory information as the system must have the capability to reason about competing explanations for a given event. In general, the EM contains knowledge such as evidence, intelligence reporting, or uncertain knowledge about actors, software, and systems, while the AM contains elements that the analyst can leverage on the basis of information in the EM. AMs correspond to DeLP programs, while EMs in this paper are abstracted away, assuming that the well-known Bayesian network model is used.

Finally, the third component is the *annotation function*, which links components in the AM with conditions over the EM (the conditions under which statements in the AM can potentially be true). We use G_{EM} to denote the sets of all ground atoms for the EM; here, we concentrate on subsets of ground atoms from G_{EM}, called *worlds*. Atoms that belong to the set are *true* in the world, while those that do not are *false* (Therefore, there are $2^{|G_{EM}|}$ possible worlds in the EM). This set is denoted with $\mathcal{W}_{EM}$. Logical formulas arise from the combination of atoms using the traditional connectives ($\wedge$, $\vee$, and $\neg$); we use $form_{EM}$ to denote the set of all possible (ground) formulas in the EM. Annotation functions then assign formulas in $form_{EM}$ to components in the AM to indicate the conditions (probabilistic events) under which they hold. In this way, each world $\lambda \in \mathcal{W}_{EM}$ *induces* a subset of the AM, comprised of all elements whose annotations are satisfied by λ; for DeLP3E program P, we denote the subset of the AM induced by λ with $P_{AM}(\lambda)$ (cf. Figure 1). Exact probabilistic query answering is carried out via Algorithm 1.

Algorithm 1: Exact probabilistic query answering

```
def compute_answer(query)
    Data: $P = (AM, EM, af)$
    Result: $[\ell, u]$
    begin
        Initialize $\ell = 0$ and $u = 1$            /* the limits of the interval */
        for EM worlds $\lambda_i$ do
            Compute the induced AM subprogram $P_{AM}(\lambda_i)$
            if the query is warranted in that program then
                $\ell \leftarrow \ell + \Pr(\lambda_i)$
            else if the negation of the query is warranted then
                $u \leftarrow u - \Pr(\lambda_i)$
        end
        return $[\ell, u]$
    end
end
```

Since the number of worlds in $\mathcal{W}_{EM}$ is exponential in the number of EM random variables, this procedure quickly becomes intractable. However, a *sound approximation* of the exact interval can be obtained by simply selecting a subset of $\mathcal{W}_{EM}$ and executing the same procedure. We refer to this algorithm as approximate query answering via *world sampling*. It is easy to see that this approximation scheme is sound since it always yields intervals $[\ell', u'] \subseteq [\ell, u]$. Section 5 is dedicated to studying the effectiveness and efficiency of this approach.

A Simple Illustrative Example

In order to clearly illustrate the model and query-answering procedure in DeLP3E, we present the following simple example of knowledge base $P = (AM, EM, af)$:

Analytical Model

$\theta_1 : L_1$
$\theta_2 : L_2$
$\theta_3 : {\sim}L_1$

Annotation Function

$af(\theta_1) : a \wedge \neg b$
$af(\theta_2) : b$
$af(\theta_3) : b$

Environmental Model

World	a	b	$P_r(\lambda_i)$
λ_1	T	T	0.25
λ_2	T	F	0.20
λ_3	F	T	0.05
λ_4	F	F	0.50

We have an AM consisting of three literals, an EM consisting of two variables, and an annotation function that relates these two models; suppose we query for the literal L_1. To compute the exact probability interval, we go world by world as described above, generating the corresponding subprogram and querying each one of them for the status of the query. Lastly, in order to arrive at the probability interval with which L_1 is warranted in P, we keep track of the probability of the worlds where the query is warranted (for the lower limit of the interval) and the probability of the worlds where the *complement* of the query is warranted (for the upper limit). In our example, the result for query L_1 is $[0.20, 0.70]$; the details of this calculation are as follows:

- **Subprograms induced in each possible world:**
  - $P_{AM}(\lambda_1) = \{L_2, {\sim}L_1\}$
  - $P_{AM}(\lambda_2) = \{L_1\}$
  - $P_{AM}(\lambda_3) = \{L_2, {\sim}L_1\}$
  - $P_{AM}(\lambda_4) = \{\emptyset\}$

  Query L_1 is, thus, clearly warranted only in world λ_2, while its complement (${\sim}L_1$) is warranted in λ_1 and λ_3.
- **Probability interval calculation:**

$$\left[\ell = \sum P_r(\lambda_2), \quad u = 1 - \sum_{i=1,3} P_r(\lambda_i)\right]$$

- **Result:** $0.20 \leq P_r(L_1) \leq 0.70$

The resulting probability interval represents *two kinds of uncertainty*: the first, called *probabilistic* uncertainty, arises from the environmental model since we have a probability distribution over possible worlds; the second, *epistemic* uncertainty, arises from the fact that we we generally have a probability interval instead of a point probability, which happens when there are worlds in which neither the query nor its complement are warranted (as is the case of world λ_4 above).

Having presented the preliminary concepts, in the next section, we illustrate the application of DeLP3E in a cybersecurity domain.

3. Cyberthreat Analysis with DeLP3E

We now present a use case leveraging several datasets developed and maintained by the MITRE Corporation (a not-for-profit organization that works with governments, industry, and academia) and National Institute of Standards and Technology (NIST) (MITRE datasets: ATT&CK (https://attack.mitre.org, accessed on 21 August 2022), CAPEC (https://capec.mitre.org, accessed on 21 August 2022), and CWE (https://cwe.mitre.org, accessed on 21 August 2022). NIST manages the National Vulnerability Database (NVD) (https://nvd.nist.gov, accessed on 21 August 2022) that includes CVE and CPE). Figure 2 shows an overview of our approach. We first describe the basic components and then show how the DeLP3E components are specified, along with two queries for addressing specific problems in the CTA domain.

The ATT&CK model is a curated knowledge base and model geared towards adversarial behavior in cybersecurity settings; it contains information on the various phases of an attack and the platforms that are most commonly targeted. The behavioral model consists of several core components:

(i) *Tactics*, denoting short-term tactical adversary goals during an attack.
(ii) *Techniques*, describing the means by which adversaries achieve tactical goals.
(iii) *Subtechniques,* describing more specific means at a lower level than that of techniques by which adversaries achieve tactical goals.
(iv) Documented *adversary usage* of techniques, their procedures, and other metadata.

The supporting datasets provide information on *attack patterns* (Common Attack Pattern Enumeration and Classification—CAPEC), software and hardware *weakness types* (Common Weakness Enumeration—CWE), and the *National Vulnerability Database* (NVD). The latter is a rich repository of data; here, we distinguish two subsets including data about *vulnerabilities* (Common Vulnerabitlities and Exposures—CVE) and *platforms* (Common Platform Enumeration—CPE).

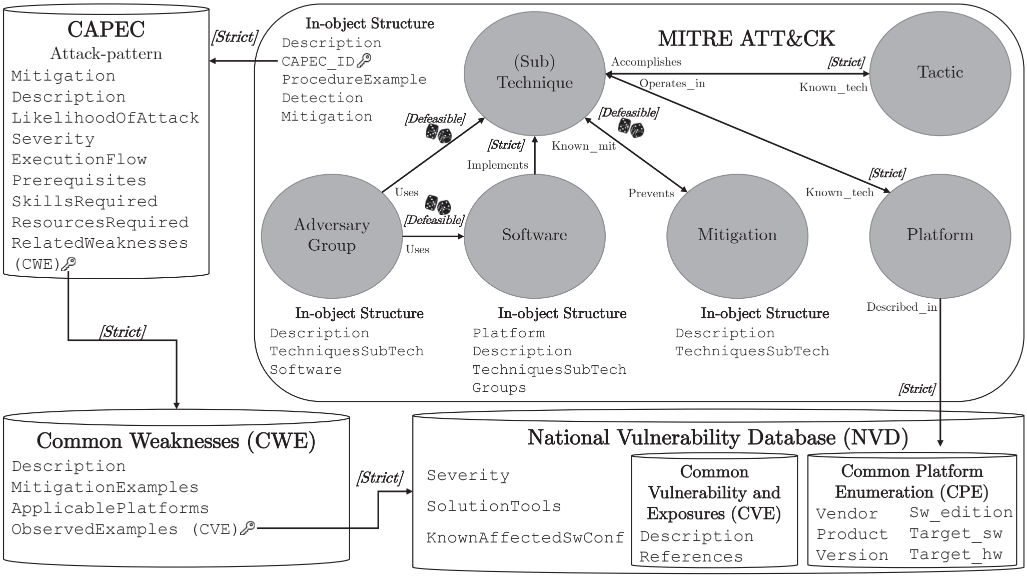


Figure 2. Designing a DeLP3E KB for cyberthreat analysis from a variety of publicly available cyber security datasets.

Figure 2 shows the information provided by each dataset, and how they are related to each other via foreign keys. For instance, attack techniques included in ATT&CK link to entries in CAPEC, which in turn link to CWE and NVD. We augmented this structure with two features towards deriving a DeLP3E KB. First, we labeled connections between datasets (and components within ATT and CK) with either "[*strict*]" or "[*defeasible*]", indicating the type of knowledge being encoded. For instance, observed examples of a weakness included in CWE are linked to CVEs included in the NVD as strict, since this is well-established knowledge. On the other hand, mitigation strategies are linked to techniques as defeasible knowledge, since the relationship between the two is tentative in nature. The second feature, which appears in the figure as a small icon depicting a pair of dice, indicates relationships that are subject to *probabilistic events*. For the purposes of this use case, we label all defeasible relations in this way.

We used all this information to create the AM, EM, and annotation function, and create a DeLP3E KB; an introductory example is shown in Listing 1. On the left-hand side, we have the elements of the AM that can be used to create arguments for and against conclusions; for instance:

$$\langle \mathcal{A}_1,\ \mathit{tech\_in\_use(account\_discovery)} \rangle, \text{ with}$$
$$\mathcal{A}_1 = \{\delta_3,\ \theta_1(\mathit{adv\_group}(\mathit{apt}29))\}$$
$$\langle \mathcal{A}_2,\ {\sim}\mathit{impl\_techsub(os\_credential\_dumping)} \rangle, \text{ with}$$
$$\mathcal{A}_2 = \{\delta_6,\ \delta_1(\mathit{prev\_techsub(os\_credential\_dumping)}),$$
$$\phi_1(\mathit{mitigation(credential\_access\_protection)})\}.$$

Listing 1. *Left*: DeLP program that comprises the AM. *Right*: Annotation function.

Θ θ_1 : *adv_group(G)*
θ_3 : *platform_available(P)*
θ_2 : *software(S)*
θ_4 : *tech_subtech(T_ST)*

Ω ω_1 : *accomp_tactic(Tactic)* $\leftarrow$ *tech_subtech(T_ST)*
ω_2 : *op_in_platform(Platform)* $\leftarrow$ *tech_subtech(T_ST)*
ω_3 : *impl_techsub(T_ST)* $\leftarrow$ *software(S)*
ω_4 : *capec_rel_weaknesses(CWE_List)* $\leftarrow$ *capec_id(T_ST)*
ω_5 : *cwe_observed(CVE_List)* $\leftarrow$ *capec_rel_weaknesses(CWE_List)*
ω_6 : *nvd_cve(Vuln_info)* $\leftarrow$ *cwe_observed(CVE_List)*
ω_7 : *known_techst(T_ST)* $\leftarrow$ *accomp_tactic(T)*
ω_8 : *known_techst(T_ST)* $\leftarrow$ *platform_available(P)*

Φ ϕ_1 : *mitigation(M)* $\prec$
ϕ_2 : *likelihoodAttack(CAPEC_ID, Value)* $\prec$

Δ δ_1 : *prev_techsub(T_ST)* $\prec$ *mitigation(M)*
δ_2 : *known_mit(M)* $\prec$ *tech_subtech(T_ST)*
δ_3 : *tech_in_use(T_ST)* $\prec$ *adv_group(G)*
δ_4 : *soft_in_use(S)* $\prec$ *adv_group(G)*
δ_5 : *pos_threat(T_ST, S)* $\prec$ *tech_in_use(T_ST), soft_in_use(S)*
δ_6 : $\sim$*impl_techsub(T_ST)* $\prec$ *prev_techsub(T_ST)*
δ_7 : *intensify_mit(M)* $\prec$ *known_mit(M), tech_in_use(T_ST),*
likelihoodAttack(T_ST, high)

$af(\phi_1) = e_1$
$af(\phi_2) = e_2$
$af(\delta_1) = e_3$
$af(\delta_2) = e_4$
$af(\delta_3) = e_5$
$af(\delta_4) = e_6$
$af(\delta_5) = e_7$
$af(\delta_6) = e_8$
$af(\delta_7) = e_9$

The former indicates that *account discovery* is used as an attack technique, since the advanced persistent threat group 29 (APT29, also known as Cozy Bear) is active and uses it. The latter refers to the use of *credential access protection* as a mitigation technique to prevent the use of *OS credential dumping*. This is a clear example of an argument that involves uncertainty, since credential access protection is not a foolproof endeavor. An example of this is the well-known *Heartbleed* vulnerability (CVE-2014-0160) that affected OpenSSL implementations, leaving them open to credential dumping. For reasons of space, in this simple example, we only label AM components with probabilistic events (e_1–e_9; elements with no annotation are simply labeled with *true*) and do not describe how they are related in the EM. One example could be to simply assume pairwise independence (as in many probabilistic database models [26]), or a Bayesian network [27], as described in Section 5.

Queries. We lastly present two queries that we revisit in the next section:

- *pos_threat*$(T1134, SO344)$:
  What is the probability that *access token manipulation* (technique T1134) uses leveraging the *Azorult* malware (software id SO344) to attack our systems?
- *intensify_mit*$(M1026)$:
  What is the probability that *privileged account management* (mitigation strategy M1026) should be deployed? M1026 mitigates T1134.

In the next two sections, we discuss the design of a software system for implementing this kind of functionalities based on DeLP3E, and a preliminary evaluation of query answering in DeLP3E via sampling techniques.

4. P-DAQAP Platform

In an early version of the platform called DAQAP [19], we developed a web-based client-server platform that offers an interface to visualize the interaction of the arguments generated from an input DeLP program via dialectical trees and graphs, as well as the abstract defeat relationships in a Dung-like graph environment. In this section, we present the extension that incorporates probabilistic reasoning based on DeLP3E knowledge bases, first briefly discussing the platform's architecture and workflow, and then moving on to presenting a set of features that could eventually support human-in-the-loop reasoning and XAI functionalities.

4.1. Architecture and Workflow

Figure 3 shows an overview of the tool's architecture and workflow that is a mock-up of a possible user interface that we are currently developing. The architecture is divided into two main modules, the front end and the back end. Within the former, there are two main sections: the DeLP and abstract argumentation section manages classical (nonprobabilistic) models and is described in detail in [19]; we focus on the DeLP3E section, which is the extension presented here. The back end is organized analogously, with the addition of three other submodules that implement the probability model (for the EM), sampling methods, and approximation algorithms.

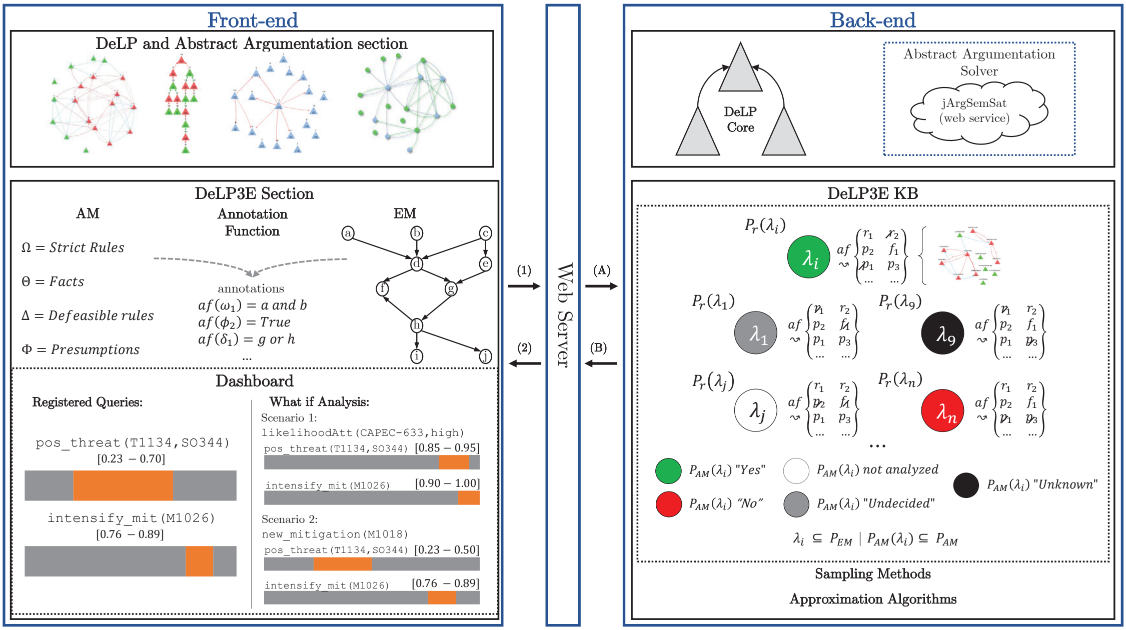


Figure 3. P-DAQAP platform architecture, including a mock-up of a dashboard for displaying query-answering results related to our use case.

Table 1 describes the workflow focused on DeLP3E tasks in the order of the steps labeled at the interaction between the two main modules in Figure 3 (1 → A → B → 2). This workflow is iterative in nature, and implements the human-in-the-loop model mentioned in Section 1. In Step B, an *anytime algorithm* approach may be applied, in which results are iteratively improved, and the user can decide when to stop the job depending on the amount of time available and/or the quality of the result currently being obtained. After Step 2, the analyst can now interact through the dashboard in response to the results received, for example by choosing to modify the DeLP3E KB, modifying the query issued in the first step, or a combination of such actions.

Table 1. P-DAQAP Workflow.

	Front-end
Step 1	Loads a DeLP3E knowledge base and specifies a *task*.
	Back-end
Step A	Web server sends the job to be executed by the Probabilistic Argumentation module.
Step B	Generate data structures and executes the job; when results become available, it returns the output data in JSON format to the web server.
	Front-end
Step 2	Client receives the response, and the data are presented to the user.

In the next section, we explore some of these functionalities, illustrating them via the use case presented in Section 3.

4.2. P-DAQAP Functionalities

We begin by describing the design of two functionalities based on our use case, which are illustrated in the *Dashboard* section of Figure 3, and then discuss the next steps to be developed.

4.2.1. Current State: *Registered Queries*

The values of a subset of the EM variables are set depending on the current state of the system (observed evidence). The analyst registers a set of queries of interest in order to monitor the associated probabilities. Consider the queries presented in Section 3; the user is interested in monitoring a possible threat and degree of application of a corresponding mitigation strategy. In Figure 3 (bottom left), we can see that in the current state the query

$$pos\_threat(T1134, SO344)$$

(referring to the probability that access token manipulation is used, levaraging Azorult) is currently warranted by the KB with probability interval $[0.23, 0.7]$; this interval is quite wide, which points to a large amount of uncertainty and lack of actionable insight.

On the other hand, the query

$$intensify\_mit(M1026)$$

(which refers to the probability that privileged account management should be deployed as a mitigation strategy) yields an interval of $[0.76, 0.89]$, which signals a high probability of the need to intensify mitigating actions associated with technique T1134.

Having this kind of insight is valuable for analysts, who can register queries regarding mitigation strategies and attack techniques of current interest. The results can inform, for instance, security alert levels and patching effort priorities for system administrators. As we discuss in Section 5, approximations can be computed whenever the cost of obtaining an exact answer is too high. In this case, the system can allow for the user to input the number of samples to be used or, given an explicit upper bound on the time that is available, decide on a budget for the sampling process.

4.2.2. "What-If" Scenarios

On the basis of the same setup as above, the user may wish to perform counterfactual reasoning, also known as *what-if* scenarios. In this case, instead of taking facts and EM variable settings from direct observations, the system allows for specifying scenarios as desired and shows the resulting probabilities.

Figure 3 illustrates this functionality with the same registered queries as before, showing how their associated probability intervals change under two scenarios. In the first, the analyst wants to know how the probabilities associated with the above queries change in case that the *token impersonation* technique is very likely to be implemented successfully, as reported by CAPEC:

$$likelihoodAttack(CAPEC\text{-}633, high).$$

The most drastic change is in the first query, which now yields a probability between 85 and 95%, while the other query's probability increases somewhat to 90–100%. This is because token impersonation (CAPEC-633) is a technique that, if it has a high likelihood of success, is directly linked to privileged account management (mitigation strategy M1026).

In the second scenario, the analyst wants to know how the probabilities would change if *user account management* is added as a new mitigation strategy ($new\_mitigation(M1018)$). Now, the query:

$$pos\_threat(T1134, SO344)$$

becomes less probable (23–50%, since the new mitigation strategy helps in preventing the T1134 technique), while for this scenario, the answer to the other query remains unchanged, since the two mitigation strategies are unrelated.

4.2.3. Next Steps: Explainability

In addition to being able to calculate query probabilities, it is possible to accompany such results with an *explanation* as to how the system arrived at that answer; explainability was recently identified as a key feature in cybersecurity domains [28]. We discuss two proposals for providing such insights into the kind of results presented in the previous sections. The first is centered on the probabilistic model (EM), while the second focuses on the rules used to derive query answers (AM).

Most Probable Scenarios. As a combination of the previous two functionalities, the system can compute a set of the k most probable scenarios given the current set of observations. In the current implementation, which uses Bayesian networks to specify the probability distribution in the EM, this set can be computed by the probabilistic model module by returning the *most probable explanations* (MPEs) of the BN given the current evidence in the EM. Then, the result of this first step can be combined with the counterfactual analysis described above and each scenario can be explored taking into account its probability of occurrence and its consequences.

Though this kind of analysis is centered on the probabilistic model, knowing the most probable scenarios is a first step towards explaining why a given query is entailed with a certain probability interval. For instance, an analyst may be interested in knowing why the upper bound is lower than expected, and being shown a high-probability scenario in which the negation of the query is entailed would be a first explanation. If further details are needed, explanations can also be derived by analyzing the rules and arguments involved in the derivations, as discussed next.

Rule-based Explanations. Another possibility is to show the arguments that support the query in the subprogram generated by a particular scenario or set of scenarios. This provides the analyst with the set of rules and facts involved in the derivation, and precisely what role they played, which may highlight the need to revise one or more of these components (for example, facts coming from an outdated data source); an approach in this direction was recently reported in [29]. Another benefit of rule-based approaches is that they can be rendered more interpretable by, for instance, using templates to translate rules into natural language, as proposed in [30]. Lastly, it is also possible to show the user minimal sets of EM elements (BN variables or worlds) that allow for the generation of supporting arguments for the query, thus pointing to the uncertain elements that play a role in the logical derivations of interest.

As a concluding remark, taking into account the general considerations of *explainable AI* approaches [2], we consider that adding a probabilistic module to a platform like DAQAP provides additional possibilities for building explanations. On the one hand, as explained in Section 2.2, the answers in P-DAQAP consist of probability intervals that represent two types of uncertainty (probabilistic and epistemic), which allows for us to provide more information about the nature of knowledge that is being processed. On the other hand, as previously detailed, it is possible to accompany the answers with different types of explanations, which demonstrates the potential of involving the probabilistic component when generating explanations. All this accompanying information provides analysts with tools that allow for them to confidently accept the obtained answer, or revise pieces of information or knowledge that do not apply to the current situation.

5. Empirical Evaluation

We now report on the results of a preliminary empirical evaluation designed to test the effectiveness and efficiency of a world sampling-based approximation to query answering in DeLP3E. We used Bayesian networks for the EM and sampled directly from the distributions they encode. The experiments focus on varying three key dimensions: *number of random variables* (which determines the number of possible worlds), *number of sampled worlds*, and the *entropy* of the probability distribution associated with the EM. Intuitively, entropy is a measure of disorder. For probability distributions, it measures how "spread out" the probability mass is over the space of possible worlds, so a low value indicates a highly concentrated mass. Extreme cases thus range from a single world having probability one, to all worlds having the same probability.

All runs were performed on a computer with an Intel Core i5-5200U CPU at 2.20GHz and 8GB of RAM under the 64-bit Debian GNU/Linux 10 OS. Probability computations were carried out using the pyAgrum (https://agrum.gitlab.io, accessed on 21 August 2022) Python library.

5.1. Experimental Setup

All problem instances (DeLP3E knowledge bases and queries) were synthetically generated to be able to adequately control the independent variables in our analysis. To obtain an instance, we first randomly generate the AM as a classical DeLP program with a balanced set of facts and rules; rule bodies and heads are generated in such a way as to ensure overlap, in order to yield nontrivial arguments (see [31] for details on such a procedure). The general design of the program generator consists of the following steps:

1. Generating the basic components on which the more complex structures are created, that is, facts and assumptions are generated first.
2. Arguments are organized in *levels*, where each level indicates the maximal number of rules used in its derivation chain until a basic element is reached.
3. Dialectical trees are generated only for top-level arguments because they have a greater number of possible points of attack, given that they have more elements in their body.

For the Bayesian networks in the EM, we randomly generated a graph on the basis of the desired number of EM variables (and a random number of edges set to the number of nodes as a maximum) using the `networkx` library (https://networkx.github.io, accessed on 21 August 2022). To control the entropy of the encoded distribution, we took each node probability table entry and randomly choose between *true* and *false*; then, we randomly assigned a probability to that outcome in the interval $[\alpha, 1]$, where α is a parameter varied in $\{0.7, 0.9\}$.

Annotation functions are lastly randomly generated by assigning to each element in the AM an element randomly chosen from the set of (possibly negated) EM variables plus "*true*" (AM elements annotated with *true* hold in all worlds).

Quality Metric. Given a probability interval $i_1 = [a, b]$, we used the following metric to gauge the quality of a sound approximation $i_2 = [c, d]$ (that is $[a, b] \subseteq [c, d]$ always holds):

$$Q_{i_1}(i_2) = \frac{1 - (d - c)}{1 - (b - a)}$$

Intuitively, this metric calculates the probability mass that is *discarded* by one interval in relation to another. The resulting value is always a real number in $[0, 1]$, where a value of zero indicates the poorest possible approximation ($[0, 1]$, which is always a sound approximation for problem instance), and a value of 1 yields the best possible approximation, which corresponds precisely with the exact interval. Thus, we generally apply this metric by using the result of the exact algorithm in the numerator and an approximation in the denominator.

5.2. Results

Figure 4a shows the average running time taken *per sample* over all configurations based on a set of 100 runs. We calculated the running time in this manner to adequately compare the times for the different EM sizes. Even though the impact of this dimension on individual running time is not significant, it may become so when sampling hundreds of thousands of worlds. For example, consider the difference between running time per sample for 1 billion worlds vs. 1 million worlds: $0.0289420 - 0.0289165 = 0.0000255$ s; for a sample size of 100,000 worlds, this difference amounts to 2.55 s. In the third column, we include an estimation of running times of the brute-force algorithm based on these values. Both running times are worst-case since optimization is possible (for instance, in our system we avoid recomputing warrant statuses of induced subprograms for which these values had been computed).

Figure 4b shows results concerning approximation quality; the metric was calculated with respect to the exact result for up to 20 EM variables (≈1M worlds). For the case of 30 EM variables (≈1B worlds), we approximated the metric using 250,000 worlds (which amounts to approximately 0.023% of the set of possible worlds), since the exact algorithm becomes intractable for instances of this size.

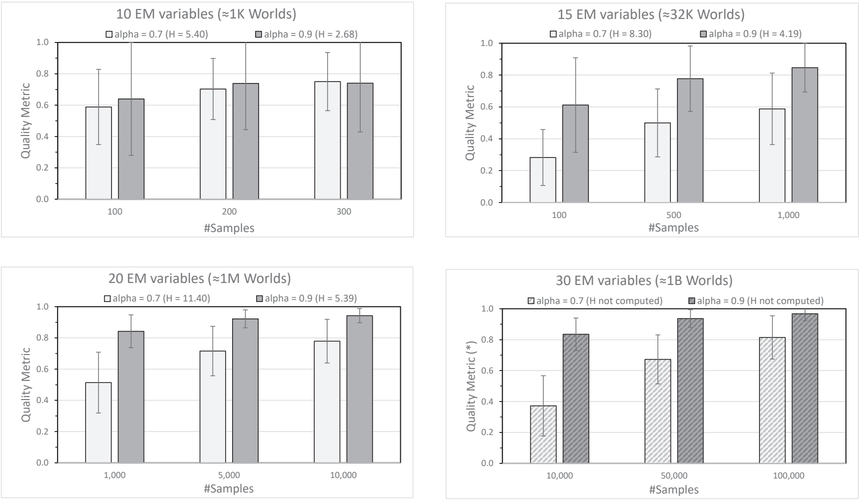
The following general observations arise from these results:

- First, sampling larger sets of worlds leads to higher quality approximations. Though this is expected, there are two interesting details:
    1. For the 20 EM variable case, the quality obtained by 5000 vs. 10,000 samples was not statistically significant (two-tailed two-sample unequal variance Student's t-tests yielded p-values greater than 0.08 for $\alpha = 0.7$ and greater than 0.16 for $\alpha = 0.9$), which means that only 5000 samples sufficed to obtain a good approximation.
    2. The proportion of repeated samples (i.e., wasted effort) was quite high for both entropy levels; for $\alpha = 0.7$ (higher entropy) on average 52% of samples were repeated, while for $\alpha = 0.9$ (lower entropy), an average of 87% were not unique. For the 20 EM variable case, the quality levels were achieved with only 2293 and 469 unique samples, respectively. Larger sample sizes also lead to lower variation in quality (shorter error bars).
- Next, entropy noticeably impacted solution quality (except for 10 EM variables, the smallest setting). Since our approximation algorithm samples worlds directly from the BN's distribution, it is natural to observe better effectiveness with lower (less spread out) entropy distributions. A smaller number of worlds represents a larger portion of the probability mass.
- Lastly, even for higher values of entropy, we observed adequate quality levels for modest numbers of samples compared to the size of the full sample space.

These results shed light on the applicability of P-DAQAP on real-world problems such as the CTA use case, given that relatively low numbers of effective (i.e., nonrepeated) samples yield good approximations of the exact values.

#EM Variables	Run. Time/Sample (seconds)	Est. Brute Force Run. Time (hours)
10 (1K worlds)	0.0286015	0.008
15 (32K worlds)	0.0288155	0.262
20 (1M worlds)	0.0289165	8.422
30 (1B worlds)	0.0289420	8632 (≈360 days)

(**a**)



(**b**)

Figure 4. (**a**) Average running times per world sampled ($n = 100$ runs). For each case, we estimate the running time (in hours) required to run the exact (brute force) algorithm. (**b**) Average solution quality varying #EM variables (log of #worlds), #samples, and the parameter that controls the entropy (H) of the probability distribution. For 30 EM variables (1B worlds, bottom right), quality is *approximated* on the basis of a sample of 250,000 worlds. Error bars correspond to standard deviation ($n > 50$ for the top charts, $n > 15$ for the bottom charts).

5.3. Results in the Context of Practical Applications

We now analyze the results we obtained in these experiments in the context of the MITRE ATT and CK data that we focused on for our use case in Section 3. For the purposes of this brief analysis, let us consider the Enterprise segment of the dataset, which contains 191 techniques and 385 subtechniques, and this translates into a large number of constants that would certainly lead to an intractable probabilistic model if tackled directly. Fortunately, there is a well-understood independence relation among such techniques, and they can, thus, be effectively pruned depending on the tactics to which they are associated. For instance, the *Privilege Escalation* tactic (TA0004) that we refer to in the use case has 13 associated techniques, while the rest of the techniques in the dataset associated at most 30 (with the exception of *Defense Evasion* (TA0005) that has 42, though additional filtering according to the specific operating system in question allows to bring this number down significantly). Our preliminary results therefore show that having the capacity to scale to 30 EM variables is within the realm of this kind of application, though further efforts are required to effectively arrive at submodels derived from the general one that can be used to solve specific query answering tasks. In this same vein, there are multiple research and development efforts to manipulate, adapt, and export data and knowledge from the ATT and CK dataset [32–35].

6. Conclusions and Future Work

We presented an extension of the DAQAP platform to incorporate probabilistic knowledge bases, giving rise to the P-DAQAP system, which is, to the best of our knowledge, the first system of its kind for probabilistic defeasible reasoning. After discussing the details of its design and describing applications to cybersecurity, we performed an empirical evaluation whose goal it was to explore the effectiveness and efficiency of world sampling-based approximate query answering. Our study showed that the entropy associated with the probability distribution over worlds has a large impact on expected solution quality, but even a modest number of samples suffices to reach good-quality approximations. Compared to classical (nonprobabilistic) approaches, the results of our experiments show that P-DAQAP allows for representing, effectively and efficiently reasoning with different types of uncertainty, modeling complex domains in more detail, and providing more informed answers that can be accompanied by explanations. In critical environments, having outputs of this kind increases credibility and trust in the system by its users.

Future work involves carrying out a broader evaluation investigating other sampling methods, avoiding repeated samples, and testing other probabilistic models. One of the goals of this research line is to develop a method to guide knowledge engineering efforts on the basis of domain features, requirements in terms of expressive power, approximation quality, and query response time.

Author Contributions: Conceptualization, M.A.L., A.J.G., P.S. and G.I.S.; methodology, M.A.L., P.S. and G.I.S.; validation, M.A.L., P.S. and G.I.S.; formal analysis, G.I.S.; investigation, M.A.L. and G.I.S.; writing—original draft, M.A.L.; writing—review and editing, M.A.L., A.J.G., P.S. and G.I.S.; project administration, G.I.S.; supervision and funding acquisition, A.J.G. and G.I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universidad Nacional del Sur (UNS) grant numbers PGI 24/ZN34 and PGI 24/N046, Universidad Nacional de Entre Ríos grant number PDTS-UNER 7066, and Agencia Nacional de Promoción Científica y Tecnológica, Argentina grant number grants PICT-2018-0475 (PRH-2014-0007). P.S. is supported by internal funding from the ASU Fulton Schools of Engineering.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AM	Analytical Model
CAPEC	Common Attack Pattern Enumeration and Classification
CPE	Common Platform Enumeration
CTA	Cyberthreat Analysis
CVE	Common Vulnerabilities and Exposures
CWE	Common Weakness Enumeration
DeLP	Defeasible Logic Programming
DeLP3E	Defeasible Logic Programming with Presumptions and Probabilistic Environments
EM	Environmental Model
KB	Knowledge Base
P-DAQAP	Probabilistic Defeasible Argumentation Query Answering Platform
NVD	National Vulnerability Database
XAI	Explainable Artificial Intelligence

References

1. Mumford, E. The story of socio-technical design: Reflections on its successes, failures and potential. *Inf. Syst. J.* **2006**, *16*, 317–342. [CrossRef]
2. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
3. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
4. Gunning, D. Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency (DARPA). 2017. Available online: https://nsarchive.gwu.edu/sites/default/files/documents/5794867/National-Security-Archive-David-Gunning-DARPA.pdf (accessed on 21 August 2022).
5. Viganò, L.; Magazzeni, D. Explainable security. In Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Genoa, Italy, 7–11 September 2020; pp. 293–300.
6. Castelvecchi, D. Can we open the black box of AI? *Nat. News* **2016**, *538*, 20. [CrossRef]
7. Mahdavifar, S.; Ghorbani, A.A. DeNNeS: Deep embedded neural network expert system for detecting cyber attacks. *Neural Comput. Appl.* **2020**, *32*, 14753–14780. [CrossRef]
8. Kuppa, A.; Le-Khac, N.A. Black Box Attacks on Explainable Artificial Intelligence (XAI) methods in Cyber Security. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
9. Szczepański, M.; Choraś, M.; Pawlicki, M.; Kozik, R. Achieving explainability of intrusion detection system by hybrid oracle-explainer approach. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
10. Malatji, M.; Sune, V.S.; Marnewick, A. Socio-technical systems cybersecurity framework. *Inf. Comput. Secur.* **2019**, *27*, 233–272. [CrossRef]
11. Alsmadi, I. *The NICE Cyber Security Framework: Cyber Security Management*; Springer Nature: Cham, Switzerland, 2020.
12. Leiva, M.A.; Simari, G.I.; Simari, G.R.; Shakarian, P. Cyber threat analysis with structured probabilistic argumentation. In Proceedings of the AI^3. CEUR-WS, Rende, Italy, 19–22 November 2019; Volume 2528, pp. 50–64.
13. Shakarian, P.; Simari, G.I.; Moores, G.; Parsons, S.; Falappa, M.A. An Argumentation-based Framework to Address the Attribution Problem in Cyber-Warfare. In Proceedings of the CyberSecurity, ASE, Stanford, CA, USA, 27–31 May 2014.
14. Kuppa, A.; Le-Khac, N.A. Adversarial xai methods in cybersecurity. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 4924–4938. [CrossRef]
15. Liu, H.; Zhong, C.; Alnusair, A.; Islam, S.R. FAIXID: A framework for enhancing ai explainability of intrusion detection results using data cleaning techniques. *J. Netw. Syst. Manag.* **2021**, *29*, 1–30. [CrossRef]
16. Srivastava, G.; Jhaveri, R.H.; Bhattacharya, S.; Pandya, S.; Rajeswari; Maddikunta, P.K.R.; Yenduri, G.; Hall, J.G.; Alazab, M.; Gadekallu, T.R. XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions. *arXiv* **2022**, arXiv:2206.03585.
17. Hariharan, S.; Velicheti, A.; Anagha, A.; Thomas, C.; Balakrishnan, N. Explainable Artificial Intelligence in Cybersecurity: A Brief Review. In Proceedings of the 2021 4th International Conference on Security and Privacy (ISEA-ISAP), Dhanbad, India, 27–30 October 2021; pp. 1–12.
18. Shakarian, P.; Simari, G.I.; Moores, G.; Paulo, D.; Parsons, S.; Falappa, M.A.; Aleali, A. Belief revision in structured probabilistic argumentation. *AMAI* **2016**, *78*, 259–301. [CrossRef]
19. Leiva, M.A.; Simari, G.I.; Gottifredi, S.; García, A.J.; Simari, G.R. DAQAP: Defeasible Argumentation Query Answering Platform. In Proceedings of the FQAS 2019, Amantea, Italy, 2–5 July 2019; pp. 126–138.
20. Simari, G.R.; Loui, R.P. A mathematical treatment of defeasible reasoning and its implementation. *Artif. Intell.* **1992**, *53*, 125–157. [CrossRef]
21. Toni, F. A tutorial on assumption-based argumentation. *Argum. Comput.* **2014**, *5*, 89–117. [CrossRef]
22. Modgil, S.; Prakken, H. The ASPIC+ framework for structured argumentation: A tutorial. *Argum. Comput.* **2014**, *5*, 31–62. [CrossRef]
23. García, A.J.; Simari, G.R. Defeasible logic programming: DeLP-servers, contextual queries, and explanations for answers. *Argum. Comput.* **2014**, *5*, 63–88. [CrossRef]
24. Besnard, P.; Garcia, A.; Hunter, A.; Modgil, S.; Prakken, H.; Simari, G.; Toni, F. Introduction to structured argumentation. *Argum. Comput.* **2014**, *5*, 1–4. [CrossRef]
25. Martinez, M.V.; García, A.J.; Simari, G.R. On the Use of Presumptions in Structured Defeasible Reasoning. In *COMMA*; Verheij, B., Szeider, S., Woltran, S., Eds.; IOS Press: Amsterdam, The Netherlands, 2012; Volume 245, pp. 185–196.
26. Suciu, D.; Olteanu, D.; Ré, C.; Koch, C. Probabilistic databases. *Synth. Lect. Data Manag.* **2011**, *3*, 1–180.
27. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Morgan Kaufmann: San Francisco, CA, USA, 1988.
28. Paredes, J.; Teze, J.C.; Simari, G.I.; Martinez, M.V. On the Importance of Domain-specific Explanations in AI-based Cybersecurity Systems (Technical Report). *arXiv* **2021**, arXiv:2108.02006.
29. Buron Brarda, M.E.; Tamargo, L.H.; García, A.J. Using Argumentation to Obtain and Explain Results in a Decision Support System. *IEEE Intell. Syst.* **2021**, *36*, 36–42. [CrossRef]

30. Grover, S.; Pulice, C.; Simari, G.I.; Subrahmanian, V.S. BEEF: Balanced English Explanations of Forecasts. *IEEE Trans. Comput. Soc. Syst.* **2019**, *6*, 350–364. [CrossRef]
31. Alfano, G.; Greco, S.; Parisi, F.; Simari, G.I.; Simari, G.R. Incremental computation for structured argumentation over dynamic DeLP knowledge bases. *Artif. Intell.* **2021**, *300*, 103553. [CrossRef]
32. Al-Shaer, R.; Spring, J.M.; Christou, E. Learning the Associations of MITRE ATT & CK Adversarial Techniques. In Proceedings of the 2020 IEEE Conference on Communications and Network Security (CNS), Avignon, France, 29 June–1 July 2020; pp. 1–9. [CrossRef]
33. Kuppa, A.; Aouad, L.; Le-Khac, N.A. Linking CVE's to MITRE ATT&CK Techniques. In Proceedings of the 16th International Conference on Availability, Reliability and Security, Vienna, Austria, 17–20 August 2021; pp. 1–12.
34. Hong, S.; Kim, K.; Kim, T. The Design and Implementation of Simulated Threat Generator based on MITRE ATT&CK for Cyber Warfare Training. *J. Korea Inst. Mil. Sci. Technol.* **2019**, *22*, 797–805.
35. Choi, S.; Yun, J.H.; Min, B.G. Probabilistic attack sequence generation and execution based on mitre att&ck for ics datasets. In Proceedings of the Cyber Security Experimentation and Test Workshop, Virtual, CA, USA, 9 August 2021; pp. 41–48.

big data and cognitive computing

Article

On the Way to Automatic Exploitation of Vulnerabilities and Validation of Systems Security through Security Chaos Engineering

Sara Palacios Chavarro [1], Pantaleone Nespoli [2], Daniel Díaz-López [1,3,*] and Yury Niño Roa [4]

[1] School of Engineering, Science and Technology, Universidad del Rosario, Bogotá 111321, D.C., Colombia
[2] Department of Information and Communications Engineering, University of Murcia, 30100 Murcia, Spain
[3] Tandon School of Engineering, New York University, Brooklyn, NY 11201, USA
[4] Cloud Infrastructure Engineering, Google, Bogotá 111321, D.C., Colombia
* Correspondence: danielo.diaz@urosario.edu.co or daniel.diaz@nyu.edu; Tel.: +57-2970200

Abstract: Software is behind the technological solutions that deliver many services to our society, which means that software security should not be considered a desirable feature anymore but more of a necessity. Protection of software is an endless labor that includes the improvement of security controls but also the understanding of the sources that induce incidents, which in many cases are due to bad implementation or assumptions of controls. As traditional methods may not be efficient in detecting those security assumptions, novel alternatives must be attempted. In this sense, Security Chaos Engineering (SCE) becomes an innovative methodology based on the definition of a steady state, a hypothesis, experiments, and metrics, which allow to identify failing components and ultimately protect assets under cyber risk scenarios. As an extension of a previous work , this paper presents ChaosXploit, an SCE-powered framework that employs a knowledge database, composed of attack trees, to expose vulnerabilities that exist in a software solution that has been previously defined as a target. The use of ChaosXploit may be part of a defensive security strategy to detect and correct software misconfigurations at an early stage. Finally, different experiments are described and executed to validate the feasibility of ChaosXploit in terms of auditing the security of cloud-managed services, i.e., Amazon buckets, which may be prone to misconfigurations and, consequently, targeted by potential cyberattacks.

Keywords: security chaos engineering; attack trees; cloud managed services; vulnerabilities

Citation: Palacios Chavarro, S.; Nespoli, P.; Díaz-López, D.; Niño Roa, Y. On the Way to Automatic Exploitation of Vulnerabilities and Validation of Systems Security through Security Chaos Engineering. *Big Data Cogn. Comput.* **2023**, *7*, 1. https://doi.org/10.3390/bdcc7010001

Academic Editors: Peter R.J. Trim and Yang-Im Lee

Received: 17 October 2022
Revised: 16 November 2022
Accepted: 22 November 2022
Published: 20 December 2022

1. Introduction

Protecting Information and Communication Technology (ICT) assets against potential threats is nowadays essential, especially with the advent of industry 4.0 and the consequent revolution. To this extent, cybersecurity aims to protect data and technological infrastructure in different spheres, e.g., personal, familiar, business, and social. In fact, different efforts have been made to contribute in such ways, for example, to protect persons against online sex offenders [1], to defend IoT devices from attacks against data or services [2], to make smart cities' infrastructure more resilient [3], to implement cybersecurity in distributed organizations [4], and to support LEA's (Law Enforcement Agencies) in the detection of malware [5] or in the prevention of cybercrimes [6]. Additionally, cybersecurity has also been considered a field of knowledge that goes beyond the validation of identity, protection of access, and monitorization of actions. Indeed, it has become a field that focuses its efforts on the consistency and resilience of systems.

Besides, Site Reliability Engineering (SRE) is a set of practices that aims to improve a system's design parameters and the conditions where it operates to supply the system with essential attributes such as scalability, reliability, and efficiency. The SRE concept originated at Google around 2003 and was rapidly adopted by other companies with strict software requirements regarding scalability and reliability [7]. SRE may be seen as one way to materialize a DevOps strategy as it offers a set of principles around automatization, quantification of business-required reliability, reduction of availability risks, and observability. SRE may be implemented through the definition of reliability goals such as SLI (Service Level Indicator) or SLO (Service Level Objective), the development of a capacity plan, and the definition and execution of a change management process, among others [8].

A relatively new approach in the scope of SRE used to test the resiliency of distributed systems has recently emerged, known as Chaos Engineering (CE). CE is used to validate a system's strengths and vulnerabilities when exposed to uncontrolled conditions. By leveraging the CE methodology, different tests may be designed and applied with the aim of validating in a measurable way the changes that the steady state of a system may experiment [9]. Furthermore, another CE principle refers to the importance of including real word events (hardware or software failures) in the experiments, especially events that have the potential to generate a high impact or may occur with some frequency. CE also remarks on the importance of automating experiments as much as possible as it allows for a better analysis of the outcomes. Lastly, CE prioritizes the execution of tests in production to guarantee authenticity in the experiments and to consider real traffic patterns, although the impact of such experiments should be carefully estimated and contained.

Following the CE methodology, a "chaotic" experiment must be designed over a controlled environment, which allows the observation of the variables that define the steady state of the target system. Additionally, such a CE experiment must be ruled by a scientific method that allows the definition and validation of a set of hypotheses [10]. Lately, CE experiments have gained importance as a way to implement SRE as it allows testing the resiliency of a system against chaotic events so that the system's weaknesses can be identified and corrected in advance. Nonetheless, the resiliency of a system should be validated not only from a perspective of availability. In fact, it should include other aspects related to the secure and correct operation of the system. Thus, the necessity of evaluating the system's resiliency in a holistic way emerges and is consistently most demanded when we evaluate distributed systems that manage sensitive information, such as secure IoT services [11] or personal data management solutions [12].

Intending to execute a security-based evaluation of a system, a fresh concept emerged in 2017 to apply CE principles to experiments that, together with the availability, evaluate the confidentiality and integrity of a system under chaotic events. That is, SCE (Security Chaos Engineering) joins the cybersecurity ecosystem, trying to defend the systems against such events. In a cybersecurity context, chaotic events may be generated by a threat agent that tries to: (i) make a system unavailable, e.g., through a Distributed Denial of Service (DDoS) attack, (ii) read sensible data hosted by a system, e.g., through an elevation of privileges that facilitate the access to restricted information, or (iii) modify users or system files that alter the operation of the system, e.g., through the remote execution of malicious code [13].

Noting that the CE methodology can significantly impact new developments by reducing vulnerabilities through the scientific method and experimentation, this paper addresses the following research question: how can SCE be used to detect application vulnerabilities automatically, not limited to a specific context and by taking into account the actions that are preferred by an attacker based on the effort expended in the exploitation?

Thus, the current paper proposes an SCE framework based on attack trees named ChaosXploit. ChaosXploit is expected to support the operations of security teams in charge of detecting and correcting in an anticipated way the vulnerabilities that an under-analysis system may contain. The defensive labor of those teams implies understanding the attack goal that an attacker may pursue as well as the offensive techniques that he/she may use to achieve such an attack goal.

Thus, the main contributions of this paper are summarized as follows:

1. The proposal of ChaosXploit, an SCE framework that leverages attack trees to address the execution of attacks. Particularly, the ChaosXploilt architecture contains three main components: an observer, an experiment runner, and a knowledge database;
2. The design of an attack tree that pursues a specific attack goal, i.e., the extraction or modification of AWS S3 buckets information that enriches the knowledge database of ChaosXploit;
3. The execution of a set of experiments that validates the feasibility of ChaosXploit to execute an attack tree over a specific target, i.e., AWS S3 bucket, exposing multiple misconfigurations.

ChaosXploit was first presented in Ref. [14], and the present paper is an extended version of such work to include the following improvements and new content:

1. An extension of Section 2 (State of the art) with a detailed analysis of the related works, including a comparative table with six identified key features;
2. The addition of Section 3 (Background), which includes a set of essential concepts that introduces the reader to SCE;
3. An extension of Section 5 (Experiments) resulting in the implementation and execution of the second branch of the proposed attack tree, which aims to extract or modify information from the AWS S3 buckets;
4. The addition of Section 6 (Discussion), which includes an analysis of the current and future adoption of SCE in the industry.

The remainder of this paper is organized as follows: Section 2 gathers the major works contributing to SCE, analyzing their strengths and weaknesses. Then, Section 3 explains the fundamentals and concepts regarding CE and SCE. In Section 4, ChaosXploit, our proposed framework to execute SCE experiments, is described. In Section 5, diverse experiments to test ChaosXploit are designed and performed. Section 6 presents an engaging discussion on the adoption of SCE in enterprises. Finally, Section 7 concludes the work, adding future work to possibly improve ChaosXploit.

2. State of the Art

Throughout the literature, CE appears as a relatively hot research topic. That is, its robust capabilities have been described in different research items while being applied in several contexts. Nonetheless, such an application and a proper definition must be clarified since they have been ambiguous so far.

Starting with Netflix's release of *Chaos Monkey* in 2011 [15], the CE paradigm has been mainly used to test the resilience and robustness of virtualized appliances, demonstrating the potentialities of the chaotic method in such scenarios.

To this extent, the work in Ref. [16] described *Pystol*, a fault injection framework to argue on the resiliency of hybrid-cloud systems in adverse events. Specifically, *Pystol* is presented as a Software Product Line (SPL) that can be mounted on top of cloud infrastructures, being able to exploit CE's capacities. The proposal is then developed in a production environment and deployed using standard Kubernetes objects (together with the corresponding APIs) and Amazon Web Services (AWS) to execute the entire cluster with three use cases. It is worth mentioning that *Pystol* has been made available as an open-source code for further community development.

Additionally, Simonsson et al. [17] proposed *ChaosOrca*, another open-source fault injection platform for system calls in containerized applications based on CE principles. In this sense, *ChaosOrca* can calculate the self-protection ability of Docker-based microservices with regard to system call errors. In particular, the system determines the steady state of the Docker container by systematically registering diverse system metrics (CPU and RAM consumption, network I/O, among others). Later, some perturbations are injected into the system calls executed by the isolated dockerized app, avoiding the possible impact on the ordinary operations of other containers. The proposal is tested in three Docker microservices scenarios, namely Torrent, Bookinfo, and Nginx, demonstrating encouraging results in noticing resilience flaws.

An interesting case study on applying the CE methodology to a real use-case scenario has been conducted in Ref. [18]. The main idea of the authors is to introduce the CE paradigm at ICE Gruppen AB, a group of companies working in the grocery market. Mainly, they started with a literature review, studying the state-of-the-art works on CE and performing explanatory interviews in the company. The resulting framework, based on a total of 27 open source CE tools, is then applied to the IT system of the company, including its e-commerce. Interestingly, among the CE categories identified during the process, the authors also indicate "network attacks" and "security attacks".

Furthermore, *ChaosMachine* is described by Zhang et al. [19]. Particularly, it can be defined as an open-source and extensible CE framework written in Java to analyze the capacities of handling exceptions in production environments. In this sense, *ChaosMachine* is able to disclose possible resilience issues of try-catch blocks, proposing an architecture composed of three parts: (i) a monitoring sidecar, (ii) a perturbation injector, and (iii) the chaos component. Then, *ChaosMachine* is tested with three voluminous open-source Java apps, totaling 630k code lines, exhibiting its capacities in production environments with realistic workloads.

Lately, the principal objective of the chaotic methodology has changed, shifting from resilience surrounding a system to enclosing security issues. Starting from the assumption that security failures will happen doubtless, SCE's primary goal is to test the system's security controls using proactive experiments and, therefore, building confidence in its capabilities to protect against potential threats.

Lamentably, since this paradigm shift has recently happened, the quantity of academic items and tools is still insufficient. In this sense, *ChaoSlingr* can be depicted as the first open-source software contribution to exhibit the potential application of the chaotic principles to information security [20]. The tool was developed to function on AWS by a team at the UnitedHealth Group, led by Aaron Rinehart, to demonstrate a simplified mode for designing security chaos experiments [21]. From the main project, several companies have started to leverage *ChaoSlingr* to execute chaotic experiments within their systems.

Moreover, Torkura et al. [13] proposed *CloudStrike*, a software architecture that measures the security of cloud environments by applying Risk-Driven Fault Injection (RDFI). For the reader's sake, the tool was first proposed in a previous article [22]. Concretely, RDFI expands the CE paradigm to contemplate cloud security without losing the resilience perspective by injecting security faults, leveraging the attack graphs representation. Such SCE tool is tested on various cloud services of principal platforms, namely, AWS, and Google Cloud Platform. Notably, the authors claim that they can calculate the risk value to which the system's assets are being exposed to by using the Common Vulnerability Scoring System (CVSS). Then, the authors used the SCE methodology to test another tool, *CSBAuditor*, a cloud security framework that can continuously monitor cloud infrastructures to identify possible ill-motivated activities [23].

Additionally, the application of SCE to enhance API security is defined in Ref. [24]. Due to the popularity of RESTful APIs in distributed applications, the authors propose utilizing this methodology to test the configuration of the API's security controls, exposing early vulnerabilities. After focusing on the OWASP (Open Web Application Security Project) list of the top 10 critical web application security risks and automated attack detection, the authors suggest the application of SCE experiments to address the abovementioned challenges. Indeed, the work is still in an early phase, but the capabilities of SCE are recognized as being valuable.

Besides, SCE experiments have been used to test System of Systems (SoS) robustness against potential attackers in [25]. Concretely, the authors used Chaos Toolkit to conduct several CE and SCE experiments on a Virtual Unmanned Aerial Vehicle (VUAV). The Attack Trees methodology is employed to better model possible attacker moves, assuming the level of access he/she would possess with a previous threat modeling phase. Precisely, two Attack Trees are developed, namely, injecting corrupted navigation service and killing ActiveMQ/WorldWind (i.e., the software tools used for communication purposes). Then, five separate experiments are executed, evaluating the performance by measuring the CPU and RAM usage. Results showed a slight increase in CPU load, while RAM was not a significant metric during tests.

Table 1 summarizes the findings of the state-of-the-art investigation. It has to be remarked that the works [16–19] refer to CE applications while [13,24,25] propose SCE employment. Consequently, one could argue that it is obvious that the attributes' value of the CE works tend to be "Resiliency", while "Security" is predominant for the SCE proposals. Nevertheless, the proposed framework in Ref. [18] adds security features to the CE requirements. Such confusion is directly derived from the ambiguous definition of CE, as previously stated.

Table 1. Comparison of the related works highlighting the main features.

Related Work	Attributes	Application Context	CE Tool	Threat Model	Experimental Data	Automation Level
Camacho et al. [16]	Resiliency	Cloud	Ad-hoc	✗	✗	✗
Simonsson et al. [17]	Resiliency	Containers (Docker)	Ad-hoc	✗	crafted	✓
Jernberg et al. [18]	Resiliency Security	Web	27 CE tools survey	✗	crafted	✗
Zhang et al. [19]	Resiliency	Java applications	Ad-hoc	✗	Public Java code	✓
Torkura et al. [13]	Security	Cloud	Ad-hoc	Could Attack Graphs	crafted	≈
Sharieh, Ferwron [24]	Security	API	✗	✗	✗	✗
Bailey et al. [25]	Security	SoS	ChaosToolkit	Attack Trees	crafted	✗
Our proposal ChaosXploit	Security	Any	ChaosToolkit based	Attack Trees	Public buckets	✓

Legend: ✓ Yes, ✗ No, ≈ Partially.

Another clear difference between CE and SCE works is that most SCE proposals leverage a threat model to map the attackers' moves within the protected system. In particular, Attack Graphs and Attack Trees seem to be a suitable choice to infer the goals of the attackers and, possibly, anticipate them.

Regarding the tool used to implement the proposals, many of them present ad hoc development of the CE/SCE framework. In this sense, one could say that, in specific situations, implementing from scratch can lead to better solutions. However, re-using already mature and tested tools should be the primary choice in order to fairly compare different proposals.

Additionally, two key aspects must be highlighted: (i) the importance of using publicly available data to perform experiments and (ii) the significance of a high automation level for CE/SCE frameworks. That is, most of the analyzed papers present crafted experiments to demonstrate their features, making the comparison challenging to execute. Then, one of the crucial characteristics of any chaotic tool is the automation level of the experiments. Since modern systems feature high complexity and distribution, automating those experiments is highly desirable.

Last but not least, the surveyed works suggest the chaos tests application only in a particular context (e.g., Cloud, containers, etc.). It is effortless to claim that the design of a full-fledged CE/SCE tool would broaden its application scope, leading to more experimentation and, perhaps, better results.

The research presented in the paper at hand uses as reference the characteristics of all these tools presented in related works and proposes an SCE-powered framework based on attack trees to detect and exploit vulnerabilities in different targets as part of an offensive security exercise. This framework, unlike those previously mentioned, can be used in any application context whether in different clouds (AWS, GCP, Azure), containers (Docker, Kubernetes), or web applications. Additionally, compared to current CE tools, our proposal develops a threat model based on attack trees since these enable modeling organized actions for more than one SCE experiment, allowing a better traceability and following the same attack goal. Another differentiating component that stands out in our proposal vs. other SCE tools is the high level of automation, since we can make a list of actions to be performed and, when launching the experiment, these will be executed in a row. Finally, we are aware that we are not reinventing the wheel, as our proposal is built on the ChaosToolKit, one of the most mature tools in CE. Lastly, our proposal is tested with common cloud services, meaning that the experiments can be easily replicated.

3. Background

For the sake of the reader, some important concepts are introduced to allow a better understanding of the context surrounding CE and SCE.

3.1. Chaos Engineering (CE)

As previously mentioned, the concept of CE emerged in 2011 when Netflix moved its services to the AWS cloud. Netflix's engineers feared that an internal instance could fail during the move, severely impacting the overall operation. For this reason, *ChaosMonkey* was created to test the stability of Netflix by injecting faults that randomly terminate internal instances [26]. A year after launching *ChaosMonkey*, Netflix added new modes that report different types of faults or detect abnormal conditions. Each of those modes were considered to be a new simian, and together they formed what is known as the *SimianArmy* [27].

In 2016, Kolton Andrus and Matthew Fornaciari founded Gremlin [28], which is recognized as a leading CE solution. Along with the creation of Gremlin, the formal definition of CE was also born as "the discipline of experimenting on a system in order to build confidence in the system's capability to withstand turbulent conditions in production" [9].

A few CE frameworks may be found in the wild. One of the most notable frameworks is the above-mentioned Gremlin, which allows one to experiment with more than 10 different attack strategies on different infrastructures. Nevertheless, not all of those strategies are free to use, and it does not have reporting capabilities. Another well-known CE framework is ChaosMesh [29], an open-source cloud-native tool built on Kubernetes Custom Resource Definition (CRD). Specifically, it allows testing several scenarios checking for network latency, system time manipulation, and resource utilization, among others. Nonetheless, this tool does not have the advantage of scheduling attacks.

Another open-source CE framework is Litmus [30] which allows developers to use a set of tools to create, facilitate, and analyze chaos in Kubernetes with automatic error detection and resilience scoring. Last but not least, it is important to mention ChaosToolkit (CTK) [31], an open-source tool that permits the automation and customization of CE experiments by defining a set of probes and actions that may be pointed to different types of targets.

It is worth remarking that the CE experiments are not chaotic at all. In fact, they are based on the scientific method and should follow the CE principles [9] that define the subsequent steps to guarantee that the experiments are correctly executed.

1. Define the behavior of the system (**observability**), which is key to measure with the purpose of approval or disapproval of an hypothesis that may be defined later;
2. Identify the **steady state** to mark out what should be considered as a normal behavior of the system;
3. Define a **hypothesis** that will be proved or refuted at the end of the experiment;
4. **Execute** the experiment by introducing real-world events such as creating instances that expose malfunctions and interrupted network connections, among others.

The fact that CE experiments have a defined method corroborates that this discipline does not consist of "breaking things on purpose". On the contrary, CE experiments are generally done in a proper testing environment with similar conditions to the ones obtained in a real environment exposed to disruptive incidents. Thus, the application of CE allows testing attributes such as availability and reliability in a controlled environment. Generally, the results that arise from conducting a CE experiment can help anticipate incidents, improve the understanding of system failure modes and reduce maintenance costs [32].

Once the method and benefits of implementing CE have been discussed, defining and implementing an experiment can be effortless. For example, in Ref. [10], one of the experiments considered a recommendation system that, as part of its functionality, stores all the searches inserted by users in a cache so that such queries may be used to redefine the recommended product that is returned to the user. The experiment uses CE to check what would happen if the communication were to fail between (i) the process (Redis Client) requesting to store the queries and (ii) the cache (Redis server) that effectively stores them. In this case, the purpose of the CE experiment is to determine if the recommendation system may still work after injecting failures, so it is defined as follows:

- **Observability**: Navigate the application and view the recommended products;
- **Steady State**: The recommended products should be returned to the user;
- **Hypothesis**: A failure in the communication with the storage component (Redis server) causes a failure in the product returned to the user by the recommendation system, even in subsequent queries when the storage component is restored.

With the **execution** of this experiment it may be possible to conclude, for example, that the hypothesis is refuted since when injecting failures in Redis Server, the recommendation system handles the error and manages to recover automatically as soon as the access to the storage system is re-established. Thus, it proves that the recommendation system is resilient to failures in the storage system.

3.2. Security Chaos Engineering

By using CE, testing security in systems with the premise that "failure is the greatest teacher" is possible. This idea was first proposed by Aaron Rinehart [21], who pursued the application of CE in cybersecurity while working as Chief Security Architecture at the UnitedHealth Group [33]. As mentioned in the previous section, CE has traditionally focused on testing system availability, while recent research is striving to apply this discipline in the field of cybersecurity. Concretely, the main goal is to apply CE concepts by testing not only the availability but also other attributes such as integrity and confidentiality to boost the concept of *Security Chaos Engineering* (SCE). SCE has been defined as "the identification of security control failures through proactive experimentation to build confidence in the system's ability to defend against malicious conditions in production" [21].

In this context, ChaosSlingr can be recognized as the first open-source framework that demonstrated the value of applying SCE to cybersecurity [34]. This tool was created by Aaron Rinehart and proposed a simple experiment. It sought to misconfigure some ports on a system and observe the behavior. Although it was a good initiative, ChaoSlinger was no longer maintained and became part of a larger project known as Verica [35].

As mentioned, while CE aims to test the resilience of a system, SCE also provides measures and experiments to provide top-notch security to the systems. By leveraging the SCE methodology, it is possible not only to corroborate assumptions or discover vulnerabilities but also to infer possible mitigations [36]. That is, SCE falls into the cybersecurity ecosystem, as it allows checking that the security controls that validate the confidentiality, integrity, and availability of the system are reliable. This check is based on identifying security flaws caused by the human component, insecure design, and lack of resilience in the system under protection. In addition, SCE experiments can identify the exact points where security flaws exist and act on time.

The methodology applied by SCE is similar to the one described for CE, as it incorporates the definition of steady state, observability, and hypothesis. However, it pursues a different objective as it aims to validate the security of a system, for example, by discovering vulnerabilities, misconfigurations, logic flaws, and insecure design, among others. In addition, if experiments are executed frequently, SCE may help in the reduction of security incidents and remediation costs, as it allows developers to: (i) understand their system, (ii) define a response plan, (iii) identify system modules failing, and (iv) note that some components were omitted during development. In addition, SCE minimizes impacts on users through experimentation, which in turn improves the ability of developers to track and measure security.

One helpful experiment to explain the SCE methodology is associated with understanding the behavior of a firewall when some associated ports are misconfigured. This was one of the experiments that were executed with *ChaoSlingr*, a framework created by a team at UnitedHealthGroup, explained in detail in Chapter 7: the journey to SCE of [21]. A brief overview of the experiment is presented below:

- **Observability**: Detection of security configuration changes that have occurred in a device;
- **Steady State**: The firewall is able to detect all changes over the ports;
- **Hypothesis**: A misconfigured port should be detected and blocked by the firewall, and such an event should be appropriately logged.

From the **execution** of this experiment, it could be possible to prove that half of the time, the hypothesis is fulfilled, and the other half of the time, the firewall does not detect and block it. In addition, a cloud configuration tool could be able to detect the failure, but this is not being logged, so it is not possible to identify that an incident has occurred. Thus, proper remediations should be undertaken to avoid the incorrect operation of the firewall.

3.3. Differences between SCE and Traditional Pentesting

At this point, one could legitimately wonder about the difference between SCE and traditional penetration testing techniques and the added value of using SCE. In order to establish these differences, Table 2 illustrates some key aspects to be considered in this comparison, which are explained in the following paragraphs.

Table 2. Main differences between traditional pentesting and SCE.

	Traditional Pentesting	**SCE**
People implementing	Executed mainly by personnel external to the organization (external red team)	Executed mainly by organization's internal personnel (internal red or blue team)
Methodology behind	ISECOM, EC-Council, OWASP, others	Chaos Engineering principles
Security approach	Offensive	Defensive
Available tools	Bunch of offensive tools	Few SCE frameworks
Grade of automatization	Mainly manual procedures	Mainly automated procedures
Expected frequency	Depend on organization policies and risk appetite, generally every 3 or 6 months	High frequency for definition, can be applied for each incremental development
Phase of SDLC where applied	Generally in production phase	Along all the SDLC
Scope of tests	Generally unitary tests	Unitary and full system tests
Kind of vulnerabilities detected	Own-system errors, misconfigurations	Own-system errors, security assumptions about the systems

As indicated in Table 2 traditional pentesting allows attacking different targets by finding and exploiting vulnerabilities and misconfigurations. On the other hand, SCE allows us not only to test for system errors but also for security assumptions about the system, which includes component misconfiguration but also human errors, so we can affirm that SCE has a bigger scope in terms of vulnerabilities that can be detected.

In addition, the pentesting process may require a set of different activities, which can be automated in a defined way, e.g., fingerprinting, scanning, and brute forcing, but the exploitation phase will generally require highly manual activities through the construction of customized exploits and payloads. Secondly, SCE strives for a high automatization in the development of experiments, so they can be reproducible and repeatable.

Additionally, traditional pentesting is generally executed by an external red team, because generally, the aim is to emulate a double-blind scenario where an attacker does not know the internal details about the system that he is attacking, and the persons in charge of protecting the system do not know when the attack will be launched [37]. In this regard, SCE offers a different approach, as SCE experiments are intended to be executed by the persons who build (developers), maintain, and secure the system, who can be part or not part of a blue team or an internal red team in case the organization has one; all of this is part of a defensive strategy.

The frequency of pentesting exercises may depend on external regulatory or internal requirements and organization risk appetite, resulting in pentesting tests developed regularly, e.g., every 3 or 6 months for the case of organizations with an intermediate maturity security level, and mainly over systems that are in the production phase. In the case of SCE, the experiments have a high frequency by definition, as SCE experiments may be designed and performed along the software development life cycle. This means it is possible to incorporate it in the early stages of development and reduce the remediation costs.

It is important to note that currently there are many tools available that can be used in different phases of pentesting, but there are not many SCE-based tools, as indicated in Section 2, so the contribution of a framework in this regard improves the traditional pentesting process as it offers an alternative way of detecting vulnerabilities in the protected assets, providing a new tactic that enriches the existing tool-set of blue and red teams. Additionally, when considering complex or distributed systems, SCE experiments help to understand the system as a whole, going beyond unit tests over specific components which is common in pentesting exercises.

Finally, methodologies behind pentesting refer to quite popular publications from ISECOM (OSSTMM methodology), EC-Council (hacking phases), or OWASP (security testing guides), among others. However, none of them are based on a scientific method, which SCE does by following the CE principles.

4. ChaosXploit Architecture

This section describes ChaosXploit, a SCE-powered framework composed of different modules that support the application of CE methodology (described in Section 3.1) to test security in different kinds of information systems. The architecture of the proposal is depicted in Figure 1. It is worth noting that a label has been assigned to each module to represent the step in the EC methodology that is executed in that module. Additionally, each internal module is described in the following sections. In particular, the Knowledge Database is described in Section 4.1, the Observer is detailed in Section 4.2, and the SCE Experiments Runner is explained in Section 4.3.

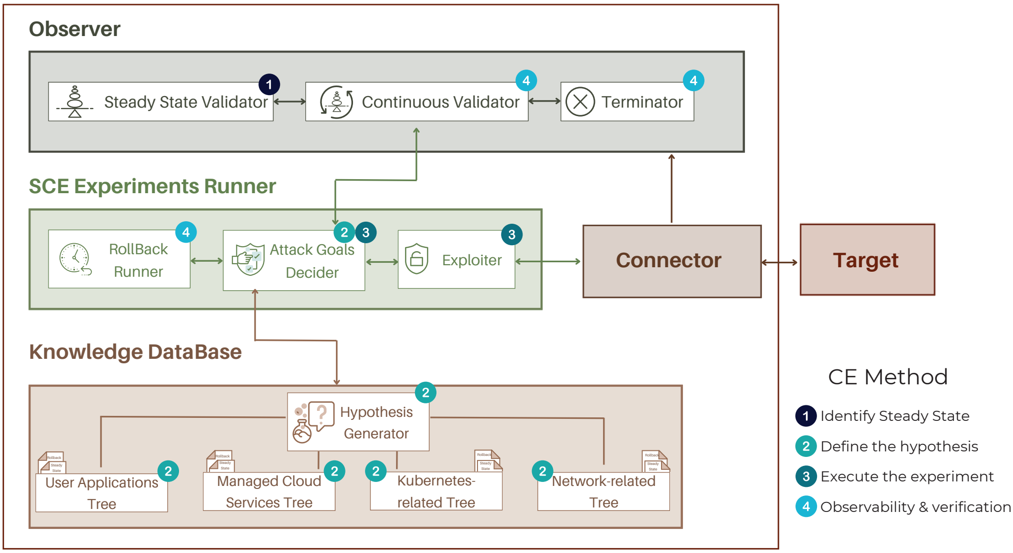


Figure 1. The proposed architecture of ChaosXploit and its relation to SCE methodology.

4.1. Knowledge Database

The knowledge database is responsible for providing the steps required to conduct an offensive SCE experiment executed by a team (blue team) interested in maturing a defensive strategy. Thus, this module is composed of a set of attack trees and a hypothesis generator, which will be some of those in charge of executing the second step of the CE methodology, i.e., defining the hypothesis for the experiment. The tasks assigned to these modules are detailed below.

4.1.1. Attack Trees

This module is in charge of delivering the intelligence for executing the SCE experiments. Such intelligence is represented by different attack trees, where each tree clusters different branches focused on achieving a specific attack goal, e.g., gaining access to data stored in a cloud storage solution. So, different attack goals may be pursued as attack trees are contained in the knowledge database. Each branch of an attack tree gathers different offensive actions that may be conducted to achieve the final attack goal, where an action may be a python script, an HTTP request, or some process to be run on the operating system. It is worth mentioning that attack trees for different types of targets may be defined, such as trees for user applications, managed cloud services, Kubernetes, and network devices, among others.

4.1.2. Hypothesis Generator

The intelligence contained in the attack trees needs to be converted to a hypothesis so that it can be consumed by the other modules of ChaosXploit. So, the Hypothesis Generator is responsible for translating the branch actions contained in an attack tree into a form readable for the module that executes the SCE experiments, i.e., the exploiter. Each hypothesis generated by this module is a statement about the system being tested that must be refuted or confirmed by the SCE experiments, e.g., an organization will not expose private data when the recognition tool Foca [38] is pointed out to the main domain.

4.2. Observer

The observer groups all the activities related to the observation of both the target and the SCE experiment. This module is important because it allows controlling the specific conditions of the target before, along, and after the execution of the SCE experiments.Therefore, this module will address, in its different components, the first step of the CE methodology: identification of the steady state and the fourth step: observability and verification. This module is composed of a steady-state validator, a continuous validator, and a terminator.

4.2.1. Steady State Validator

The steady-state validator is responsible for verifying the steady-state hypothesis on the target representing the steady-state conditions, which allows us to create a direct association with the first step of the CE methodology. These conditions will depend on the target of the attack and the hypothesis defined in the hypothesis generator. For example, a normal condition may be a well-formed response from a web server or an assumption about the system.

4.2.2. Continuous Validator

The continuous validator is activated once the experiment starts and is constantly checked until the end of the experiment. It allows for verifying specific signals detected from the target, which makes it possible to determine the results of an interaction between the exploiter and the target. These signals are especially important because they can indicate whether a current action included in a branch of an attack tree has succeeded, so the following action in the branch should be triggered, or they can indicate that the target is not vulnerable and the other actions in the branch should not be executed. This leads us to categorize it as one of the components that perform the last step of the CE methodology, as it allows us to observe and verify the behavior of the experiment.

4.2.3. Terminator

Each time the execution of an action is completed, the experiment status is updated and the terminator validation is performed. This module observes the failure states of the SCE experiment to define the actions to be taken accordingly, thus it is associated with the last step of the CE methodology. For example, if the target stops responding due to the execution of an SCE experiment, the experiment status is updated to failed and the terminator will be able to inform the Rollback Runner so that it can restore the target.

4.3. SCE Experiments Runner

The SCE Experiments Runner is in charge of the SCE experiment's execution over a target to validate or refute a hypothesis. This component is fundamental because it not only leads the interaction with the target but also centralizes the communication with the observer and knowledge database. Although it is an execution module, it also includes elements that contribute to the development of the other steps of the CE methodology. It consists of three main elements: attack goal decider, exploiter, and rollback runner.

4.3.1. Attack Goal Decider

The attack goal decider receives a defined goal attack as input to be tested over a target. Such an attack goal may be contributed by the user of ChaosXploit who is interested in probing if a particular system is susceptible to a specific attack. Then, the attack goal decider requests the knowledge database for the proper attack tree that matches such a defined goal. This request implies that the module is involved in the hypothesis generation process (step 2 of the CE methodology). In addition, when asking for the information from the knowledge database, it will receive the actions to be performed to execute the experiment, which allows it to be associated with the third step of the methodology as well.

4.3.2. Exploiter

The exploiter executes the SCE experiment over a target to validate or refute a hypothesis. This is directly associated with the third step of the methodology. With such purpose, the exploiter performs the offensive actions defined previously by the attack tree obtained from the knowledge database. Besides, it is also able to collect information about specific responses coming from the target to define the next step in an attack.

4.3.3. Rollback Runner

An experiment may contain a sequence of actions that reverse what was undone during the execution; this allows us to identify the points where failures were generated. Thus, the Rollback Runner is supported by the last phase of the methodology. The set of actions will be called by the Rollback Runner after the Continuous Validator finishes its execution regardless of whether an error occurred in the process or not.

4.4. Connector

The connector is responsible for searching for the most suitable extension to connect to the target on which the user wants to run the experiment. Once an extension has been defined, the connector establishes the link with the target and tests that the scenario is adequate to run the SCE experiment.

While ChaosXploit has a high level of automation, some previous activities are required before executing the experiments. First, the security team in charge of testing an under-analysis system must define the attack goal to be tested in the experiments and draw an hypothesis with its corresponding steady state. Then, an attack tree consistent with the previously defined attack goal is needed, which may come from an external cyberthreat intelligence provider (in cases where the under-analysis system is common and sufficiently known by the provider) or from the security team that builds it as a way to understand the possible steps an attacker could perform to achieve the attack goal. After the attack tree is defined, ChaosXploit will automatically perform all necessary actions, i.e., identify the vulnerability type, do the exploitation from the tree and measure steady-state, to conclude the SCE experiments. In case the results have not been satisfactorily completed, the type of vulnerability found will be indicated by ChaosXploit.

The interactions between the components of ChaosXploit are shown in Figure 2. First, the user of ChaosXploit requests the Attack Goal Decider for the execution of a SCE experiment, informing the attack goal to be considered and the target where the SCE experiment should be addressed. Then, the Attack Goal Decider retrieves from the knowledge database the steady-state of the experiment, the rollback procedure, and the most proper hypothesis (a branch in the attack tree) that matches the attack goal desired by the user. The Attack Goal Decider also requests to the Connector the preparation of the extension for the target informed by the user. When a connection to the target is established and a hypothesis is defined, the Attack Goal Decider then performs the following actions: (i) It establishes the steady state of the experiment in the Observer and tests it in an initial phase. Therefore, in this step, it is necessary to establish a new connection to validate its stability. In case this action fails, the state of the experiment is updated to failed and it is terminated; (ii) it starts the execution of the steps defined in the selected branch of the attack tree with the help of the Exploiter, and (iii) it keeps continuous communication with the Continuous Validator to monitor the execution of the exploitation in progress and in that way be aware if the attack goal is achieved. If the Continuous Validation fails, then the termination process is activated by the Terminator. The experiment ends with the execution of the Rollback Runner to restore everything.

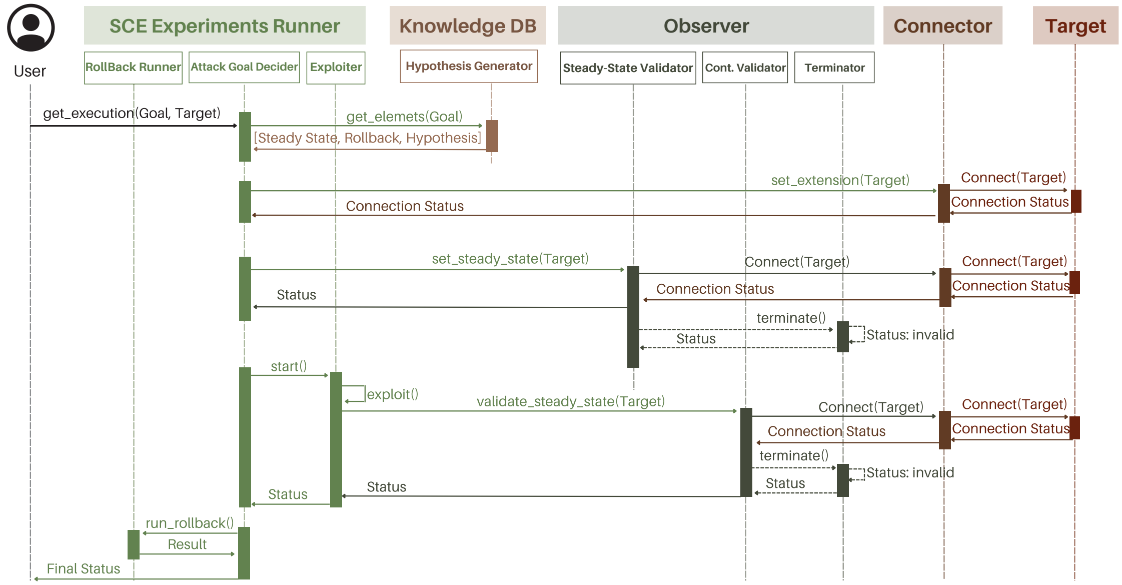


Figure 2. Flow diagram of the execution of a SCE experiment in ChaosXploit.

5. Experiments

Multiple experiments have been conducted using the ChaosXploit proposal mentioned in Section 4, which are also available in the public repository of the project [39]. Based on the fact that AWS S3 buckets and Elasticsearch databases account for nearly 45% of the cloud misconfigured and compromised technologies [40], the proposed session of ChaosXploit experiments focuses on evaluating the security of the AWS S3 service. It considers the possible configurations and whether they permit establishing a connection, whether they are public or private buckets, or whether they permit getting the configured Access Control Lists (ACLs) which allow managing the access to the buckets and their objects. These lists define which AWS accounts or groups have access and what kind of permissions they have.

This section of experiments comprises the following subsections: Settings, Section 5.1, in which the hardware and software requirements to develop the experiment, are specified. Definition of the knowledge database, Section 5.2, in which the attack tree is presented together with the specification of the branches chosen for the experiments. Sections 5.3 and 5.4 describe the implementation of the first and second branches of the attack tree. Each of them contains the definition of the steady-state and the hypothesis, as well as the input parameters and the monitored variables. Additionally, each of them includes a subsection for result analysis.

5.1. Settings

The following setup was used to execute the above-mentioned experimental session using ChaosXploit:

- Hardware: the experiments were executed on a Fedora OS with AMD Ryzen 5 3500U CPU, 8 GB RAM, and 512 GB SSD;
- Internal Components: Some of the components of ChaosXploit have been built over existing modules of ChaosToolkit, as it is an open-source framework that allows its extension and improvement to make it oriented to security purposes. ChaosToolkit was chosen since this tool simply allows automation of the experiments using *json* files. The connection to the different targets (buckets) was done using boto3 (SDK for python);
- Environment: The first version of ChaosXploit should be installed on a virtual environment with *python3.7* and *Chaostoolkit* installed.

5.2. Definition of the Knowledge Database

In Figure 3 it is possible to observe the attack tree designed for this experimental session. In this case, ChaosXploit is used as an internal auditing tool where a user with the role of an attacker can follow the four paths shown in the attack tree. These paths are described as:

- **Branch 1**: In this case, the intruder first locates public buckets by either listing the names or by using search engines such as the Wayback Machine. The next step aims to verify whether the attacker is successful in connecting to the bucket. Once inside, he has access to look at the storage system’s objects, and read the Access Control Lists (ACL). The attacker will be able to accomplish the attack objective if these ACLs have permissions that are available to the general public;
- **Branch 2**: In this route, the attacker tries to access private buckets using privilege escalation after failing to recognize public buckets. A policy rollback in this situation, where a user with permission to restore a previous policy is requested, presents a chance for privilege escalation. In a perfect world, this user would have had administrator rights or full access to the S3 service;
- **Branch 3**: in which the attacker can use brute-forcing techniques to compromise other user’s credentials and thereby gain access;
- **Branch 4**: where the attacker can use social engineering techniques such as phishing to compromise credentials and gain access.

It is important to note that the execution of the first and second branches was included in the scope of this project, as the actions included in such branches were easier to automate. Other branches could also be implemented through a combination of manual and automatic actions.

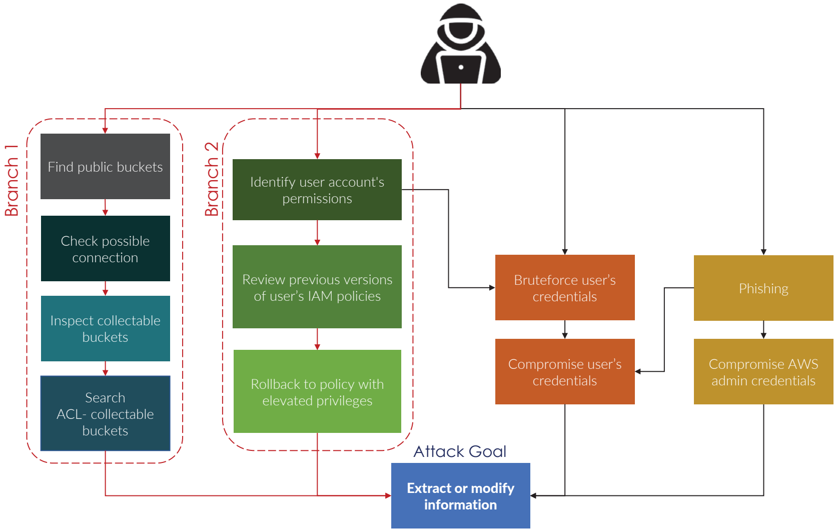

Branch 1
Find public buckets
Check possible connection
Inspect collectable buckets
Search ACL- collectable buckets
Branch 2
Identify user account's permissions
Review previous versions of user's IAM policies
Rollback to policy with elevated privileges
Bruteforce user's credentials
Compromise user's credentials
Phishing
Compromise AWS admin credentials
Attack Goal
Extract or modify information

Figure 3. Designed attack tree for the experimental scenario, highlighting the implemented branches.

5.3. Results of ChaosXploit's Execution of Branch 1: Exploitation of Public Buckets

5.3.1. Description

The reason for this experiment is that data can be stored on Amazon S3 and safeguarded from illegal access using encryption techniques and access management software. However, the shared responsibility model of cloud services has caused security configuration errors by the designers of this sort of storage. Exposing the data to the public endangers its availability, confidentiality, and integrity.

Based on the goal of the attack tree (Extract or modify Information), it is possible to define this first experiment following the CE method as follows:

- **Observability**: AWS S3 Buckets that can be found publicly;
- **Steady State**: The buckets to be analyzed suggest having the access controls properly configured;
- **Hypothesis**: If you try to access the objects stored in the buckets, then you will not be able to see their contents or the associated access controls since they are properly configured to prevent information leaks.

Below is a description of how the first branch of the attack tree specified for this scenario was implemented and **executed**. First, by taking regular expressions into account, public buckets were discovered using enumeration approaches. Since Amazon S3 has established some specifications for the bucket names, it is quite simple for an attacker to compile a list of them. Then, boto3, the AWS SDK for Python, was used to carry out the connection check. This stage allowed us to clean up the buckets, removing any that were empty or had incorrect names. Then, ChaosXploit looks at the buckets to see if their objects can be read, and lastly, it checks to see if any buckets provide access to the ACLs.

As shown in Table 3, three monitored variables were considered: (i) **Object-Collectable-Buckets**, which are the buckets that have public files such as pictures, documents, executable files, among others, which may be gathered through the experiment, (ii) **ACL-Collectable-Buckets** which refers to those buckets that have public ACLs, and can be accessed by anyone, and (iii) the **Permissions** obtained from the ACLs.

Table 3. Monitored variables and input parameters considered along the execution of branch 1 by ChaosXploit.

Monitored Variables	
Name	**Description**
Object-Collectable-Buckets	N° of buckets that have public objects and are accessible by anyone
ACL-Collectable-Buckets	N° of buckets that have public ACLs and are accessible by anyone
Permissions	N° of permissions obtained from the ACL.
Input Parameters	
Name	**Description**
Domain (Optional)	Domain name to which you want to identify the buckets
Threads	N° of execution threads
Mode	Object-Collectable-Buckets or ACL-Collectable-Buckets
Output	Output File

Regarding the input values, four were needed to execute the experiment. First, the *domain* is an optional input that should contain the name of the organization to be analyzed. We have considered this option since ChaosXploit can be used as an internal audit tool. Therefore, with this argument, the enumeration of the buckets will be limited to all those that are related to the given domain. In case this input is not provided, ChaosXploit will generate a list of names using brute-force, wordlists, and bucket naming rules defined by AWS. Second, the number of *threads* is considered as an input, so that the process of connecting and reading the buckets' information may be performed in parallel on the different cores, according to the defined thread's value. Third, the *mode* indicates the type of analysis to be performed, whether it aims to find *Object-Collectable-Buckets* or *ACL-Collectable-Buckets*. The last input, *output*, is a file name used to store the results and feed the ChaosXploit continuous validator.

5.3.2. Results Analysis

ChaosXploit's functionality was tested using a list of 3k buckets obtained through a bucket name enumeration process, which can be performed using automated tools.

As seen in the upper left part of Figure 4, all possible actions of the first branch of the attack tree presented in Section 5.2 were executed by ChaosXploit. It is possible to identify that for the second action (Check possible connection), out of the 3k buckets listed, 271 did not allow a connection. This is because the bucket no longer existed or had an invalid name, e.g., it did not follow the common bucket naming characteristics proposed by AWS. This leaves us with 2729 buckets to be tested.

In the case of the third act of the branch (Inspect collectible buckets), 2454 buckets were well configured and passed the steady-state defined in our experiment, since they did not allow reading files or permissions listed in the ACLs. However, 275 did not pass validation.

The lower left part of Figure 4 shows the file extensions that were extracted from the 252 Object-Collectable-Buckets. From each bucket, only the first 50 objects were collected, since some buckets had more than 100,000 files stored, for a total of 7465 collected files. Of all these files it was possible to identify that more than 2000 were images (jpg and png) and approximately 1250 were categorized as others because they could be log files, folders, or had no extension.

To analyze the users and user groups associated with each bucket we first need to know that Amazon S3 has a set of predefined groups:

- **AuthenticatedUsers group** representing all AWS accounts;
- **AllUsers group** allowing anyone in the world to access the resource;
- **LogDelivery group** allowing access logs to be written to the bucket.

Additionally, AWS also defines the following types of permissions:

- **READ** Allows the grantee to list the objects in the bucket;
- **WRITE** Allows the grantee to create new objects in the bucket. For the bucket and object owners of existing objects, it also allows deletions and overwrites of those objects;

- **READ_ACP** Allows the grantee to read the bucket ACL;
- **WRITE_ACP** Allows the grantee to write the ACL for the applicable bucket;
- **FULL_CONTROL** Allows the grantee the READ, WRITE, READ_ACP, and WRITE_ACP permissions on the bucket

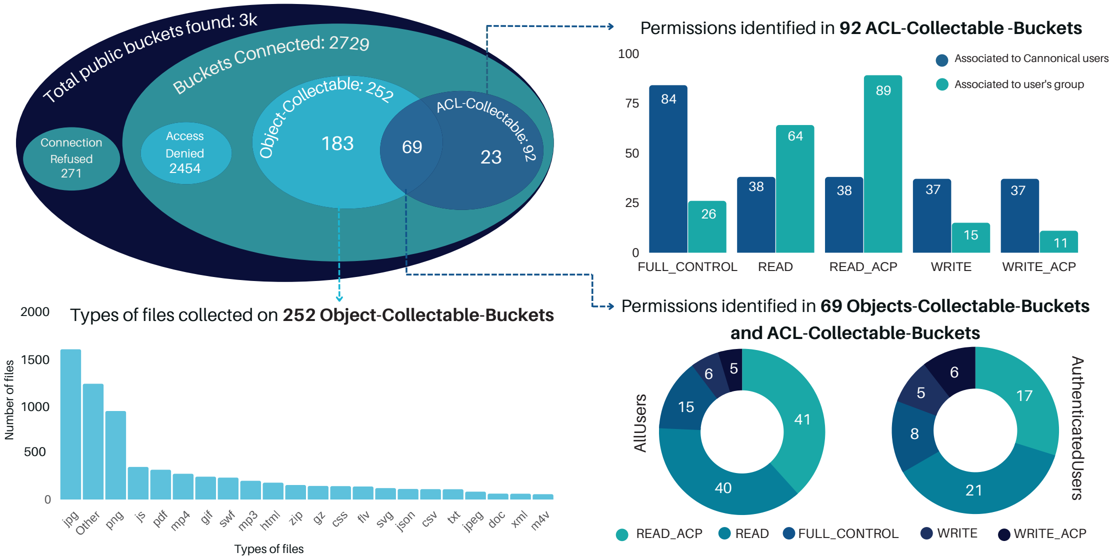


Figure 4. Results of the execution of ChaosXploit to achieve the defined attack goal (extract or modify information) through the branch .

In the upper right part of Figure 4 is possible to identify that 92 of the 257 buckets allowed the extraction of the ACLs. Up to 13 permissions per bucket were identified. Some of them showed information about the user who owned the bucket (known as **CanonicalUser** by AWS); others showed data about the users who belong to one of the predefined groups by AWS and had access to the bucket. Then, it is worth noting that for the information associated with canonical users, the FULL_CONTROL permission was enabled for 84 buckets (91.3%). In the case of the data associated with the users who belong to any of the groups, 64 (69.5%) of them allow the reading of the stored objects (READ permission) and 89 (96.7%) allow the reading of the ACLs (READ_ACP permission).

Finally, we analyze the results of those buckets that allowed the extraction of both objects and ACLs. As seen in the lower right part of Figure 4, 69 buckets (25%) allowed both tasks to be performed. These were filtered by the *AllUsers* and *AuthenticatedUsers* user groups and it was identified that 41 (38.3%) from the *AllUsers* group and 17 (29.8%) from the *AuthenticatedUsers* group were allowed to read the ACLs and the objects. Nevertheless, it was identified that 11 buckets (10.3%) from the *AllUsers* group and 11 buckets (19.3%) from the *AuthenticatedUsers* group allowed the modification of their content (WRITE permission) and the alteration of the ACLs (WRITE_ACP permission), indicating a big flaw that could severely compromise the confidentiality, integrity, and availability of the stored data.

With these results, we have noticed the importance of not only providing a tool for the detection of flaws or vulnerabilities but also seeing it as an aid to infer possible mitigations to prevent the exploitation of such vulnerabilities.

Table 4 shows the summary of the results considering the differences between traditional pentesting and SCE presented in Section 3.3. In this case, it is important to highlight that different tools (s3enum https://github.com/koenrh/s3enum (accessed on 11 October 2022), Sublist3r https://github.com/aboul3la/Sublist3r (accessed on 11 October 2022), bucketkicker https://github.com/craighays/bucketkicker (accessed on 11 October 2022)) may be integrated to ChaosXploit to execute this experiment, which allows us to enumerate the names of the buckets in an optimal way. After the bucket names are identified, ChaosXploit may perform the rest of the actions in a completely automated way. In addition, as we have refuted the hypothesis, ChaosXploit allows us to report a vulnerability related to misconfiguration because the security assumptions on the buckets have not passed the validation of the steady state of the experiment.

Table 4. Results of ChaosXploit's execution of branch 1 in terms of differences between traditional pentesting and SCE.

	SCE
People implementing	Executed by ChaosXploit's team
Methodology behind	Chaos Engineering principles
Security approach	Defensive
Available tools	ChaosXploit
Grade of automatization	All actions to be performed in this branch of the tree have been automated
Expected frequency	By definition, high frequency
Phase of SDLC where applied	Along all the SDLC
Scope of tests	Full test on the buckets list
Kind of vulnerabilities detected	Security assumptions about the configurations of the buckets

5.4. Results of ChaosXploit's Execution of Branch 2: Exploitation of Private Buckets

5.4.1. Description

This second branch refers to scenarios where the AWS policy administration in an organization is not working properly, and a user account maintains unnecessary policies, e.g., when a user changes role or area in a company. This scenario, caused by a misconfiguration in the IAM module, may be more critical when such a policy enables the user account to restore policies. Thus, the user may cause an elevation of privileges that allow him/her access to services and data in an unauthorized way. As part of the security inspection that a cybersecurity team could execute over a business infrastructure, one may assume that an internal attacker, e.g., an employee or contractor, could be interested in validating if his/her account allows the execution of policies additional to the required ones for the role. In addition, in the case of an external attacker, he/she could be interested in validating if some previously compromised AWS account, which contains limited permissions, can be elevated.

Considering the previous scenario, the following SCE definitions aligned to the scientific method are posed:

- **Observability**: List and status of the policies assigned to an AWS user account under analysis;
- **Steady State**: The AWS user account under analysis has policies assigned to him/her according to minimum privilege and need-to-know policies specific to his/her role in the organization;
- **Hypothesis**: Policies assigned to an AWS user account should not be modified in an unauthorized way.

For the **execution** of the second branch of the tree, ChaosXploit checks the policies assigned to the user account's profile defined for the experiment setup. If it identifies that the user account has the permission to restore previous versions of its policies, then it lists all the policy versions and searches for the one with elevated permissions to gain access to a privileged service, i.e., the AWS managed storage service (S3). This will achieve the goal of the attack tree: to extract or modify information. If the user does not have such a permission, ChaosXploit will start the execution of the third branch of the presented attack tree.

The upper part of Table 5 shows the two main variables that were monitored through the experiments of branch 2, i.e., *Attached-User-Policies* and *Current-Policy*. First, *Attached-User-Policies* is used at two moments of the branch execution: (i) at the beginning of branch 2 to identify all policies associated with a user account, and (ii) at the middle of branch 2 to identify permission associated with the user account that allows for the restoration of the previous version of policies and a previous policy that may be a suitable candidate to be restored, e.g., a policy that allows for the extraction and modification of information in the AWS S3 service. Second, *Current-Policy* represents the current version of the user's policy set, so this variable verifies whether the previous policy's restoration was successful.

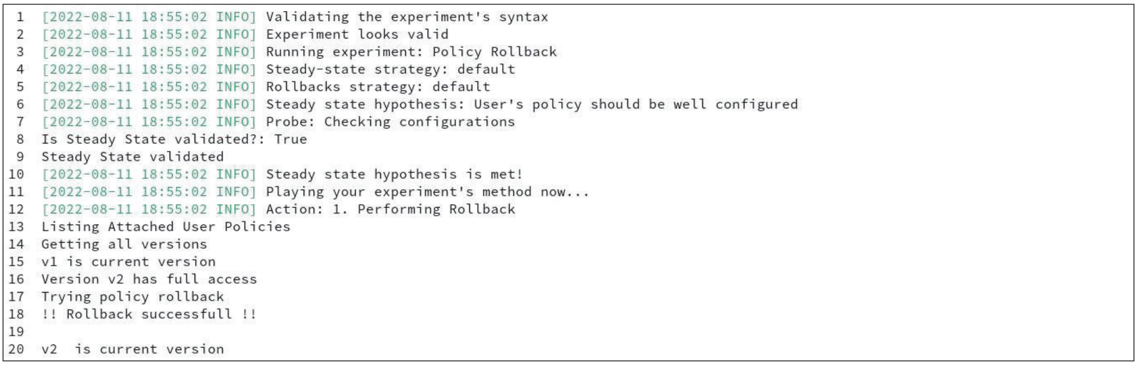
Table 5. Monitored variables and input parameters considered along the execution of branch 2 by ChaosXploit.

Monitored Variables	
Name	**Description**
Attached-User-Policies	Listing of policies assigned to a user
Current-Policy	Policy currently assigned to the user
Input Parameters	
Name	**Description**
User-Account	User account from which the actions will be performed
Output	Output file

On the other hand, the lower part of Table 5 shows the input elements that ChaosXploit receives for the execution of this branch. In this case, ChaosXploit uses the name of the user account (user account) for whom the security inspection must be performed. In addition, ChaosXploit takes as a parameter the name of the output file ($output$) for where to store the results.

5.4.2. Results Analysis

Figure 5 shows the execution of ChaosXploit for branch 2, which includes (i) the setup of ChaosXploit (lines 1–5), (ii) the steady state validation which assumes a correct configuration of the policies assigned to the user account under analysis (lines 6–10), (iii) execution of the actions that allow validating the hypothesis through an attempt to restore a previous policy (lines 11–20). This last set of lines includes listing the user policies (line 13–14), validating the current version (line 15), identifying the version that allows the privilege escalation (line 16), restoring the desired policy (line 17–18) and validation of the restore (line 20).



```
1   [2022-08-11 18:55:02 INFO] Validating the experiment's syntax
2   [2022-08-11 18:55:02 INFO] Experiment looks valid
3   [2022-08-11 18:55:02 INFO] Running experiment: Policy Rollback
4   [2022-08-11 18:55:02 INFO] Steady-state strategy: default
5   [2022-08-11 18:55:02 INFO] Rollbacks strategy: default
6   [2022-08-11 18:55:02 INFO] Steady state hypothesis: User's policy should be well configured
7   [2022-08-11 18:55:02 INFO] Probe: Checking configurations
8   Is Steady State validated?: True
9   Steady State validated
10  [2022-08-11 18:55:02 INFO] Steady state hypothesis is met!
11  [2022-08-11 18:55:02 INFO] Playing your experiment's method now...
12  [2022-08-11 18:55:02 INFO] Action: 1. Performing Rollback
13  Listing Attached User Policies
14  Getting all versions
15  v1 is current version
16  Version v2 has full access
17  Trying policy rollback
18  !! Rollback successfull !!
19
20  v2  is current version
```

Figure 5. Validation of the steady state and elevation of privileges achieved by ChaosXploit through branch 2.

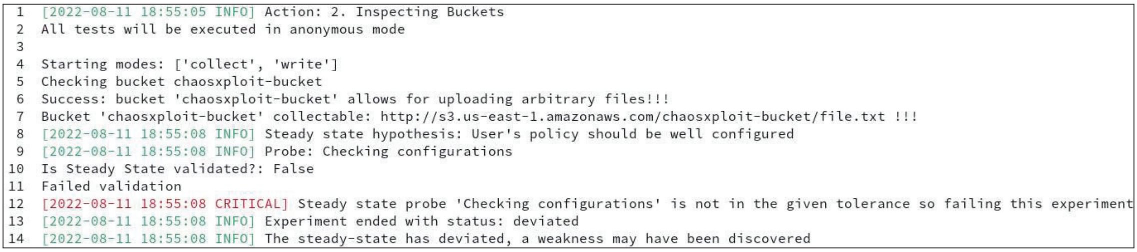
Table 6 shows the details of each of the policy versions found by ChaosXploit for the user account under analysis. This table lists the policy versions, the effects on the actions (either allow or deny access), the actions that indicate what the user can or cannot do, the resources on which the action may be applied, and additional conditions under which the policy has an effect. The current policy version (1) has limited actions related to the IAM service, but it still allows to change the policy through the action $SetDefaultPolicyVersion$.

It is also possible to identify the policy version 5, which includes some actions to manage the AWS S3 service. However, such actions would not allow reaching the attack goal because they do not allow modifying information. Finally, the version chosen by ChaosXploit (2) to be restored was the one that allows any action on any resource without any condition.

Table 6. Policy versions found by ChaosXploit through branch 2.

Version	Effect	Action(s)	Resource(s)	Condition
1 (Current)	Allow	"iam:Get*", "iam:List*", "iam:SetDefaultPolicyVersion"	*	None
2	Allow	*	*	None
3	Deny	*	*	IP Condition
4	Allow	"iam:Get*"	*	Time Condition
5	Allow	"s3:ListBucket", "s3:GetObject", "s3:ListAllMyBuckets"	*	None

Once the previous policy is restored, as shown in Figure 5, ChaosXploit initiates the actions shown in Figure 6. Between the first actions, ChaosXploit establishes the connection to the target and defines the *collect* mode to inspect the files in the bucket and the *write* mode to write a new file (lines 1–4). Additionally, ChaosXploit creates new files in the S3 bucket, as this experiment was being executed in its own controlled environment (lines 5–6). The validation of the steady state at lines 8–10 failed in this case as the policy settings can be manipulated and used to alter the information.



```
 1  [2022-08-11 18:55:05 INFO] Action: 2. Inspecting Buckets
 2  All tests will be executed in anonymous mode
 3
 4  Starting modes: ['collect', 'write']
 5  Checking bucket chaosxploit-bucket
 6  Success: bucket 'chaosxploit-bucket' allows for uploading arbitrary files!!!
 7  Bucket 'chaosxploit-bucket' collectable: http://s3.us-east-1.amazonaws.com/chaosxploit-bucket/file.txt !!!
 8  [2022-08-11 18:55:08 INFO] Steady state hypothesis: User's policy should be well configured
 9  [2022-08-11 18:55:08 INFO] Probe: Checking configurations
10  Is Steady State validated?: False
11  Failed validation
12  [2022-08-11 18:55:08 CRITICAL] Steady state probe 'Checking configurations' is not in the given tolerance so failing this experiment
13  [2022-08-11 18:55:08 INFO] Experiment ended with status: deviated
14  [2022-08-11 18:55:08 INFO] The steady-state has deviated, a weakness may have been discovered
```

Figure 6. Attack goal (extract or modify information) achieved by ChaosXploit through branch 2.

In experiments executed along branch 1 (Section 5.3) and branch 2 (Section 5.4), the attack goal was achieved so the experiments ended in a **critical** state similar to the one seen in line 11 at Figure 6. Table 7 shows the summary of the results for this second experiment, considering the differences between traditional pentesting and SCE presented in Section 3.3. In this case, we highlight the ChaosXploit capabilities to develop this kind of experiment that exploits the AWS authorization module. Additionally, we define the scope of the experiment only to users belonging to the same IAM account. Finally, as the experiment ended in a critical state, we report a vulnerability associated with privilege escalation, which allows a user to pass from few to many permissions, putting the confidentiality and integrity of the information available in the different AWS services at risk.

Table 7. Results of ChaosXploit's execution of branch 2 in terms of the differences between traditional pentesting and SCE.

	SCE
People implementing	Executed by ChaosXploit's team
Methodology behind	Chaos Engineering principles
Security approach	Defensive
Available tools	ChaosXploit
Expected frequency	By definition, high frequency
Phase of SDLC where applied	Along all the SDLC
Scope of tests	Users belonging to the IAM account
Kind of vulnerabilities detected	Privilege escalation considering the policy versions assigned to users

6. Toward an Adoption of SCE in Industry

With the growing adoption of CE, many companies have included it as a discipline for improving reliability. According to InfoQ [41], the appropriation of CE practices to inject failures and generate resilience has evolved to the "Early Majority stage", which means that its adoption is about one-third of the overall population. Gremlin, Litmus, and Steadybit are some key CE initiatives that have contributed to this achievement.

The stories of the adoption of CE reported by companies such as Capital One, Linkedin, Google, and Microsoft [34] are examples of its wide acceptance. The appropriation of CE as a common discipline to inject failures and generate resilience provides arguments to justify the success of this discipline between industry and academia.

Not only have the failures of the infrastructure attracted the attention of practitioners, but data breaches and security incidents have risen in recent years [42]. Failure to implement basic configurations and appropriate security controls have led to causes that contribute to the security incidents. Undoubtedly organizations are being asked to produce with extremely high throughput and with very little resources to maintain the security status quo. All the while, there is a divergent gap in how we design and build distributed systems and approach security engineering.

In this sense, SCE serves as a foundation for developing a learning culture around how organizations build, operate, instrument, and secure their systems. The goal of these experiments is to move security in practice from subjective assessment into objective measurements. As they do in CE, Security Chaos experiments allow security teams to reduce the "unknown unknowns" and replace "known unknowns" with information that can drive improvements to security posture. The promise in terms of adoption and sophistication is immense.

Even though introducing false positives into production networks and other infrastructures under the context of CE is a common practice nowadays, SCE is still seen as more of an academic research topic than industry practice. Nevertheless, in recent years, SCE is starting to become known in the industry. One example is the Thoughtworks report [43], which documented an evolution around this technique migrating SCE from a phase of "Assess" to "Trial", which means that SCE could be eventually used in a controlled way and validated that the security policies in place are robust enough to handle common security failure modes.

Another remarkable example of the application of SCE in the industry was documented by Jamie Dicken [44]. She wrote about her SCE journey at Cardinal Health, a global Fortune 20 healthcare manufacturer and distributor of medical and laboratory products and a provider of performance and data solutions for healthcare facilities. Cardinal Health needed an applied security model to protect critical infrastructure and data as it was moving to the cloud, and SCE became the most appropriate answer. Cardinal Health created a process named Continuous Verification and Validation (CVV) that, by using SCE, allowed them to continuously verify that security controls were working correctly and as expected.

Adopting SCE first requires a solid understanding of the principles of chaos. For example, insufficient observability of the chaotic experiments would impede drawing reliable statements about a hypothesis. After understanding the fundamentals, the next step should start by developing competency and confidence in the methods and tools needed to perform the SCE experiments. For this, a new SCE practitioner may decide to start designing small and manual experiments. In case the hypothesis is not disproved, we can automate the experiment. Here, ChaosXploit may play a key role as one of the few SCE platforms existing nowadays that may enable the industry to design and execute experiments aimed at the automatic and controlled exploitation of vulnerabilities and validation of systems security. Security validations can also be achieved progressively through security chaos game days that allow players to advance in this path without causing a security incident on production.

On the side, diverse teams should know and try SCE since it is no longer a limited concept for Security Engineers or security teams. We believe that if SCE begins as an engineering practice, it could be quickly adopted by other roles (Cloud Engineers, Software Engineers, Site Reliability Engineers) and teams (platform, infrastructure, operations, and application development) as it would allow them to improve the reliability of their applications through proactive testing of their own security.

7. Conclusions and Future Work

The digital revolution, or digital transformation, as it has been called in recent years, has proven to be an incredible driving factor in our society. Thanks to this revolution, our society was able to handle some of the most serious restrictions that the recent pandemic put on different essential services, e.g., the use of highly interactive e-health services in response to the restrictions regarding in-person medical consultations, exploitation of e-learning platforms to face the limitations in the physical access to formal educational services, enabling e-payments as an alternative to the use of traditional financial services, among many others.

On the downside, such a change also implies the existence of ill-motivated entities that constantly try to attack connected systems to damage the confidentiality, integrity, or availability of the provided online services. Such threat entities use increasingly advanced techniques, for example, based on malware campaigns [45] or threats addressed to a specific technology [46].

Over the last years, a novel paradigm has emerged, the so-called Chaos Engineering (CE), whose main objective consists of testing the resiliency of distributed and complex systems through continuous observation and experimentation. More recently, the paradigm has evolved to embrace the entire cybersecurity ecosystem, i.e., Security Chaos Engineering (SCE) comes into play to defend the system assets against cyberattacks through continuous and rigorous experimentations on possible security holes and consequent mitigations.

In this paper, we proposed ChaosXploit, an SCE-powered framework that can conduct SCE experiments on different target architectures. Based on the hypothesis generated by the knowledge database and the attack representations, ChaosXploit executes SCE experiments over a target to find a potential security problem as an ultimate goal. In addition, ChaosXploit features an observer that is in charge of verifying the change between the steady state of a certain hypothesis and the current state of the system. To prove the capabilities of ChaosXploit, a set of experiments was conducted on several AWS S3 buckets, evaluating their security characteristics with SCE. The results demonstrated that our approach could be successful, highlighting several unprotected buckets for a specific attack path. To foster its adoption, ChaosXploit was made publicly available for the cybersecurity community through the repository of the project [39].

Future work will explore the possibility of widening the ChaosXploit framework target architectures to include other use cases, systems, or providers. That is, the extension of the Attack Trees knowledge base is considered mandatory to include a number of different application scenarios, which can lead to the potential improvements of ChaosXploit, too. Particularly, one could easily argue that using a standardized attack modeling methodology (e.g., MITRE ATTC&K [47]) would be beneficial for the proposed SCE framework, even if some adjustments are needed to achieve full compliance. Besides, integrating a recommendation module to suggest countermeasures once a security flaw is discovered is worth investigating. In this sense, several attack models have been proposed in the literature so far, and some of them already integrate the Attack Trees representation adding countermeasures (e.g., Attack Countermeasures Trees [48], Attack Response Trees [49], etc.). Thus, ChaosXploit may incorporate those representations in the Knowledge base and select the optimal reaction to fire against the threat based on specific criteria [50]. Moreover, the performance of ChaosXploit should be further evaluated to prove its usefulness in performance-demanding or critical scenarios. Expressly, the assessment of the response time and resource consumption is essential to argue the applicability of the presented framework in scenarios where the threat discovery procedure must be executed in real-time or with limited computation capabilities.

Author Contributions: Conceptualization, S.P.C., P.N. and D.D.-L.; methodology, S.P.C., P.N. and D.D.-L.; software, S.P.C.; validation, P.N. and D.D.-L.; formal analysis, S.P.C., P.N. and D.D.-L.; investigation, S.P.C. and P.N.; resources, D.D.-L.; data curation, S.P.C.; writing—original draft preparation, S.P.C., P.N., D.D.-L. and Y.N.R.; writing—review and editing, P.N. and D.D.-L.; visualization, S.P.C.; supervision, P.N. and D.D.-L.; project administration, D.D.-L.; funding acquisition, D.D.-L. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by Universidad del Rosario (Bogotá) through the project "IV-TFA043—Developing Cyber Intelligence Capacities for the Prevention of Crime" and through "Becas para Estancias de Docencia e Investigación. Universidad del Rosario".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

SCE	Security Chaos Engineering
ICT	Information and Communication Technology
LEA	Law Enforcement Agencies
SRE	Site Reliability Engineering
SLI	Service Level Indicator
SLO	Service Level Objective
CE	Chaos Engineering
SPL	Software Product Line
API	Application Programming Interface
AWS	Amazon Web Services
GCP	Google Cloud Platform
CVSS	Common Vulnerability Scoring System
OWASP	Open Web Application Project
SoS	System of Systems
VUAV	Virtual Unmanned Aerial Vehicle
CRD	Custom Resource Definition
CTK	ChaosToolKit
ACL	Access Control List
SDK	Software Development Kit
CVV	Continuous Verification and Validation

References

1. Rodríguez, J.I.; Durán, S.R.; Díaz-López, D.; Pastor-Galindo, J.; Mármol, F.G. C3-Sex: A Conversational Agent to Detect Online Sex Offenders. *Electronics* **2020**, *9*, 1779. [CrossRef]
2. Sánchez, P.; Huertas, A.; Bovet, G.; Martínez, G.; Stille, B. An ML and Behavior Fingerprinting-based Framework for Cyberattack Detection in IoT Crowdsensing Platforms. In Proceedings of the VII Jornadas Nacionales de Investigación en Ciberseguridad (JNIC), Bilbao, Spain, 27–29 June 2022; Volume 1, pp. 188–191.
3. Botello, J.V.; Mesa, A.P.; Rodríguez, F.A.; Díaz-López, D.; Nespoli, P.; Mármol, F.G. BlockSIEM: Protecting Smart City Services through a Blockchain-based and Distributed SIEM. *Sensors* **2020**, *20*, 4636. [CrossRef] [PubMed]
4. Díaz-López, D.; Dólera-Tormo, G.; Gómez-Mármol, F.; Martínez-Pérez, G. Managing XACML systems in distributed environments through Meta-Policies. *Comput. Secur.* **2015**, *48*, 92–115. [CrossRef]
5. Useche-Peláez, D.E.; Sepúlveda-Alzate, D.; Díaz-López, D.O.; Cabuya-Padilla, D.E. Building malware classificators usable by State security agencies. *Iteckne* **2018**, *15*, 107–121. [CrossRef]
6. Pastor-Galindo, J.; Sáez, R.; Maestre, J.; Sotelo, M.; Gómez, F.; Martínez, G. Designing a platform for discovering TOR onion services. In Proceedings of the VII Jornadas Nacionales de Investigación en Ciberseguridad (JNIC), Bilbao, Spain, 27–29 June 2022; Volume 1, pp. 30–33.

7. Beyer, B.; Jones, C.; Petoff, J.; Murphy, N.R. *Site Reliability Engineering: How Google Runs Production Systems*, 1st ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2016.
8. Beyer, B.; Murphy, N.; Rensin, D.; Kawahara, K.; Thorne, S. *The Site Reliability Workbook: Practical Ways to Implement SRE*; O'Reilly Media: Sebastopol, CA, USA, 2018.
9. Principles of Chaos Engineering. Available online: https://principlesofchaos.org/ (accessed on 9 November 2022).
10. Pawlikowski, M. *Chaos Engineering: Site Reliability through Controlled Disruption*; Manning: Shelter Island, NY, USA, 2021.
11. Díaz-López, D.; Blanco Uribe, M.; Santiago Cely, C.; Tarquino Murgueitio, D.; Garcia Garcia, E.; Nespoli, P.; Gómez Mármol, F. Developing Secure IoT Services: A Security-Oriented Review of IoT Platforms. *Symmetry* **2018**, *10*, 669. [CrossRef]
12. Díaz-López, D.; Dólera Tormo, G.; Gómez Mármol, F.; Alcaraz Calero, J.M.; Martínez Pérez, G. Live digital, remember digital: State of the art and research challenges. *Comput. Electr. Eng.* **2014**, *40*, 109–120. [CrossRef]
13. Torkura, K.A.; Sukmana, M.I.; Cheng, F.; Meinel, C. CloudStrike: Chaos Engineering for Security and Resiliency in Cloud Infrastructure. *IEEE Access* **2020**, *8*, 123044–123060. [CrossRef]
14. Palacios, S.; Díaz-López, D.; Nespoli, P. ChaosXploit: A Security Chaos Engineering framework based on Attack Trees. In Proceedings of the VII Jornadas Nacionales de Investigación en Ciberseguridad (JNIC), Bilbao, Spain, 27–29 June 2022; Volume 1, pp. 130–137.
15. Basiri, A.; Behnam, N.; de Rooij, R.; Hochstein, L.; Kosewski, L.; Reynolds, J.; Rosenthal, C. Chaos Engineering. *IEEE Softw.* **2016**, *33*, 35–41. [CrossRef]
16. Camacho, C.; Cañizares, P.C.; Llana, L.; Núñez, A. Chaos as a Software Product Line—A platform for improving open hybrid-cloud systems resiliency. In *Software—Practice and Experience*; Wiley: Hoboken, NJ, USA, 2022; pp. 1–34. [CrossRef]
17. Simonsson, J.; Zhang, L.; Morin, B.; Baudry, B.; Monperrus, M. Observability and chaos engineering on system calls for containerized applications in Docker. *Future Gener. Comput. Syst.* **2021**, *122*, 117–129.
18. Jernberg, H.; Runeson, P.; Engström, E. Getting started with chaos engineering—Design of an implementation framework in practice. In Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'20), Bari, Italy, 5–9 October 2020; Volume 1211704, p. 10. [CrossRef]
19. Zhang, L.; Morin, B.; Haller, P.; Baudry, B.; Monperrus, M. A Chaos Engineering System for Live Analysis and Falsification of Exception-Handling in the JVM. *IEEE Trans. Softw. Eng.* **2021**, *47*, 2534–2548.
20. ChaoSlingr: Introducing Security into Chaos Testing. Available online: https://github.com/Optum/ChaoSlingr (accessed on 9 November 2022).
21. Rinehart, A.; Shortridge, K. *Security Chaos Engineering Gaining Confidence in Resilience and Safety at Speed and Scale*; Technical Report; O'Reilly Media: Sebastopol, CA, USA, 2021.
22. Torkura, K.A.; Sukmana, M.I.; Cheng, F.; Meinel, C. Security Chaos Engineering for Cloud Services: Work in Progress. In Proceedings of the 2019 IEEE 18th International Symposium on Network Computing and Applications, NCA 2019, Cambridge, MA, USA, 26–28 September 2019. [CrossRef]
23. Torkura, K.A.; Sukmana, M.; Cheng, F.; Meinel, C. Continuous auditing and threat detection in multi-cloud infrastructure. *Comput. Secur.* **2021**, *102*, 102124. [CrossRef]
24. Sharieh, S.; Ferworn, A. Securing APIs and Chaos Engineering. In Proceedings of the 2021 IEEE Conference on Communications and Network Security (CNS), Tempe, AZ, USA, 4–6 October 2021; pp. 290–294. [CrossRef]
25. Bailey, T.; Marchione, P.; Swartz, P.; Salih, R.; Clark, M.; Denz, R. Measuring resiliency of system of systems using chaos engineering experiments. In Proceedings of the 2022 SPIE 12117, Disruptive Technologies in Information Sciences VI, Orlando, FL, USA, 3–7 April 2022; Volume 1211704, p. 26. [CrossRef]
26. Pierce, T.; Schanck, J.; Groeger, A.; Salih, R.; Clark, M.R. Chaos engineering experiments in middleware systems using targeted network degradation and automatic fault injection. In Proceedings of the Open Architecture/Open Business Model Net-Centric Systems and Defense Transformation 2021, Online, 12–17 April 2021; Suresh, R., Ed.; International Society for Optics and Photonics, SPIE: Bellingham, WA, USA, 2021; Volume 11753, p. 117530A. [CrossRef]
27. The Netflix Simian Army. Available online: https://netflixtechblog.com/the-netflix-simian-army-16e57fbab116 (accessed on 14 March 2022).
28. Gremlin. Available online: https://www.gremlin.com/ (accessed on 10 November 2022).
29. Chaos Mesh. Available online: https://chaos-mesh.org/ (accessed on 10 November 2022).
30. Litmus. Available online: https://litmuschaos.io/ (accessed on 10 November 2022).
31. ChaosToolkit. Available online: https://chaostoolkit.org/ (accessed on 10 November 2022).
32. Chaos Engineering: The History, Principles, and Practice. Available online: https://www.gremlin.com/community/tutorials/chaos-engineering-the-history-principles-and-practice/ (accessed on 21 March 2022).
33. UnitedHealthGroup. Available online: https://www.unitedhealthgroup.com/ (accessed on 14 March 2022).
34. Rosenthal, C.; Jones, N. *Chaos Engineering: System Resiliency in Practice* ; O'Reilly Media: Sebastopol, CA, USA, 2020.
35. Verica. Available online: https://www.verica.io/ (accessed on 14 March 2022).
36. Nespoli, P.; Papamartzivanos, D.; Mármol, F.G.; Kambourakis, G. Optimal Countermeasures Selection Against Cyber Attacks: A Comprehensive Survey on Reaction Frameworks. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 1361–1396. [CrossRef]
37. Raj, S.; Walia, N.K. A Study on Metasploit Framework: A Pen-Testing Tool. In Proceedings of the 2020 International Conference on Computational Performance Evaluation (ComPE), Shillong, India, 2–4 July 2020; pp. 296–302. [CrossRef]

38. FOCA (Fingerprinting Organizations with Collected Archives). Available online: https://github.com/ElevenPaths/FOCA (accessed on 14 March 2022).
39. ChaosXploit. Available online: https://github.com/SaraPalaciosCh/ChaosXploit (accessed on 10 November 2022).
40. Rapid7. *2021 Cloud Misconfiguration Report*; Rapid7: Boston, MA, USA, 2021.
41. Wiggers, S.J. DevOps and Cloud InfoQ Trends Report. Available online: https://www.infoq.com/articles/devops-and-cloud-trends-2022/ (accessed on 10 November 2022).
42. *2018 Cost of Data Breach Study: Impact of Business Continuity Management*; Technical Report; Benchmark research sponsored by IBM; Ponemon Institute LLC: Traverse City, MI, USA, 2018.
43. ThougthWorks. Security Chaos Engineering. Available online: https://www.thoughtworks.com/radar/techniques/security-chaos-engineering (accessed on 10 November 2022).
44. Rinehart, A.; Shortridge, K.; Safari, a.O.M.C. *Security Chaos Engineering*; O'Reilly Media, Incorporated: Sebastopol, CA, USA, 2020.
45. Martínez Martínez, I.; Florián Quitián, A.; Díaz-López, D.; Nespoli, P.; Gómez Mármol, F. MalSEIRS: Forecasting Malware Spread Based on Compartmental Models in Epidemiology. *Complexity* **2021**, *2021*. [CrossRef]
46. Nespoli, P.; Díaz-López, D.; Gómez Mármol, F. Cyberprotection in IoT environments: A dynamic rule-based solution to defend smart devices. *J. Inf. Secur. Appl.* **2021**, *60*, 102878. [CrossRef]
47. Ahmed, M.; Panda, S.; Xenakis, C.; Panaousis, E. MITRE ATT&CK-Driven Cyber Risk Assessment. In Proceedings of the 17th International Conference on Availability, Reliability and Security, Vienna, Austria, 23–26 August 2022. [CrossRef]
48. Roy, A.; Kim, D.S.; Trivedi, K.S. Attack countermeasure trees (ACT): Towards unifying the constructs of attack and defense trees. *Secur. Commun. Netw.* **2012**, *5*, 929–943.
49. Zonouz, S.A.; Khurana, H.; Sanders, W.H.; Yardley, T.M. RRE: A Game-Theoretic Intrusion Response and Recovery Engine. *IEEE Trans. Parallel Distrib. Syst.* **2014**, *25*, 395–406. [CrossRef]
50. Nespoli, P.; Mármol, F.G.; Vidal, J.M. A Bio-Inspired Reaction Against Cyberattacks: AIS-Powered Optimal Countermeasures Selection. *IEEE Access* **2021**, *9*, 60971–60996. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

big data and cognitive computing

Article

Assessment of Security KPIs for 5G Network Slices for Special Groups of Subscribers

Roman Odarchenko [1], Maksim Iavich [2], Giorgi Iashvili [2], Solomiia Fedushko [3,4,*] and Yuriy Syerov [3,4]

[1] Department of Telecommunication and Radioelectronic Systems, National Aviation University, 03058 Kyiv, Ukraine; odarchenko.r.s@ukr.net

[2] Department of Computer Science, Caucasus University, 0102 Tbilisi, Georgia; miavich@cu.edu.ge (M.I.); giiashvili@cu.edu.ge (G.I.)

[3] Department of Social Communication and Information Activity, Lviv Polytechnic National University, 79000 Lviv, Ukraine; yurii.o.sierov@lpnu.ua

[4] Department of Information Systems, Faculty of Management, Comenius University in Bratislava, 820 05 Bratislava, Slovakia

* Correspondence: solomiia.s.fedushko@lpnu.ua

Abstract: It is clear that 5G networks have already become integral to our present. However, a significant issue lies in the fact that current 5G communication systems are incapable of fully ensuring the required quality of service and the security of transmitted data, especially in government networks that operate in the context of the Internet of Things, hostilities, hybrid warfare, and cyberwarfare. The use of 5G extends to critical infrastructure operators and special users such as law enforcement, governments, and the military. Adapting modern cellular networks to meet the specific needs of these special users is not only feasible but also necessary. In doing so, these networks must meet additional stringent requirements for reliability, performance, and, most importantly, data security. This scientific paper is dedicated to addressing the challenges associated with ensuring cybersecurity in this context. To effectively improve or ensure a sufficient level of cybersecurity, it is essential to measure the primary indicators of the effectiveness of the security system. At the moment, there are no comprehensive lists of these key indicators that require priority monitoring. Therefore, this article first analyzed the existing similar indicators and presented a list of them, which will make it possible to continuously monitor the state of cybersecurity systems of 5G cellular networks with the aim of using them for groups of special users. Based on this list of cybersecurity KPIs, as a result, this article presents a model to identify and evaluate these indicators. To develop this model, we comprehensively analyzed potential groups of performance indicators, selected the most relevant ones, and introduced a mathematical framework for their quantitative assessment. Furthermore, as part of our research efforts, we proposed enhancements to the core of the 4G/5G network. These enhancements enable data collection and statistical analysis through specialized sensors and existing servers, contributing to improved cybersecurity within these networks. Thus, the approach proposed in the article opens up an opportunity for continuous monitoring and, accordingly, improving the performance indicators of cybersecurity systems, which in turn makes it possible to use them for the maintenance of critical infrastructure and other users whose service presents increased requirements for cybersecurity systems.

Keywords: 5G network; communication systems; transmitted data; hybrid warfare; cybersecurity; security systems; cellular networks

Citation: Odarchenko, R.; Iavich, M.; Iashvili, G.; Fedushko, S.; Syerov, Y. Assessment of Security KPIs for 5G Network Slices for Special Groups of Subscribers. *Big Data Cogn. Comput.* **2023**, *7*, 169. https://doi.org/10.3390/bdcc7040169

Academic Editors: Peter R.J. Trim and Yang-Im Lee

Received: 17 September 2023
Revised: 22 October 2023
Accepted: 23 October 2023
Published: 26 October 2023

1. Introduction

It is clear that 5G networks have become an integral part of today's digital society. This technology is already implemented in many places worldwide and continues to be implemented rapidly, offering many benefits for ordinary users of cellular networks (standard services) and business and specialized services (government communications, military, firefighters, etc.). In the context of the latest special user introductions, 5G provides high throughput, low latency, and fairly high levels of reliability, opening up many opportunities for special missions and entirely new use cases. For example, 5G technology allows specific services to provide mission-critical communications whenever needed. It is clear that as specialized users implement more sensors, services, and subscribers, there may be additional operational needs, such as cybersecurity. It has become critical in the modern world, full of all kinds of threats, from single hackers to entire groups and even states. In this case, a single converged network capable of managing all of these functions gives operators the flexibility and control to manage high-bandwidth and low-latency applications while maintaining the required level of cybersecurity.

With emerging technologies such as artificial intelligence and machine learning, 5G's potential is truly impressive. It can provide special users with improved situational awareness, allowing entire units and platforms to respond faster and more accurately to threats in a dynamic environment. Furthermore, 5G's below-millisecond latency and reliability mean it can fit into various military and other government use cases.

The problem is that existing 5G communication systems cannot fully ensure the required quality of government line data service and the security of transmission in the widespread use of the concept of the Internet of Things, as well as in the context of hostilities, hybrid warfare, and cyberwar. Now, it is possible to intercept text messages, listen to conversations, and then use the data obtained against individuals and the military, government, etc. In addition, a remarkable landscape of other cyberattacks has appeared over the last decade. The current 5G network increases the range and adaptability of various services but also faces numerous security and privacy issues from attackers inside and outside the system perimeter. For example, 35 types of cyber threats were identified that pose significant risks in different areas of cybersecurity [1,2]: confidentiality, authentication, integrity, and availability in networks. This creates new serious threats that may become critical in the future. For example, an attacker can initiate eavesdropping to intercept data packets, conduct man-in-the-middle attacks to obtain session keys, or conduct location-tracking attacks on legitimate subscribers. These external threats that undermine the security of services for special users, the Internet of Things, etc., are the main security threats for every component in the structure of the modern 5G network, which is focused on providing high-quality services to its users. All this indicates the low efficiency of the applied methods of 5G network planning, the imperfection of the applied security technologies for the most secure data transmission, and the lack of ability to respond quickly to cyber incidents, etc.

The most spread-specific challenges and vulnerabilities in existing 5G communication systems that hinder the quality of service and data security for government lines and IoT applications were collected and reflected in Table 1.

Table 1. Specific challenges and vulnerabilities in existing 5G communication systems.

# Challenge	Security Threat	Target Point/Network Element	Effected Technology			Links	Privacy
			SDN	NFV	Cloud		
1.	DoS attack	Centralized control elements	+	+	+		
2.	Hijacking attacks	SDN controller, hypervisor	+	+			
3.	Signaling storms	5G core network elements			+	+	
4.	Resource (slice) theft	Hypervisor, shared cloud resources		+	+		
5.	Configuration attacks	SDN (virtual) switches, routers	+	+			

Table 1. *Cont.*

# Challenge	Security Threat	Target Point/Network Element	Effected Technology			Links	Privacy
			SDN	NFV	Cloud		
6.	Saturation attacks	SDN controller and switches	+				
7.	Penetration attacks	Virtual resources, clouds	+		+		
8.	User identity theft	User information data bases			+		+
9.	TCP level attacks	SDN controller-switch communication	+			+	
10.	Man-in-the-middle attack	SDN controller-communication	+			+	+
11.	Reset and IP spoofing	Control channels				+	
12.	Scanning attacks	Open air interfaces				+	+
13.	Security keys exposure	Unencrypted channels				+	
14.	Semantic information attacks	Subscriber location				+	+
15.	Timing attacks	Subscriber location			+		+
16.	Boundary attacks	Subscriber location					+
17.	IMSI catching attacks	Base station, identity registers				+	+

Therefore, scientifically based planning and optimization of cellular network security systems that provide the requested services with specified performance indicators for special groups of subscribers (transmission speed, delay, security of transmitted data) is a very complex scientific, technical, and economic problem, without which it is impossible to create an information infrastructure that meets the needs of a developed world-class information society.

As a leading standardization body in the field, 3GPPP pays great attention to the problem of network slice management in 5G [3]. Then, 5GPPP considered network slice KPIs and issued the White Paper on KPI Measurement Tools from KPI Definition to KPI Validation Enablement. Complete 5G projects, or parts of them, are dedicated to managing network slices and monitoring them. For example, 5G-DRIVE [4] was partially dedicated to researching critical innovations in networking slicing, network virtualization, etc. Moreover, 5G-MoNArch [5] in Work Package 3 worked on resilience and security and therefore developed secure network services and slices for them.

Leading manufacturers of telecommunications equipment also pay significant attention to this topic. For example, Juniper Networks described their end-to-end solution to manage service quality [6], Accedian paid attention to the active monitoring of network slices and the appropriate tools [7], Emblasoft developed flexible testing and active monitoring for 5G slices [8], and Huawei issued a white paper on 5G network cutting self-management [9]. Also, many research papers are devoted to monitoring network slices, the measurement of KPIs, level of security, etc. [10], focusing on the security challenges of the implementation of network slices in 5G networks [11,12]. The authors proposed that network slice controllers support security by enabling security controls at different network layers. The researchers [13] proposed the AI-based approach for cybersecurity in network slices and provided a comprehensive analysis [14] of the division of the network to develop commercial needs and challenges in the network. In [15], the authors considered the strategy for deploying and integrating one or more network management software with managed services. Furthermore, in [16], the authors proposed a principally novel framework for 6G network slices.

As we found from the analysis of the above projects and articles, insufficient attention is paid to the problems of monitoring the performance indicators of network layer security systems.

The article offers an analysis of key performance indicators (KPIs) and provides security KPIs. The calculation model and the study of the corresponding KPIs are provided. The paper also offers the architecture of the system to collect and estimate security KPIs and make the most appropriate decision. The algorithm was developed that automatically checks the organization's security KPIs based on the corresponding parameters.

The rest of the paper is organized as follows. The next section of the paper analyzes existing related resources and concludes with a problem statement, the goals of the paper, and the establishment of subtasks.

2. Review of the Literature

In the paper [17], the authors propose minimized sets of security KPIs, focusing mainly on computing and memory resources. In the article, certain key performance indicators (KPIs) are intricately linked with the Management and Orchestration (MANO) framework, necessitating their definition as integral components of the said MANO orchestration.

In the paper [18], the authors define the main requirements and KPIs of 5G networks. The offered methodology's primary focus is providing diverse vertical sectors with ultra-reliable communication and minimizing latency. As a result, the authors provide the requirements and key performance indicators for 5G networks.

In the article [19], the main objective of the study is to stimulate future research towards the secure implementation of Machine Learning (ML) methodologies within 5G infrastructures and prospective wireless networks. In the papers [20,21], the authors offer an approach to increase the flexibility of key performance indicators in 5G networks. However, one of the crucial indicators, Network Availability, is not considered in the mentioned papers. This indicator's emphasis on network availability aligns with existing 5G practices that prioritize high availability through network slicing and virtualization. This technique ensures that critical services remain operational, even during security incidents or disruptions. In the papers [22–24], the security aspect of 5G networks is not fully covered.

In the paper [25], the main focus is on understanding and managing the quality and performance of services to meet the technical quality of service (QoS) and the quality of experience (QoE). One of the critical security KPIs of 5G networks is Mean Time to Detect (MTTD), which shows 5G's advanced monitoring capabilities, AI-driven analytics, and machine learning algorithms to contribute to a shorter MTTD than traditional methods. This enables security teams to identify potential threats faster and respond proactively. This security KPI is not used in the above-mentioned paper. Another essential security KPI is the Mean Time to Respond (MTTR). This KPI gives 5G's improved data processing capabilities and network speed, leading to a quicker MTTR when compared to conventional response methods. Faster data analysis and communication enable efficient incident investigation and remediation. The mentioned KPI can significantly increase the security of the level of services to fulfill the technical quality of the service working with QoS/QoE.

Another important KPI is Data Leakage Rate, which makes 5G's implementation of advanced encryption protocols and secure communication channels reduce the data leakage rate compared to less secure approaches. Robust encryption ensures the confidentiality of sensitive information during transmission, which is essential for the security level in

5G networks and is not presented in the articles [26,27], in which the authors perform experiments on optimizing monitoring processes in 5G networks.

Several key performance indicators (KPIs) for security are not completely represented in the articles [28,29]. Compared to traditional network security approaches, incident response time is not used in the documents. In addition, 5G's incident response time benefits from lower latency and higher data transfer rates. This allows security teams to detect and respond to incidents more quickly, reducing the time between identifying a threat and taking appropriate actions to mitigate it.

Key performance indicator Security Patch Management ensures faster and more efficient distribution of security patches and updates. It offers 5G's more rapid data transfer rates, enabling more efficient security patch management compared to slower network technologies. In the papers [30–32], the authors offer 5G network functions and characterize the performance of location management functions in 5G core networks. Security patch management provides better distribution of security patches, reducing exposure to known vulnerabilities and enhancing the network's overall security while working with the mentioned functions. In the papers [33,34], the security aspect is not fully covered, which is one of the essential aspects of building a 5G network infrastructure. The compliance indicator with security standards is vital for 5G network security. The security concepts of the 5G network are designed with security standards in mind, making them more compliant than the older approaches. Adherence to security standards ensures that best security practices are followed, reducing the likelihood of vulnerabilities.

In the paper [35], the authors show the open challenge of integrating satellites into 5G cellular networks. During the investigation of the open challenges of satellite integration into 5G networks, comparing the 5G network security KPIs with existing approaches is an important aspect, demonstrating how 5G leverages its inherent technological advantages to strengthen network security [36,37]. Integrating faster data transfer, improved data processing, and advanced security mechanisms contribute to better incident response, threat detection, authentication, intrusion prevention, data protection, and compliance with security standards.

Problem Statement

The main goal of this work is to develop a system to monitor security KPIs in fifth-generation and subsequent-generation cellular networks. It will give the possibility of continuous control and optimization of the network.

Achieving the set goal requires solving the following tasks:

1. Analysis of key performance indicators of 5G cellular networks.
2. Selection of optimal indicators that describe the state of cyber security in the cellular network.
3. Development of a mathematical apparatus to evaluate safety KPIs.
4. Improvement of the 4G/5G network core to ensure continuous monitoring of security KPIs.
5. Development of an algorithm and pseudocode for continuous monitoring and evaluation of safety KPIs.

3. Definition of Performance and Security KPIs

The development of advanced communication networks is based on the establishment of internationally accepted standards to ensure compatibility, cost-effectiveness, and widespread adoption. This collaboration aims to empower the European industry to lead the advancement of 5G standards and secure a minimum of 20% of the 5G SEP (standard essential patents) for development and use.

We have identified the benchmarks for the new network's operational characteristics:

- A thousand-fold increase in mobile data volume per unit area.
- Ten to a hundred times more connected devices.
- Ten to a hundred times higher average user data rate.

- A tenfold reduction in energy consumption.
- End-to-end latency of less than one millisecond.
- Universal 5G access, even in low-density regions.

This high-performance network will operate through a scalable management framework that enables the rapid deployment of innovative applications, including sensor-based solutions. It will also reduce network management operating expenses by at least 20% compared to current standards. Furthermore, the network will incorporate new lightweight yet robust security and authentication measures designed to address the challenges posed by pervasive multidomain visualized networks and services in the modern era.

The main categories of 5G key performance indicators (KPIs) typically include the following.

Enhanced Mobile Broadband (eMBB): This category focuses on improving mobile broadband services. Ultra-Reliable and Low-Latency Communications (URLLC) emphasizes reliable and low-latency communication, crucial for applications such as autonomous vehicles or remote surgery. Massive Machine-Type Communications (mMTC): This category addresses the requirements for connecting many IoT devices. ITU, NGMN, and 3GPP have globally characterized 5G use cases and related requirements since their development. Some 5G technology use cases include broadband access in densely populated areas, high user mobility, massive IoT connectivity, tactile Internet, support during natural disasters, electronic health services, and broadcast services.

Table 2 below summarizes the KPIs for 5G wireless technology at the ITU level, representing the minimum performance requirements:

Table 2. KPIs for 5G wireless technology at the ITU level [38].

The Type of 5G Performance Requirement	Minimum KPI Requirement and Category
Peak Spectral Efficiency	The downlink spectral efficiency is 30 bits per second per hertz (bps/Hz), whereas the uplink spectral efficiency is 15 bits per second per hertz (bps/Hz). (eMBB)
Peak Data Rate	The downlink speed for data transmission is 20 Gbps, while the uplink speed is 10 Gbps. (eMBB)
Area Traffic Capacity	In an indoor hotspot, the downlink data rate is 10 Mbps per square meter. (eMBB test environment)
Data Rate of User Experience	The downlink speed for data transmission is 100 Mbps, while the uplink speed is 50 Mbps. (eMBB)
Connection Density	106 devices/Km^2 (mMTC)
Latency (Control Plane)	The specified target latency is 20 milliseconds, with 10 milliseconds being the encouraged latency whenever possible. (eMBB, URLLC)
Latency (User Plane)	The specified latency requirement for enhanced mobile broadband (eMBB) is 4 milliseconds, whereas for ultrareliable low latency communications (URLLC), the latency target is 1 millisecond. (eMBB, URLLC)
Average Spectral Efficiency	Indoor coverage area with high-speed Internet: Download (DL) speed of 9 Mbps and upload (UL) speed of 6.75 Mbps. Dense urban coverage area: Download (DL) speed of 7.8 Mbps and upload (UL) speed of 5.4 Mbps. Rural coverage area: Download (DL) speed of 3.3 Mbps and Upload (UL) speed of 1.6 Mbps. (eMBB)
Reliability	1×10^{-5} the probability of successfully transmitting a layer-2 protocol data unit (PDU) consisting of 32 bytes in a 1 millisecond timeframe in an urban macro-URLLC test environment with edge channel coverage quality. (URLLC)
Energy Efficiency	Demonstrating Efficient Data Transmission (Loaded Case): The effectiveness of data transmission can be assessed by evaluating the "average spectral efficiency" metric. Minimizing Energy Consumption (No-Data Case): This test case aims to support a high sleep ratio and long sleep duration to achieve low energy consumption. It is designed to optimize the system for scenarios without data transmission. (eMBB)

Table 2. *Cont.*

The Type of 5G Performance Requirement	Minimum KPI Requirement and Category
Mobility	In a dense urban environment, the maximum speed considered is up to 30 Km/h, while in a rural setting it can reach up to 500 Km/h. (eMBB)
Mobility Interruption Time	0 ms (eMBB, URLLC)
Bandwidth (Maximum Aggregated System)	For operation in high-frequency bands (above 6 GHz), the minimum required bandwidth is at least 100 MHz, while the maximum supported bandwidth can reach up to 1 GHz. (IMT-2020)

Here are some of the key challenges and vulnerabilities that must be addressed during the design and deployment of 5G network services for special groups of subscribers.

1. Security concerns:
   - Spectrum vulnerability—the use of shared and unlicensed spectrum in 5G networks can make them susceptible to interference and jamming, which can disrupt government and IoT communications.
   - Cyberattacks—with more connected devices and a larger attack surface, the risk of cyberattacks, such as distributed denial of service (DDoS) attacks, increases, potentially affecting government and IoT services.
   - Device Vulnerabilities—IoT devices often have limited security features and can be vulnerable to hacking, compromising data security.
2. Privacy Concerns:
   - Data privacy—the massive amount of data generated by IoT devices, including personal information, can raise concerns about data privacy and unauthorized access, particularly in government applications.
   - Data Localization—governments may require data to be stored within their borders, creating challenges for global IoT deployments.
3. Compatibility and Interoperability:
   - Legacy systems—Integrating 5G with existing communication systems can be challenging, particularly for government agencies with legacy infrastructure.
   - IoT standards—The lack of universal IoT standards can hinder interoperability and create compatibility issues.
4. Risks to the supply chain:
   - Vendor Dependencies: Relying on specific vendors for 5G infrastructure or IoT devices can create supply chain vulnerabilities, especially if the vendors are from countries with conflicting interests.
5. Regulatory and Compliance Challenges:
   - Spectrum Regulations—Regulations and licensing for spectrum use can vary by region, complicating IoT device deployment and government communication systems.
   - Security and Privacy Regulations—Compliance with data security and privacy regulations, such as GDPR or HIPAA, can be complex, especially in cross-border scenarios.

Addressing these challenges and vulnerabilities in 5G communication systems for government lines and IoT applications requires a comprehensive approach that includes robust security measures, privacy protections, resilience, and interoperability. Collaboration between governments, industry stakeholders, and standardization bodies is crucial to effectively implement secure and reliable 5G and IoT solutions.

For today’s 5G networks, a new cybersecurity approach must be defined, and precise metrics must be established to inform all stakeholders about potential threats and breaches. Typically, the leaders of large cellular service consumers are looking for clear security metrics that demonstrate costs and anticipated potential impacts on their business goals. The following study results can be cited as an example of such losses. A breach lasting more than two hundred days has been shown to cost an organization 4.56 million USD, which is 37% more than the cost of a breach lasting less than two hundred days (3.34 million USD) [39].

Furthermore, the results of the study [39] showed that 44% of those surveyed said that their organization's security approach has improved significantly in recent years. Figure 1 lists the specific metrics companies used to measure this improvement. They mainly include the number of attacks prevented [40], the time taken to identify the incident, and the time required to locate the incident.

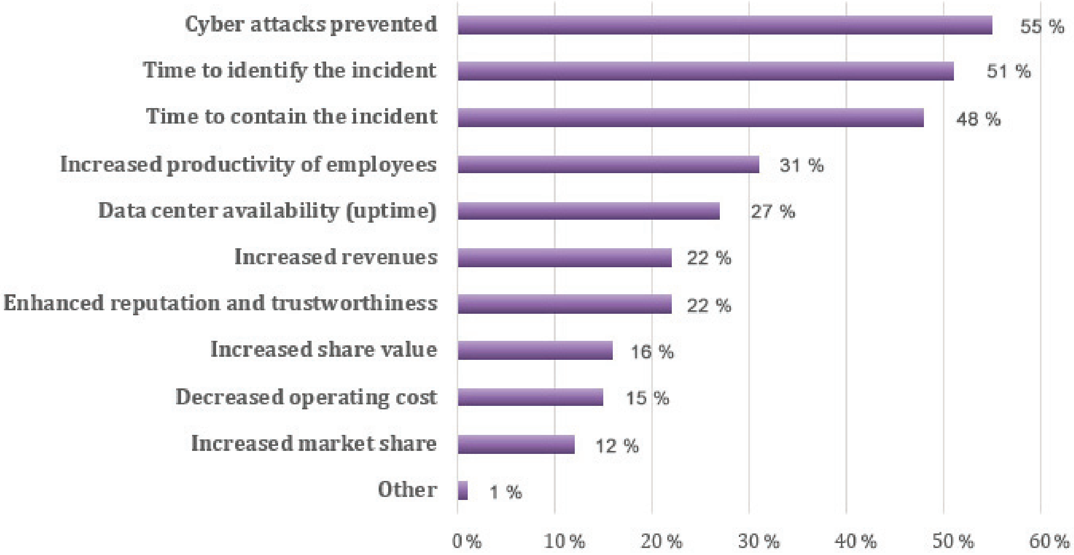


Figure 1. Results of the cyber security survey [41].

These KPIs outline the performance requirements for 5G wireless technology according to the ITU.

It is essential to determine security KPIs for 5G wireless networks. Security key performance indicators (KPIs) for 5G networks can help assess the effectiveness and efficiency of the security measures in place. Based on our research, we have identified the following security KPIs for 5G networks:

1. Incident Response Time: Measures the time taken to detect and respond to security incidents, such as network breaches or unauthorized access attempts.
2. Mean Time to Detect (MTTD): Measures the average time to detect security incidents or anomalies within the 5G network.
3. Mean Time to Respond (MTTR): Measures the average time it takes to respond and resolve security incidents or vulnerabilities identified within the 5G network.
4. Network availability: Measures the percentage of time the 5G network is available and operational without any security-related disruptions.
5. Network Resilience: Measures the ability of the 5G network to withstand and recover from security attacks or incidents without significant impact on network performance.
6. Authentication Failure Rate: Measures the percentage of failed authentication attempts within the 5G network, which can indicate potential security breaches or unauthorized access attempts.
7. Intrusion Detection and Prevention Effectiveness: Measures the accuracy and effectiveness of intrusion detection and prevention systems deployed within the 5G network in detecting and blocking security threats.

8. Data Leakage Rate: Measures the occurrence of data leaks or unauthorized access to sensitive information within the 5G network.
9. Compliance with Security Standards: Measures the level of compliance with security standards and regulations relevant to 5G networks, such as the 3GPP security specifications or industry best practices.
10. Security Patch Management: Measures the frequency and timeliness of applying security patches and updates to network equipment and software within the 5G network.

It is important to note that specific security KPIs may vary depending on the network operator, service provider, or organization that implements the 5G network. These KPIs can be tailored to suit the network infrastructure's specific security goals and requirements. To ensure the success of the concrete 5G business, it is crucial to establish a well-defined cybersecurity approach and use accurate metrics to inform relevant stakeholders. C-level executives and board members are actively looking for security metrics that clearly understand the costs involved and the anticipated impact on their business objectives. According to the IBM research findings [39], organizations experience a significantly higher cost of 4.56 million USD when a breach lasts more than two hundred days. This amount is 37% greater than the cost incurred when a breach is resolved in a shorter period, which is 3.34 million USD.

Furthermore, the study highlights that 44% of the respondents surveyed reported notable improvements in their organization's security approaches during the past 12 months. These metrics include primarily the number of prevented attacks, the time required to identify an incident, and the time required to contain an incident. Approximately 55%, 51%, and 48% of companies use these respective metrics for measurement purposes. Based on this study, we can identify the security KPIs for 5G networks. To effectively assess security operations, metrics such as Mean Time to Identification (MTTI) and Mean Time To Contain (MTTC) are considered essential to measure cybersecurity intrusions or incidents in 5G networks. Based on related articles, we have identified a set of main KPIs for security measures (Table 3).

Table 3. The most relevant 5G cybersecurity KPIs.

The Type of 5G Security Requirement	Minimum Security KPI Requirements	Formula/Symbol	Challenges Addressed (from Table 1)
Intrusion Attempts [42]	As a cybersecurity operative, you must monitor intrusion attempts on your organization's network. Similarly, you can regularly review your firewall logs to see if anyone has unauthorized access to the network.	NIA	1–3, 5–7, 9–17
Number of Security Incidents [43]	This KPI quantifies the total number of security incidents or breaches detected in the 5G network over a specific period. Monitoring this metric helps to track the security posture and identify trends or patterns.	NSI	1–17
Mean Time to Identification (MTTI) [43]	The whole process must take a maximum of 12 h.	$MTTI = \frac{S_{IT}}{N_I}$, where: S_{IT}—sum of identification times; N_I—number of incidents.	1–17
Mean Time To Contain (MTIC) [43]	The entire process must take a maximum of 12 h.	$MTIC = \frac{S_{CT}}{N_I}$, where: S_{CT}—sum of contain times N_I—number of incidents.	1–17
Mean Time to Recover (MTTR) [44]	This KPI measures the average recovery time from a security incident or breach. A shorter MTTR indicates effective incident response and recovery capabilities, minimizing impact on network operations.	$MTTR = \frac{S_{TTR}}{N_I}$, where: S_{TTR}—total time taken to recover from incidents N_I—number of incidents.	1–17

Table 3. *Cont.*

The Type of 5G Security Requirement	Minimum Security KPI Requirements	Formula/Symbol	Challenges Addressed (from Table 1)
Incident Response Time [43]	Aim for a rapid incident response time to ensure timely detection and mitigation of security incidents. A specific target can be set, such as responding to critical incidents within a defined timeframe (e.g., within 1 h).	$IRT = Timestamp_{IR} - Timestamp_{ID}$, where: $Timestamp_{IR}$—time of incident resolution; $Timestamp_{ID}$—time of incident detection.	1–3, 5–7, 9–17
Mean Time to Detect (MTTD) [45]	Strive to minimize the average time taken to detect security incidents. Setting a target, such as keeping the MTTD below a certain threshold (e.g., within 30 min), can help promptly identify potential threats.	$MTTD = \frac{S_{ID}}{N_I}$, where: S_{ID}—sum of detection times; N_I—number of incidents.	1–3, 5–7, 9–17
Mean Time to Respond (MTTRes) [45]	The aim is to minimize the average time taken to respond and resolve security incidents. Establishing a target, such as keeping the MTTR below a specific value (e.g., within 2 h), can help expedite incident resolution.	$MTTR = \frac{S_{RT}}{N_I}$, where: S_{RT}—sum of respond times; N_I—number of incidents.	1–3, 5–7, 9–17
Network Availability [46]	Aim for high network availability to minimize disruptions due to security incidents. Setting a target, such as maintaining network availability at a high percentage (e.g., 99.99%), ensures that security events do not significantly impact network services.	$NA = \frac{t_{up}}{t_{total}}$, where: t_{up}—total uptime; t_{total}—total time.	1–3, 9, 11
Authentication Failure Rate [47]	Try to keep the authentication failure rate as low as possible. Although the acceptable rate may depend on the specific network context, aiming for a minimal failure rate (for example, less than 1%) helps reduce the risk of unauthorized access.	$AFR = \frac{N_{AF}}{N_{AA}}$, where: N_{AF}—number of authentication failures; N_{AA}—total number of authentication attempts.	10, 17
Intrusion Detection and Prevention Effectiveness [48]	Implement robust intrusion detection and prevention systems with high accuracy rates. Regularly assess and monitor the effectiveness of these systems, with a goal of a high detection and prevention rate (for example, above 95%).	$TPR = \frac{N_{TP}}{N_{AI}}$, $TPR = \frac{N_{FP}}{N_{AI}}$, where: N_{TP}—number of true positives; N_{FP}—number of false positives; N_{AI}—total number of actual intrusions.	4, 7, 8, 10, 17
Data Leakage Rate [49]	Aim for a minimal data leakage rate within the 5G network. This can be achieved through solid access controls, encryption, and monitoring mechanisms. Setting a target, such as keeping the data leakage rate below a specific value (e.g., 0.5%), helps ensure data protection.	$DLR = \frac{N_{DLI}}{TV_{DH}}$, where: N_{DLI}—number of data leakage incidents; TV_{DH}—total volume of data handled.	13
Threat Detection Time [50]	This KPI measures the time it takes to detect and identify a security threat or intrusion on the 5G network. A shorter detection time indicates a more proactive and effective security system.	$T_D = T_{TD} - T_{TO}$, where: T_{TD}—time of threat detection; T_{TO}—time of threat occurrence.	1–17
Patching and Vulnerability Management [51]	This KPI evaluates the time to apply security patches and updates to address known vulnerabilities in the 5G network infrastructure. Correct patching helps minimize the risk of exploitation. Time taken to apply patches and updates	PVMT	1–17
Compliance with Security Standards [52]	This KPI evaluates the extent to which the 5G network adheres to relevant cybersecurity standards and regulations. Compliance with standards such as the 3GPP security specifications ensures a robust security posture.	$CR = \frac{N_{CRM}}{N_{TNCR}}$, where: N_{CRM}—number of compliance requirements met; N_{TNCR}—total number of compliance requirements.	1–17

Table 3 is a set of performance indicators for cybersecurity systems in cellular 4G/5G networks. It contains indicators that describe the state of security in the network as a whole and individual elements that describe the state of individual network elements. The table also includes both indicators (Intrusion Attempts) that need to be constantly measured. Their deviation may indicate the occurrence of a cybersecurity incident, as well as indicators that are measured over time and therefore require preliminary collection (accumulation) of information (number of Security Incidents, Mean Time To Identification, Mean Time To Contain, Mean Time to Identification, Mean Time to Detect, Mean Time to Respond, Network Availability, Authentication Failure Rate, Intrusion Detection and Prevention Effectiveness, Data Leakage Rate, Threat Detection Time, Patching, and Vulnerability Management). Their assessment indicates the need for comprehensive changes (possibly a revision of current approaches) in the security system. Such a KPI, like "Compliance with Security Standards", has to be fully satisfied and continuously reviewed (Table 4).

Table 4. Table of threshold values of security KPIs.

Security KPIs	**Network Slice-Type Thresholds**			
	Slice 1 (i.e., eMBB)	**Slice 2 (i.e., MCC)**	...	**Slice *N***
NIA	*NIA1*	*NIA2*	...	*NIAN*
NSI	*NSI1*	*NSI2*	...	*NSI2*
MTTI	*MTTI1*	*MTTI2*	...	*MTTIN*
MTTR	*MTTR1*	*MTTR2*	...	*MTTRN*
MTTD	*MTTD1*	*MTTD2*	...	*MTTDN*
MTTRes	*MTTRes1*	*MTTRes2*	...	*MTTResN*
NA	*NA1*	*NA2*	...	*NA3*
AFR	*AFR1*	*AFR2*	...	*AFRN*
TPR	*TPR1*	*TPR2*	...	*TPRN*
FPR	*FPR1*	*FPR2*	...	*FPRN*
DLR	*DLR1*	*DLR2*	...	*DLRN*
TDT	*TDT1*	*TDT2*	...	*TDTN*
PVMT	*PVMT1*	*PVMT2*	...	*PVMTN*
CR	*CR1*	*CR2*	...	*CRN*

Minimal KPI requirements can vary depending on the organization's specific risk appetite and security objectives.

4. Development of Architecture

To achieve low latency, high data transfer rates, and a higher level of security, the concept of network cutting was defined in 5G. This technology allows network operators to divide their physical infrastructure into multiple logical networks, each configured according to its characteristics and needs. As shown in Figure 2, each network layer is an independent virtual subnet from end to end and can even be owned by different tenants (or vertical markets) that manage the physical, virtualized, and service layers with different key performance indicators (KPIs), including security metrics.

Using emerging advances in virtualization and network management, such as software-defined networking (SDN) and network function virtualization (NFV), network partitioning creates virtual networks that provide a customized network experience that meets pre-defined key performance indicators (KPIs). Therefore, there are known security issues associated with these underlying SDN and NFV technologies and access networks. Thus, the central part of the security in the division of the network is to determine what constitutes the main potential threats to this segment, the establishment of minimum requirements, and their mandatory implementation. In this case, it is imperative to define isolation attributes, create an abstraction layer to provide end-to-end isolation at a particular level, and introduce appropriate security policies for each layer.

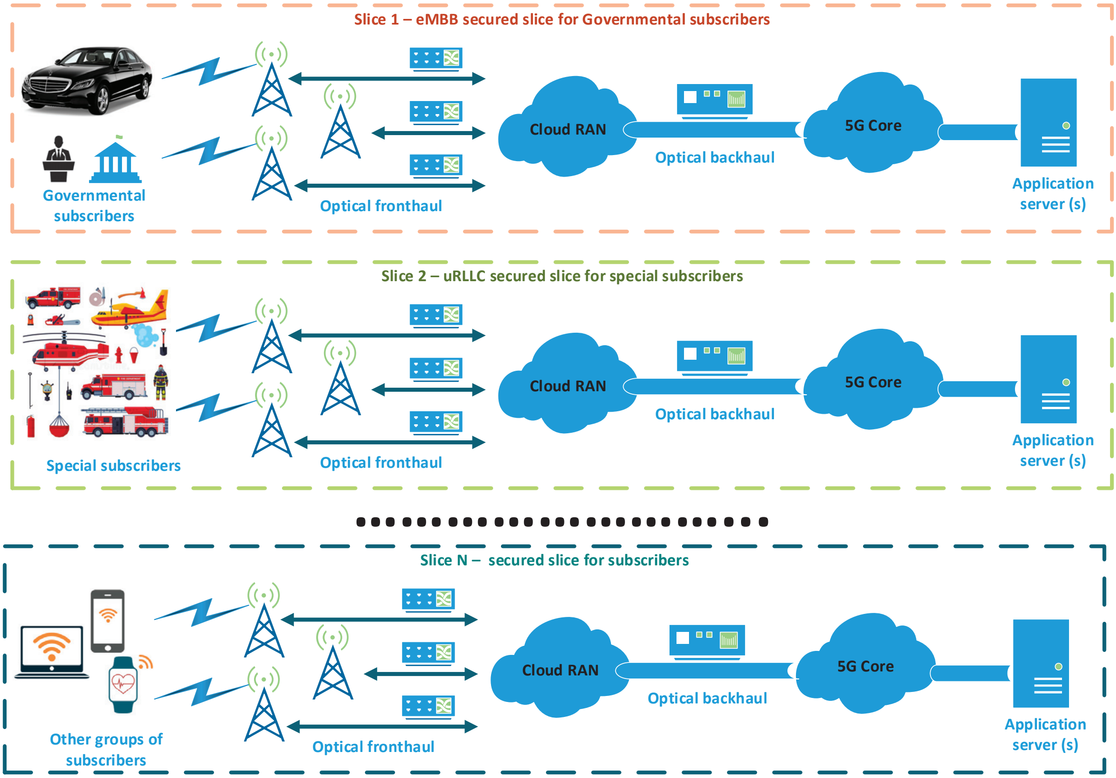


Figure 2. Network slices concept for the special subscribers' groups.

Therefore, an effective network partitioning solution requires integrated management, performance, and security considerations. In this case, attacks directed against one segment must not affect others. Therefore, security functions must act independently for each layer. Thus, the main challenge in designing a network partitioning solution is to satisfy all the requirements of the segment owner while ensuring the security of each segment independently.

As illustrated in Figure 3, a 5G network may accommodate different use cases, and each can be served by single or multiple network slices, which can be applied to monitoring mechanisms [53]. When the subscribers are geographically dispersed, dedicated or shared network slices can also serve the horizontal use cases. Each network slice owns logically isolated computation and storage resources to perform data processing and storage tasks for all use cases that receive their services. Each network layer, which must serve a specific group of subscribers to ensure the required quality of service and secure data transmission, is characterized by its specific network characteristics and network security indicators (KPIs). To respond immediately to emerging anomalies, degradation of service quality, or lowering the level of information security, it is necessary to continuously monitor the above parameters. This process is reflected in Figure 3. In addition, also it is also possible to perform forced penetration tests of layers. For these two procedures, a specialized network slices monitoring server can be used (Figure 3).

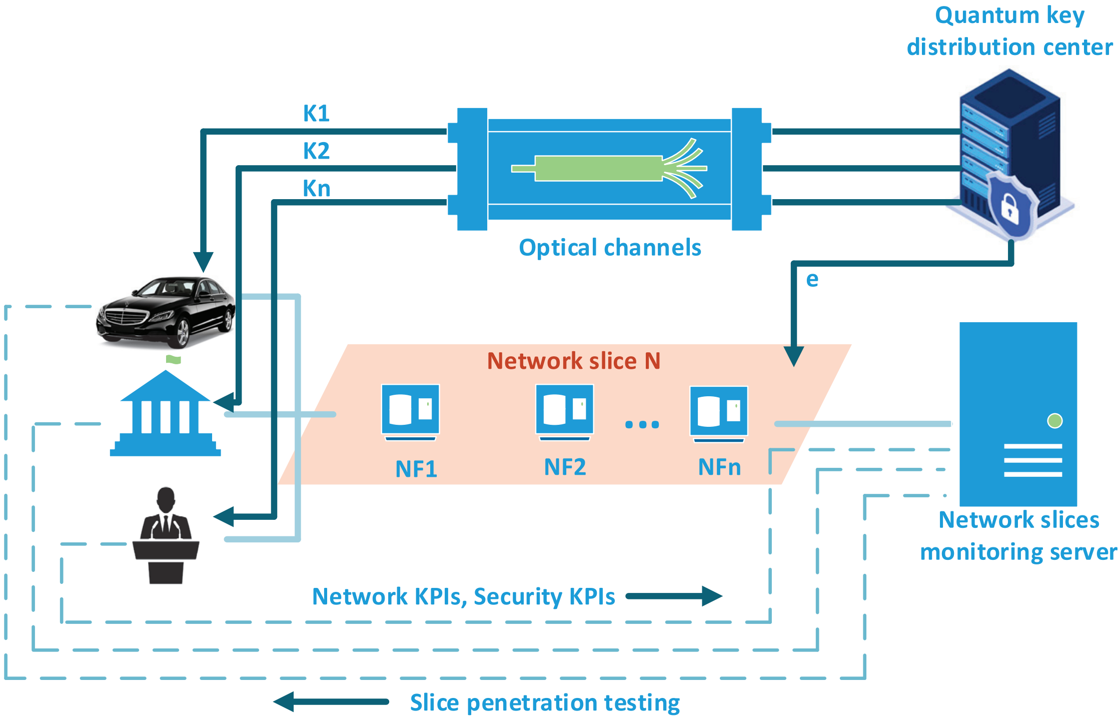

Quantum key distribution center
K1
K2
Kn
Optical channels
e
Network slice N
NF1
NF2
...
NFn
Network slices monitoring server
Network KPIs, Security KPIs
Slice penetration testing

Figure 3. Graphical representation of delivering security credentials in the key management scheme.

The operation of this system obviously must be in synchronization with the cybersecurity systems. As an example, the figure shows a case of potential use of a quantum key distribution system, described in detail in [54], to increase the confidentiality level of transmitted data. Thus, in the case of measuring security indicators and identifying problems, for example, with confidentiality, quantum fundamental distribution mechanisms can be used. However, in general, the study aims to describe a generalized model and, accordingly, the architecture of the monitoring system that will ensure the main security principles, traditionally categorized as confidentiality, authentication, authorization, availability, and integrity.

5. The Offered Model

Based on the above, using the security KPIs from Table 2, a set of safety KPIs for the evaluation analysis model is proposed, which can be objectively evaluated. There is a set of network layers for which both the QoS quality of service indicators and the security KPI indicators are clearly defined.

$$\left\{\overset{n}{\underset{i=1}{\cup}} Slice_i\right\} = \{Slice_1, Slice_2, I, Slice_n\},$$

where

- network layers: $Slice_i \subseteq Slice$, $(i = \underline{1, n})$, n is the number of these layers.
- $KPI_i^{sec} = \left\{\cup_{j=1}^{m_i} KPI_{i,j}^{sec}\right\} = \left\{KPI_{i,1}^{sec}, KPI_{i,2}^{sec}, \ldots KPI_{i,m_i}^{sec}\right\}$, $KPI_{i,j}^{sec}\,(j = \underline{1, m_i})$ is a subset KPI for cyber security systems (Table 2).

In order to collect information about any operations that occur on the network, analyze them, and, accordingly, make decisions based on the assessments made, it is proposed to add either an additional network function to the core of the network, which will contain all the functionality necessary for this or, more straightforward at first, especially for testing the system, is to add an external server that will be connected to the network core via standard interfaces. This approach is reflected in Figure 4.

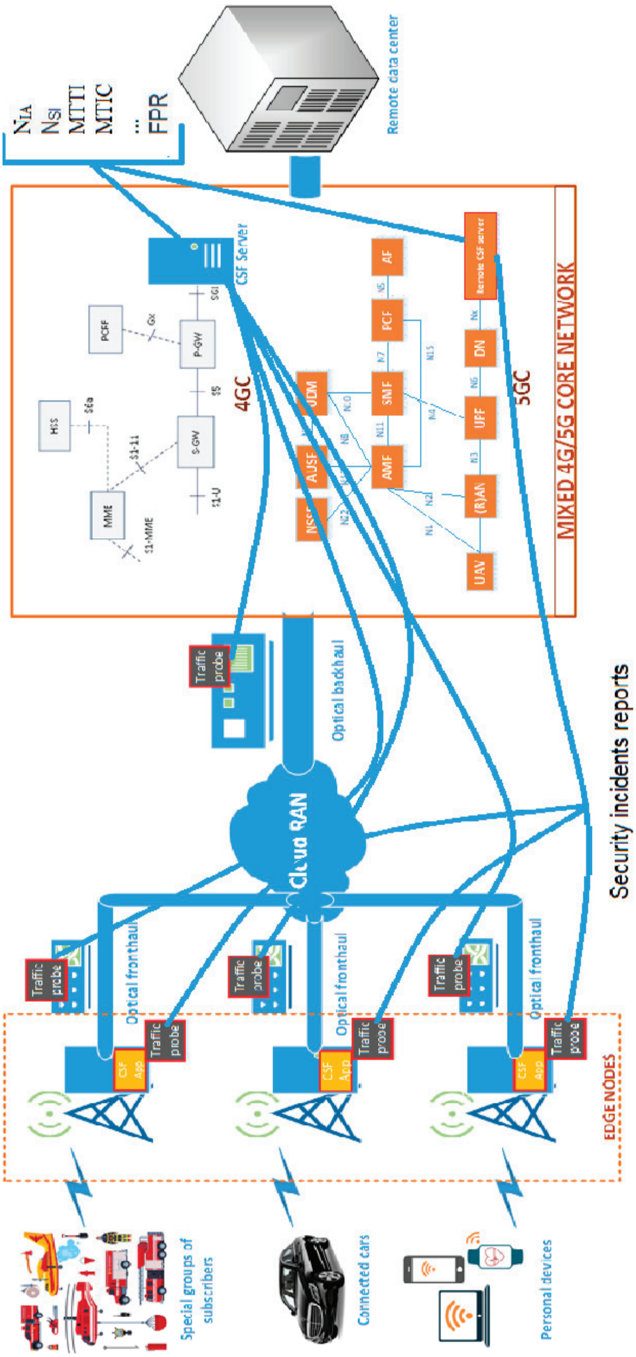


Figure 4. Continuous security KPIs monitoring system for 4G/5G/6G.

Thus, all the KPIs mentioned above will be collected in different parts of the network (different nodes) and stored in a specialized database that can be combined with the Cybersecurity Function Server (CSF) (Figure 4).

Furthermore, due to constant monitoring, the database will be filled in real-time with primary security KPIs, for which statistics on the number of incidents, their impact, scale, duration, etc., can be used. In the future, these primary indicators can be used to estimate secondary parameters using the mathematical apparatus in Table 2. The following pseudocode defines the algorithm developed for this assessment.

```
class Secure_KPI():
def __init__(self):
#defining the dictionary with the security KPIs as the keys and lists of desired parameters for the corresponding KPI for the concrete organization.
self.KPI={NIA:[parameters], NSI:[parameters], MTTI:[parameters], MTTR:[parameters], MTTD:[parameters], MTTRes:[parameters], NA:[parameters], AFR:[parameters], TPR:[parameters], FPR:[parameters], DLR:[parameters], TDT:[parameters], PVMT:[parameters]}
def input_data(self):
    # the array for storing the lists of parameters
self.data=[]
# appending the list with the needed number of empty lists
    for i in range(len(self.KPI)):
        self.KPI.append([])
    #filling the lists with the secure KPIs data for the concrete organization
    for kpi in self.KPI:
        number=0
        for i in self.KPI[kpi]:
            d=input("the desired data for your organization")
            self.data[number].append(d)
        number=number+1
#checking security KPIs with the defined formulas
def check(self):
    for kpi in self.KPI:
        for i in self.KPI[kpi]:
Calculate the corresponding security kpi according to the formulas in Table 2.
If security kpi > corresponding element in data list:
    alert(self.kpi)
#taking the security measures to mitigate the corresponding vulnerability, it will be defined in future works
def alert(self, problematic_kpi):
    taking the corresponding measures
#creating the object of the concrete organization
organization_x=Secure_KPI()
#inputting the data of the organization
organization_x.input_data()
#calculating and checking security KPI
organization.check()
```

The pseudocode offered is divided into 5 stages. The class is named Secure_KPI, designed to manage and assess key performance indicators (KPIs) related to security for a specific organization.

1. Initialization (Constructor):

The __init__ method is the constructor that initializes the class. Inside it, a dictionary called KPI is defined. This dictionary stores security KPIs as keys and lists the desired parameters for those KPIs as values.

2. Input Data:

   The input_data method is intended to gather data related to the security KPIs for the organization. Create an empty list called self.data and append it multiple times based on the number of KPIs in the KPI dictionary. Then, it iterates over each KPI, asking for user input to populate the lists in self.data with the desired data for each KPI.

3. Checking Security KPIs:

   The check method is used to assess the security KPIs. It iterates through the KPIs in the KPI dictionary and compares each KPI to the corresponding data from the self.data list; if a security KPI is greater than the corresponding element in the data list, it calls the alert method with the problematic KPI as an argument.

4. Alerting:

   The alert method is intended to take appropriate security measures to mitigate vulnerabilities when a problematic KPI is detected. However, implementing this method is incomplete, and it mentions taking measures not defined in the provided code.

5. Creating an Organization and Using the Class:

   At the end of the code, an instance of the Secure_KPI class is created, named organization_x.
   Data for the organization are input using the input_data method.
   The security KPIs are calculated and checked using the check method.

Additionally, the database should contain threshold values for the parameters of each layer (Table 3).

Based on the comparison of actual measured (estimated) KPIs with threshold values, a decision is made on the need to improve certain parameters (D), if necessary, based on Decision Rules (DR) matrices for each KPI.

$$D = \begin{cases} Rule_1 & if & cond_1 = true \\ \dots & \dots & \dots \\ Rule_N & if & cond_N = true \end{cases},$$

where

$$DR = \begin{pmatrix} Rule_1\ cond_1 \\ \dots \quad \dots \\ Rule_N cond_N \end{pmatrix}$$

where $Rule_N$ is the action that has to be applied if the condition $cond_N$ is true.

These formulas are introduced to complete the work of the approach in a comprehensive way. In the future, specific rules will be developed for certain conditions corresponding to deviations in the measured indicators.

6. Conclusions

In conclusion, 5th generation cellular networks actively replace communication in many areas of human life. The number of industries decreases, in which it is impossible or impractical to use 5G networks. Operators of critical infrastructure, special users (such as the police), governments, and the military are not the exception. Modern cellular networks can and must be easily adapted to the needs of special users. In this scenario, the network is subject to more stringent demands regarding reliability, performance, and, most importantly, data security. This scientific article focuses on the challenges related to ensuring cybersecurity.

To effectively increase the level of cybersecurity or ensure its sufficient level, it is necessary to measure the leading indicators of the effectiveness of security systems. At the moment, there are no comprehensive lists of these key indicators that require priority monitoring. Therefore, this article first analyzed the existing similar indicators and presented their list, which will make it possible to continuously monitor the state of cyber security systems of 4G/5G cellular networks with the aim of using them for groups of special users.

Therefore, this article proposed a method to determine these indicators and their evaluation. For this method, a meaningful analysis of possible groups of performance indicators was performed, the most relevant ones were selected, and a mathematical apparatus was proposed for their quantitative evaluation. Furthermore, within the framework of solving research problems, improvements were proposed for the core of the 4G/5G network, which allows data and performing statistical analysis at the expense of special sensors and the existing server.

Thus, to improve cybersecurity in critical infrastructure, government, military, and particular user networks using 5G technology, it is necessary to continuously monitor the performance of security systems. The first step is to ensure that the security architecture and practices comply with all the regulations governing the special user groups. After this, it is necessary to continuously monitor the presence of cyber incidents, log any violations, and perform more comprehensive assessments of the cybersecurity parameters in Table 3. If thresholds are exceeded, these assessments should become the basis for making decisions about an immediate response to cybersecurity problems or a comprehensive change to cybersecurity approaches.

Thus, the approach proposed in the article opens up an opportunity for continuous monitoring and, accordingly, improving the performance indicators of cybersecurity systems, which in turn makes it possible to use them for the maintenance of critical infrastructure and other users whose service requires increased requirements for cybersecurity systems.

Future scientific research will be directed toward implementing the proposed method and evaluating its validity. Additionally, there are plans to take advantage of artificial intelligence to process large datasets and make informed decisions based on established rules.

Author Contributions: Conceptualization, R.O., M.I., G.I., S.F. and Y.S.; methodology, R.O., M.I., G.I., S.F. and Y.S.; software, R.O., M.I., G.I., S.F. and Y.S.; validation, R.O., M.I., G.I., S.F. and Y.S.; formal analysis, R.O., M.I., G.I., S.F. and Y.S.; investigation, R.O., M.I., G.I., S.F. and Y.S.; resources, R.O., M.I., G.I., S.F. and Y.S.; data curation, R.O., M.I., G.I., S.F. and Y.S.; writing—original draft preparation, R.O., M.I., G.I., S.F. and Y.S.; writing—review and editing, R.O., S.F. and Y.S.; visualization, R.O., M.I., G.I., S.F. and Y.S.; project administration, R.O. and S.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by the Shota Rustaveli National Foundation of Georgia (SRNSFG) (NFR-22-14060), the National Scholarship Programme of the Slovak Republic and EU Next Generation EU through the Recovery and Resilience Plan for Slovakia under project No. 09I03-03-V01-000153.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pateria, J.; Ahuja, L.; Som, S.; Seth, A. Applying Clustering to Predict Attackers Trace in Deceptive Ecosystem by Harmonizing Multiple Decoys Interactions Logs. *Int. J. Inf. Technol. Comput. Sci.* **2023**, *15*, 35–44. [CrossRef]
2. Khaleefah, A.D.; Al-Mashhadi, H.M. Methodologies, Requirements and Challenges of Cybersecurity Frameworks: A Review. *Int. J. Wirel. Microw. Technol.* **2023**, *13*, 1–13. [CrossRef]
3. 5G Network Slice Management. Available online: https://www.3gpp.org/technologies/slice-management (accessed on 10 July 2023).
4. 5G-Trials—From 5G Experiments to Business Validation. Available online: https://5g-drive.eu/ (accessed on 9 September 2023).
5. 5G-MoNArch: 5G Mobile Network Architecture for Diverse Services, Use Cases, and Applications in 5G and Beyond. Available online: https://5g-ppp.eu/5g-monarch/ (accessed on 17 June 2022).
6. Juniper Networks Whitepaper. Managing 5G Slice Quality of Service End-to-End. Available online: https://www.juniper.net/content/dam/www/assets/flyers/us/en/managing-5g-slice-quality-of-service-end-to-end.pdf (accessed on 22 April 2021).
7. Hallé, C. Why Network Slicing Requires Active Monitoring, Passive Monitoring AND True APM. Available online: https://accedian.com/blog/why-network-slicing-requires-active-monitoring-passive-monitoring-and-true-apm/ (accessed on 16 November 2020).

8. Emblasoft. Innovate, Validate, Operate. Available online: https://emblasoft.com/ (accessed on 6 December 2022).
9. 5G Network Slicing Self-Management White Paper. Available online: https://www-file.huawei.com/-/media/corporate/pdf/news/5g-network-slicing-self-management-white-paper.pdf?la=en (accessed on 19 October 2020).
10. Wichary, T.; Mongay Batalla, J.; Mavromoustakis, C.X.; Żurek, J.; Mastorakis, G. Network Slicing Security Controls and Assurance for Verticals. *Electronics* **2022**, *11*, 222. [CrossRef]
11. Ogidiaka, E.; Ogwueleka, F.N.; Irhebhude, M.E. Game-Theoretic Resource Allocation Algorithms for Device-to-Device Communications in Fifth Generation Cellular Networks: A Review. *Int. J. Inf. Eng. Electron. Bus.* **2021**, *13*, 44–51. [CrossRef]
12. Mallipudi, C.C.; Chandra, S.; Prakash, P.; Arya, R.; Husain, A.; Qamar, S. Reinforcement Learning Based Efficient Power Control and Spectrum Utilization for D2D Communication in 5G Network. *Int. J. Comput. Netw. Inf. Secur.* **2023**, *15*, 13–24. [CrossRef]
13. Majeed, A.; Alnajim, A.M.; Waseem, A.; Khaliq, A.; Naveed, A.; Habib, S.; Islam, M.; Khan, S. Deep Learning-Based Symptomizing Cyber Threats Using Adaptive 5G Shared Slice Security Approaches. *Future Internet* **2023**, *15*, 193. [CrossRef]
14. Zahoor, S.; Ahmad, I.; Othman, M.; Mamoon, A.; Rehman, A.U.; Shafiq, M.; Hamam, H. Comprehensive Analysis of Network Slicing for the Developing Commercial Needs and Networking Challenges. *Sensors* **2022**, *22*, 6623. [CrossRef]
15. De Jesus Martins, R.; Wickboldt, J.A.; Granville, L.Z. Assisted Monitoring and Security Provisioning for 5G Microservices-Based Network Slices with SWEETEN. *J. Netw. Syst. Manag.* **2023**, *31*, 36. [CrossRef]
16. Kuklinski, S.; Tomaszewski, L.; Kolakowski, R.; Chemouil, P. 6G-LEGO: A framework for 6G network slices. *J. Commun. Netw.* **2021**, *23*, 442–453. [CrossRef]
17. Kukliński, S.; Tomaszewski, L. Key Performance Indicators for 5G network slicing. In Proceedings of the IEEE Conference on Network Softwarization (NetSoft), Paris, France, 24–28 June 2019; pp. 464–471. [CrossRef]
18. El Azzaoui, A.; Singh, S.K.; Pan, Y.; Park, J.H. Block5GIntell: Blockchain for AI-Enabled 5G Networks. *IEEE Access* **2020**, *8*, 145918–145935. [CrossRef]
19. Suomalainen, J.; Juhola, A.; Shahabuddin, S.; Mammela, A.; Ahmad, I. Machine Learning Threatens 5G Security. *IEEE Access* **2020**, *8*, 190822–190842. [CrossRef]
20. Zhang, S. An Overview of Network Slicing for 5G. *IEEE Wirel. Commun.* **2019**, *26*, 111–117. [CrossRef]
21. Koumaras, H.; Tsolkas, D.; Gardikis, G.; Gomez, P.M.; Frascolla, V.; Triantafyllopoulou, D.; Emmelmann, M.; Koumaras, V.; Osma, M.L.G.; Munaretto, D.; et al. 5GENESIS: The Genesis of a flexible 5G Facility. In Proceedings of the 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Barcelona, Spain, 17–19 September 2018; pp. 1–6. [CrossRef]
22. Doukoglou, T.; Gezerlis, V.; Trichias, K.; Kostopoulos, N.; Vrakas, N.; Bougioukos, M.; Legouable, R. Vertical Industries Requirements Analysis & Targeted KPIs for Advanced 5G Trials. In Proceedings of the 2019 European Conference on Networks and Communications (EuCNC), Valencia, Spain, 18–21 June 2019; pp. 95–100. [CrossRef]
23. Gupta, M.; Legouable, R.; Rosello, M.M.; Cecchi, M.; Alonso, J.R.; Lorenzo, M.; Kosmatos, E.; Boldi, M.R.; Carrozzo, G. The 5G EVE End-to-End 5G Facility for Extensive Trials. In Proceedings of the 2019 IEEE International Conference on Communications Workshops (ICC Workshops), Shanghai, China, 20–24 May 2019; pp. 1–5. [CrossRef]
24. Boero, L.; Bruschi, R.; Davoli, F.; Marchese, M.; Patrone, F. Satellite Networking Integration in the 5G Ecosystem: Research Trends and Open Challenges. *IEEE Netw.* **2018**, *32*, 9–15. [CrossRef]
25. Banović-Ćurguz, N.; Ilišević, D. Mapping of QoS/QoE in 5G Networks. In Proceedings of the 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 20–24 May 2019; pp. 404–408. [CrossRef]
26. Christopoulou, M.; Xilouris, G.; Sarlas, A.; Koumaras, H.; Kourtis, M.-A.; Anagnostopoulos, T. 5G Experimentation: The Experience of the Athens 5GENESIS Facility. In Proceedings of the 2021 IFIP/IEEE International Symposium on Integrated Network Management (IM), Bordeaux, France, 17–21 May 2021; pp. 783–787.
27. Saha, N.; James, A.; Shahriar, N.; Boutaba, R.; Saleh, A. Demonstrating Network Slice KPI Monitoring in a 5G Testbed. In Proceedings of the NOMS 2022–2022 IEEE/IFIP Network Operations and Management Symposium, Budapest, Hungary, 25–29 April 2022; pp. 1–3. [CrossRef]
28. Xie, M.; Gonzalez, A.J.; Gronsund, P.; Lonsethagen, H.; Waldemar, P.; Tranoris, C.; Denazis, S.; Elmokashfi, A. Practically Deploying Multiple Vertical Services into 5G Networks with Network Slicing. *IEEE Netw.* **2022**, *36*, 32–39. [CrossRef]
29. Lagen, S.; Bojovic, B.; Koutlia, K.; Zhang, X.; Wang, P.; Qu, Q. QoS Management for XR Traffic in 5G NR: A Multi-Layer System View & End-to-End Evaluation. *IEEE Commun. Mag.* **2023**, 1–7. [CrossRef]
30. Vordonis, D.; Giannopoulos, D.; Papaioannou, P.; Tranoris, C.; Denazis, S.; Rahav, R.; Altman, B.; Bosneag, A.-M.; Jain, S.; Margolin, U.; et al. Monitoring and Evaluation of 5G Key Performance Indicators in Media Vertical Applications. In Proceedings of the 2022 IEEE International Mediterranean Conference on Communications and Networking (MeditCom), Athens, Greece, 5–8 September 2022; pp. 203–208. [CrossRef]
31. Bolla, R.; Bruschi, R.; Davoli, F.; Lombardo, C.; Pajo, J.F.; Siccardi, B. Machine-Learning-Based 5G Network Function Scaling via Black- and White-Box KPIs. In Proceedings of the 21st Mediterranean Communication and Computer Networking Conference (MedComNet), Island of Ponza, Italy, 13–15 June 2023; pp. 143–150. [CrossRef]
32. Pinto, A.; Santaromita, G.; Fiandrino, C.; Giustiniano, D.; Esposito, F. Characterizing Location Management Function Performance in 5G Core Networks. In Proceedings of the IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), Phoenix, AZ, USA, 14–16 November 2022; pp. 66–71. [CrossRef]

33. Abdellatif, A.A.; Mohamed, A.; Erbad, A.; Guizani, M. Dynamic Network Slicing and Resource Allocation for 5G-and-Beyond Networks. In Proceedings of the 2022 IEEE Wireless Communications and Networking Conference (WCNC), Austin, TX, USA, 10–13 April 2022; pp. 262–267. [CrossRef]
34. Beaubrun, R. Technical Challenges and Categorization of 5G Mobile Services. In Proceedings of the 2022 Thirteenth International Conference on Ubiquitous and Future Networks (ICUFN), Barcelona, Spain, 5–8 July 2022; pp. 345–350. [CrossRef]
35. De Gaudenzi, R.; Luise, M.; Sanguinetti, L. The Open Challenge of Integrating Satellites into (Beyond-) 5G Cellular Networks. *IEEE Netw.* **2022**, *36*, 168–174. [CrossRef]
36. Fkih, F.; Al-Turaif, G. Threat Modelling and Detection Using Semantic Network for Improving Social Media Safety. *Int. J. Comput. Netw. Inf. Secur.* **2023**, *15*, 39–53. [CrossRef]
37. Shaikh, N.S.; Yasin, A.; Fatima, R. Ontologies as Building Blocks of Cloud Security. *Int. J. Inf. Technol. Comput. Sci.* **2022**, *14*, 52–61. [CrossRef]
38. Redefining Security KPIs for 5G Service Providers. Available online: https://www.helpnetsecurity.com/2019/11/19/5g-security-kpis/ (accessed on 19 November 2019).
39. Help Net Security. Average Data Breach Cost Has Risen to $3.92 Million. Available online: https://www.helpnetsecurity.com/2019/07/24/data-breach-cost/ (accessed on 24 July 2011).
40. Avkurova, Z.; Gnatyuk, S.; Abduraimova, B.; Fedushko, S.; Syerov, Y.; Trach, O. Models for early web-attacks detection and intruders identification based on fuzzy logic. *Procedia Comput. Sci.* **2022**, *198*, 694–699. [CrossRef]
41. Aurobindo, S. An introduction to intrusion detection. *Crossroads* **1996**, *2*, 3–7.
42. Kuypers, M.A.; Maillart, T.; Paté-Cornell, E. *An Empirical Analysis of Cyber Security Incidents at a Large Organization*; Department of Management Science and Engineering, Stanford University, School of Information: Stanford, CA, USA, 2016.
43. Doerrfeld, B. 5 Mean-Time Reliability Metrics to Follow. 7 July 2022. Available online: https://devops.com/5-mean-time-reliability-metrics-to-follow (accessed on 7 July 2023).
44. Hou, L.; Lao, Y.; Wang, Y.; Zhang, Z.; Zhang, Y.; Li, Z. Modeling freeway incident response time: A mechanism-based approach. *Transp. Res. Part C Emerg. Technol.* **2013**, *28*, 87–100. [CrossRef]
45. Oggerino, C. *High Availability Network Fundamentals*; Cisco Press: Indianapolis, IN, USA, 2001; 25p, ISBN 1-58713-017-3.
46. Azenkot, S.; Rector, K.; Ladner, R.; Wobbrock, J. PassChords: Secure multi-touch authentication for blind people. In Proceedings of the 14th international ACM SIGACCESS conference on Computers and Accessibility, Boulder, CO, USA, 22–24 October 2012; pp. 159–166.
47. Campos, L.M.; Ribeiro, L.; Karydis, I.; Karagiannis, S.; Pedro, D.; Martins, J.; Marques, C.; Armada, A.G.; Leal, R.P.; Lopez-Morales, M.J.; et al. Reference Scenarios and Key Performance Indicators for 5G Ultra-dense Networks. In Proceedings of the 12th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP), Porto, Portugal, 20–22 July 2020; pp. 1–5. [CrossRef]
48. Patel, A.; Qassim, Q.; Wills, C. A survey of intrusion detection and prevention systems. *Inf. Manag. Comput. Secur.* **2010**, *18*, 277–290. [CrossRef]
49. Alneyadi, S.; Sithirasenan, E.; Muthukkumarasamy, V. A survey on data leakage prevention systems. *J. Netw. Comput. Appl.* **2016**, *62*, 137–152. [CrossRef]
50. Lobato, A.G.P.; Lopez, M.A.; Sanz, I.J.; Cardenas, A.A.; Duarte, O.C.M.; Pujolle, G. An adaptive real-time architecture for zero-day threat detection. In Proceedings of the IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; pp. 1–6.
51. Kitchen, J.T.; Coogan, D.R.; Christian, K.H. The Evolution of Legal Risks Pertaining to Patch Management and Vulnerability Management. *Duq. L. Rev.* **2021**, *59*, 269.
52. Susanto, H.; Almunawar, M.N. *Information Security Management Systems: A Novel Framework and Software as a Tool for Compliance with Information Security Standard*; CRC Press: Boca Raton, FL, USA, 2018; 302p, ISBN 1771885777.
53. Perez, R.; Garcia-Reinoso, J.; Zabala, A.; Serrano, P.; Banchs, A. A monitoring framework for multi-site 5G platforms. In Proceedings of the IEEE European Conference on Networks and Communications (EuCNC), Dubrovnik, Croatia, 15–18 June 2020; pp. 52–56. [CrossRef]
54. Porambage, P.; Miche, Y.; Kalliola, A.; Liyanage, M.; Ylianttila, M. Secure Keying Scheme for Network Slicing in 5G Architecture. In Proceedings of the IEEE Conference on Standards for Communications and Networking (CSCN), Granada, Spain, 28–30 October 2019; pp. 1–6. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

big data and
cognitive computing

Article

Threat Hunting Architecture Using a Machine Learning Approach for Critical Infrastructures Protection

Mario Aragonés Lozano *, Israel Pérez Llopis and Manuel Esteve Domingo

Department of Communications, Universitat Politècnica de València, 46022 Valencia, Spain;
ispello0@upvnet.upv.es (I.P.L.); mesteve@dcom.upv.es (M.E.D.)
* Correspondence: maarlo9@teleco.upv.es

Abstract: The number and the diversity in nature of daily cyber-attacks have increased in the last few years, and trends show that both will grow exponentially in the near future. Critical Infrastructures (CI) operators are not excluded from these issues; therefore, CIs' Security Departments must have their own group of IT specialists to prevent and respond to cyber-attacks. To introduce more challenges in the existing cyber security landscape, many attacks are unknown until they spawn, even a long time after their initial actions, posing increasing difficulties on their detection and remediation. To be reactive against those cyber-attacks, usually defined as zero-day attacks, organizations must have Threat Hunters at their security departments that must be aware of unusual behaviors and Modus Operandi. Threat Hunters must face vast amounts of data (mainly benign and repetitive, and following predictable patterns) in short periods to detect any anomaly, with the associated cognitive overwhelming. The application of Artificial Intelligence, specifically Machine Learning (ML) techniques, can remarkably impact the real-time analysis of those data. Not only that, but providing the specialists with useful visualizations can significantly increase the Threat Hunters' understanding of the issues that they are facing. Both of these can help to discriminate between harmless data and malicious data, alleviating analysts from the above-mentioned overload and providing means to enhance their Cyber Situational Awareness (CSA). This work aims to design a system architecture that helps Threat Hunters, using a Machine Learning approach and applying state-of-the-art visualization techniques in order to protect Critical Infrastructures based on a distributed, scalable and online configurable framework of interconnected modular components.

Keywords: critical infrastructures protection; cyberattacks; machine learning; threat hunting; visualization models; architecture

Citation: Aragonés Lozano, M.; Pérez Llopis, I.; Esteve Domingo, M. Threat Hunting Architecture Using a Machine Learning Approach for Critical Infrastructures Protection. *Big Data Cogn. Comput.* **2023**, *7*, 65. https://doi.org/10.3390/bdcc7020065

Academic Editors: Peter R.J. Trim and Yang-Im Lee

Received: 8 February 2023
Revised: 10 March 2023
Accepted: 23 March 2023
Published: 30 March 2023

1. Introduction

In today's hyper-connected world, the dependency on the internet of production processes and activities is absolute, leaving useless any service offered, not only by big companies, agencies and SMEs (Small and Medium Enterprises), but also by critical infrastructures if internet access is lost, even for a few hours, thus leading to substantial economic losses and high severity cascading effects. This fact is well-known and exploited by cybercriminals who set cyber-attacks the order of the day.

To prevent cyber-attacks or, at least, to address them properly, critical infrastructures are investing big amounts of money in the improvement of their Information Technology (IT) security departments by making them bigger. The desired outcome is to avoid data loss, data exfiltration, maintain the reputation, and, probably the most important concern, minimize any impact in business continuity. Whether or not the previously stated desired outcomes are achieved by increasing in number the employee workforce, it is needed to continuously invest in highly skilled and specialized personnel who, without specific and useful tools, may end up overflowed by vast amounts of near real-time data and are unable to spot complex attacks, which are very quiet and remain in the protected infrastructure for a long time.

Nevertheless, a huge amount of the actionable data, both in the network and host, are related to harmless actions of the employees (such as DNS requests or WEB browsing). Moreover, surveys conducted with Threat Hunters [1] on the traits of those datasets concluded that there were specific and characterizable patterns for each of the studied actions, resulting in them being harmless or potentially dangerous. Being that Machine Learning is a scientific field characterized by providing outstanding techniques and procedures in extracting models from raw data [2], it follows that using well-designed, adequately tuned and scenario-customized ML algorithms can be helpful in classifying data samples according to how benign or malign they are.

Furthermore, according to several studies [3–5], human cognition tends to predict words, patterns, etc. strongly influenced by the context [6], even further if they seem to be under stress conditions [7]. In fact, those stressful conditions are suffered by Threat Hunters when they must face big amounts of data in highly dynamic scenarios where the smallest mistake can have a very high impact. Moreover, Threat Hunting is a complex decision-making process that encompasses many uncontrolled factors, typically working with limited and incomplete information and possibly facing unknown scenarios, for instance, zero-day attacks [8]. As a consequence, paying attention to the previously stated strong dependency on context in prediction by human cognition, an attack quite similar in behavior to a non-attack could be seen as such due to human bias; however, a Machine Learning system could discriminate between both more accurately than humans do. Thus, with all the data provided by the output of ML systems (such as likelihoods, feasibility thresholds, etc.), Threat Hunters could be able to understand better what is going on at the operations theater.

Moreover, it is well known that the human brain processes visual patterns more quickly and accurately than any textual or speech report, gaining understanding at a glimpse, and this, naturally, also happens in cybersecurity [9,10]; as a consequence, representing the data (both raw and ML processed data) properly is also a decisive factor for Threat Hunters in order to achieve Situational Awareness [11,12] and therefore an early detection of any threat. Some studies have been trying to classify which advanced visualization fits best for each kind of attack [13,14].

Lastly, using both Machine Learning and specifically defined data visualizations, Threat Hunters will be able to generate hypotheses about what is going on in their systems and networks, being able to quickly detect any threat and even have enough context information to deal with it.

Systems capable of gathering all those huge amounts of data, processing them (including Machine Learning techniques) and providing insightful visualization techniques must be developed following a properly designed architecture in accordance to the challenges that such an ambitious approach must face. The most relevant contribution of this work is an architecture proposal and its implementation devoted to fulfill the stated needs. The proposed architecture must provide means for dynamic and adaptable addition of ML techniques at will and the selection of which to use from the existing ones at a given moment. In addition, *big data* must be taken into account for vast amounts of data that must be stored and analyzed. Moreover, due to the time-consuming nature of ML processing, the architecture must enforce parallelization of as many processes as possible; therefore, architecture components must be orchestrated to maximize this parallelism. Furthermore, asymmetric scalability must be enforced in order to be efficient; thus, means should be instantiated to guarantee that only necessary components are working at a certain time. The architecture must be implemented in a distributed approach; therefore, communications, synchronization and decoupling of components and processing must be carefully envisioned and designed. Lastly, but not least, the whole system must be secured regarding the type of data it will process.

2. Motivation and Previous Work

The use of Machine Learning techniques in the field of Threat Hunting is booming: The research *An enhanced stacked LSTM method with no random initialization for malware threat hunting in safety and time-critical systems* [15] is focused on Time-Critical systems, paying attention to the conditions of those fast-paced situations, benefiting from the automation and effectiveness of malware detection that ML can provide. Both *Intelligent threat hunting in software-defined networking* [16] and *Advanced threat hunting over software-defined networks in smart cities* [17] are focused on developing intelligent Threat Hunting approaches on Software-Defined Networks (SDNs). In contrast, other efforts such as *A deep recurrent neural network based approach for internet of things malware threat hunting* [18] and *A survey on cross-architectural IoT malware threat hunting* [19] are more oriented toward the Internet of Things (IoT), a relevant area in the Threat Hunting community where the ML approaches provide benefits for the IoT specificities, for instance, resource scarceness as computational capabilities, among others. Finally, there also are works existing in the literature which try to solve the problem in a general perspective of ML applied to Threat Hunting, such as *Know abnormal, find evil: frequent pattern mining for ransomware threat hunting and intelligence* [20] and *Cyber threat hunting through automated hypothesis and multi-criteria decision making* [21].

Studies trying to develop a Threat Hunting architecture using an ML approach have already been conducted. First of all, the article *ETIP: An Enriched Threat Intelligence Platform for improving OSINT correlation, analysis, visualization and sharing capabilities* [22] can be found in the literature. In that work, an architecture which includes all steps, from data collection to data shown, is proposed; despite that, it is focused on generating IoCs (Indicators of Compromise) and it suggests using ML in some steps of the process. Another interesting work is *PURE: Generating Quality Threat Intelligence by Clustering and Correlating OSINT* [23]. This work, similar to the previous one, tries to develop an architecture to generate and enrich IoCs using ML at some steps. It gives another perspective on how to do it, despite the fact that it does not take into account the visualization of the results. It is interesting to highlight that neither of them define how to generate hypotheses using the generated data. Finally, the approach *SYNAPSE: A framework for the collection, classification, and aggregation of tweets for security purposes* [24] offers a wide and well-designed architecture, from data collectors to contents in visualization, although it is developed for a very specific data source (Twitter). Notwithstanding all the efforts already done, there are no specific studies about Threat Hunting using a Machine Learning approach for Critical Infrastructures in which an architecture is due to cope with all the stated needs that are proposed and neither the definition of useful nor specific visualizations are provided.

Regarding useful and specific visualizations for Cyber Situational Awareness, there is a very relevant work done in *Cyber Defense and Situational Awareness* [25] which states that "Visual analytics focuses on analytical reasoning using interactive visualizations". In order to support the previous statement, there is a comprehensive and complete survey on the cognitive foundations of visual analytics done in *Cognitive foundations for visual analytics* [26]. There is a wide variety of visualization techniques. Firstly, basic visualization charts, which include scatter plots [27–29], bar charts [30–32], pie charts [31] and line charts [32–34]. Another kind of simple visualization include word clouds [35,36] and decision trees [37,38]. On the other end of the spectrum, there are advanced visualizations. First are those oriented for pattern detection [39–43]. In addition, there are geo-referenced visualization charts for assets [41,43,44], risks [45–47] and threats [41,44]. Furthermore, there are also immersive visualization techniques using 3D models instead of 2D models which have been designed for optimum visualization with an ultra-wide high-definition screen, wrap-around screen or three-dimensional Virtual-Reality (VR) goggles, which allows the user to look around 360 degrees while moving [42,44,48–50].

All of them state the difficulties of the Threat Hunting process in terms of situation understanding in a broad threat-characterization landscape, with fast-changing conditions, sometimes unknown new threats, incomplete information and hidden features. Furthermore, several examples of enhancing the process by using ML techniques and useful visualizations can be found.

Besides academia, companies are also trying to develop specific Machine Learning techniques and algorithms for their Threat Hunting products to enrich current visualizations used to understand the cyber situational awareness of the monitored systems. Some offered products that implement ML algorithms are systems for Security Information and Event Management (SIEM), Firewalls, Antiviruses, Instrusion Detection System (IDS) and Intrusion Prevention System (IPS). A few examples are those like Splunk [51], Palo Alto next generation smart Firewalls [52], IBM immune system-based approach to cyber security (IBM X-Force Exchange [53,54]) or even Anomali ThreatStream [55].

After conducting deep research on the current state-of-the-art in the area, it can be concluded that, despite having made several outstanding efforts towards solving specific areas of the problem, there is no effort to define an architecture where implementation is rich enough to generate hypotheses about what is going on the system monitored. As a consequence, there is a lack (1) in the design of a particular unified architecture to help Threat Hunters with a Machine Learning approach with capabilities to define and generate (manually or automatically) hypotheses about what is going on and (2) in the provision of specific and useful visualizations, particularly in the issues detected for Critical Infrastructures (as might be the case of business continuity) and coping with all detected and envisioned scenarios. To fill this gap, an architecture with a specific component to define and generate hypotheses is proposed that must ensure security, scalability, modularity and upgradeability. It must also constitute a proper framework for developing platforms for Threat Hunting based on flexible and adaptable Machine Learning over the time. This work aims to solve this problem and fill the detected gap, mainly in terms of providing a unified framework that interrelates existing different components from data acquisition to knowledge generation (emphasizing the hypothesis generation) and visualization, which, despite being generic, is particularized for Critical Infrastructures Protection.

3. Outline of the System

In a brief and simplified view, a Threat Hunting tool can be seen as a closed-loop system. The system receives continuous and real-time feeds with, potentially, high-volume and diverse data inputs and, by means of some aiding subsystems (in this architecture machine-learning fuelled components), it provides and generates hypotheses on what is going on with confidence estimators or metrics. Those hypotheses and suggestions are provided to the end user, which closes the loop by providing feedback by selecting some selection branches more than others and even pruning complete branches, while seeking what is more likely to be going on with the given data.

The architecture proposed to help Threat Hunters by using a Machine Learning approach for Critical Infrastructures Protection is described in the following section. It is composed of five main layers interconnected in a stacked manner, as shown in Figure 1. The components within a layer can only communicate one with each other or to other components in adjacent layers. Moreover, components will provide standardized interfaces to communicate among themselves, and reusability will be enforced for their design and implementation.

It is important to state that bias can be introduced in the Threat Hunting process due to the well-known phenomena in interactive hypothesis-confirmation processes such as the valley effect for local versus global searches [56], among others, shown in areas as optimization or genetic algorithm evolutionary fitting [57].

Secondly, this architecture aims to be modular, efficient, and scalable. It is generic enough that it is able to be used in any kind of Critical Infrastructure but never loses focus on the main problems that must be tackled. By defining architecture-wide Application Programming Interfaces (APIs) that must be implemented at any component, creating new ones (components) is straightforward; the only requirement needed is to implement the corresponding interface and to provide mechanisms to notify the rest of the components about its availability. In addition, another relevant requirement is that each component must be completely stateless to allow decoupling and parallelization of processes. Moreover, with the components being stateless, the order of actions to do a simple process is not relevant, and therefore it can be a pool of available elements that dequeue pending tasks and, properly orchestrated, proceed to its completion, receiving all the required metadata (the state) itself.

The proposed architecture is flexible and scalable in terms of resources for its deployment. If resources are scarce, for instance, in debugging or testing or for an SME setup, every involved component can reside in a docker container [58] or in virtual machines [59], and the overall architecture can reside in a single machine. At the other end of the spectrum, where we can find setups with huge amounts of resources, the setup can be clustered using Kubernetes [60] or via cloud using AWS [61] or Azure [61]. From the components perspective, the type of deployment is transparent and seamless.

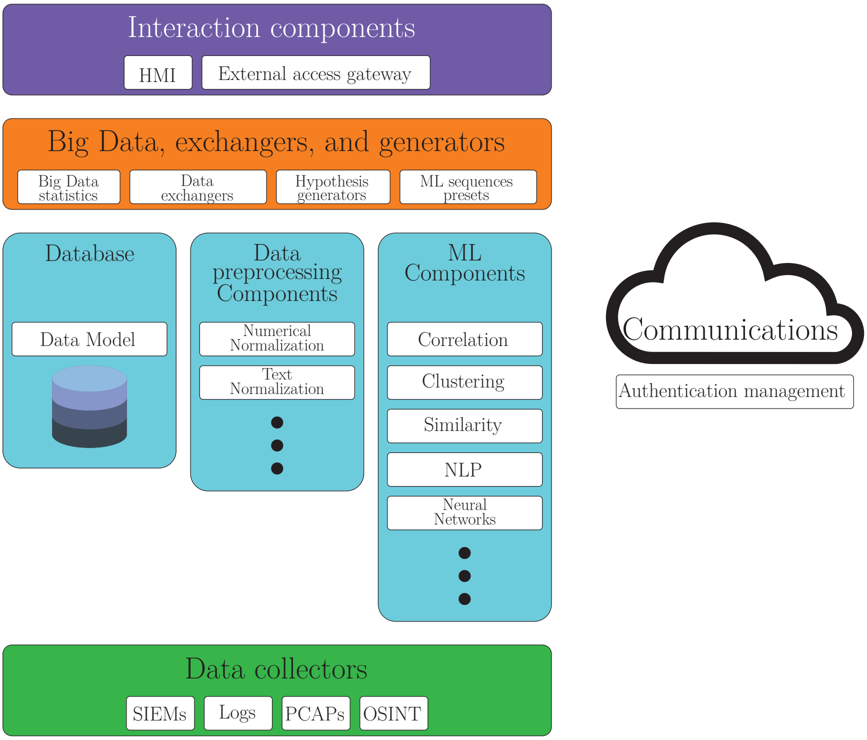


Figure 1. Proposed architecture. Groups of components from bottom to top and from left to right: Sections 4.1–4.8.

To achieve that goal, components must be completely decoupled, only knowing the existence of others on a per-needs basis on an orchestrated schema and communicating on standardized and predefined interfaces and mechanisms. That way, inner features of the component are completely isolated to the rest, and flexibility and decoupling can be reached.

This is one outstanding feature of the architecture that can provide flexibility and scalability for easily adapting to different and dynamically changing scenarios, depending on needs and resources. In addition, being able to provide flexibility also makes the architecture optimum for all kinds of Critical Infrastructures, deploying only the modules required for each specific one.

Another essential feature that must enforce the proposed architecture is the capability of providing High Availability (HA) [62] to guarantee service continuity (one of the main concerns of Critical Infrastructures) even in degraded conditions. To achieve that goal, load-balancing schemas are proposed within the component orchestrator, and, for the key elements (tagged as **crucial** through the following exposition) whose service must be guaranteed at all stakes for the rest to be able to work, backup instances should be ready in the background to replace the running ones if any issue is detected, therefore avoiding overall system service interruption.

Security is a crucial concern for any cyber security tool. Therefore, the architecture will establish security mechanisms to provide Agreed Security Service Levels in terms of security guarantees. Initially, these Security Service Levels Agreements (SSLAs) will be oriented to the capability of exchanging messages among components, and each component will ensure the authenticity [63] of the transmission; in short, the source's identity is confirmed and the requested action is allowed.

Another key part of the architecture is the interconnection within platforms implementing it or even with external sources. It does not matter how complex the developed architecture is; if the Section 4.6.2 is deployed, the implemented system will never lose the capability of being interconnected and sharing all kind of knowledge.

If several systems are deployed, creating a federation, the architecture will also provide the ability of sharing data regarding which items are the current active attacks, their input vectors, the IoC, etc. to warn other members of the federation if the system detects similar devices on the monitored network or even alert Threat Hunters which devices might be compromised. This feature is very important because a cyber-attack affecting a Critical Infrastructure can be propagated to another Critical Infrastructure [64].

In a brief summary, the proposed architecture aims to be distributed, self-adaptative, resilient and autopoietic [65], achieving that goal by being flexible, modular, and scalable but never losing the main objective of solving the detected problems in a fast and secure way.

The architecture will enforce the usage of standards at all levels to guarantee interoperability capabilities of the system, both in terms of data acquisition and, eventually, data export. Moreover, the usage of standards will provide sustainability of the life-cycle of developments, both at the hardware and the software faces, as well as flexibility and modularity in the selection and insertion of new elements and the replacement of existing ones. To do so, many different standards are proposed to be implemented and they will be specified in the corresponding sections. Among others, standard COTS (Commercial off-the-shelf) [66] mechanisms will be enforced at several layers of the architecture.

Several data sources will be implemented and feedback from Threat Hunters will be received in order to generate proactive security against threats. All this information, correctly processed, can be used to measure the security levels of the analyzed Critical Infrastructure.

4. System Architecture

The purpose of each layer is described hereunder from the bottom of Figure 1 to the top.

4.1. Layer 1: Data Collectors

The first layer contains the data collectors which are in charge of gathering data to feed the overall system. The collected data will be stored and it will be used by the other components within the system to process it. Both the raw and the processed data will be used to generate hypotheses about what is going on in the monitored infrastructure.

Any kind of data source is suitable to be implemented if it is interesting for Threat Hunters. Some examples of data sources could be:

- SIEMs, such as AlienVault [67] or IBM QRadar [68].
- Logs, such as Syslogs from the Operating System (OS), logs from network hardware devices, etc.

- PCAPs (Packet Captures, files with information about network traffic) [69].
- Threat Management Platforms (TMP), such as MISP [70].
- Incident Response Systems, such as The Hive [71] or RT-IR [72].
- Advanced Persistent Threat (APT) [73] management tools.
- OSINT (Open Source Intelligence [74]) sources, with their specific need in terms of normalization due to the wide variety of data typologies.

4.2. Layer 2: Database

The data gathered by the collectors will be stored in the database. In addition, every required metadata, which must be persistent over time, will also be stored in the database. Furthermore, the database must provide means for the rest of the components to access the stored data in an efficient and seamless way. Due to the previous statements, the database is a critical element and mandatory to be up and running for all the rest of the components to be working. Therefore, it is considered and shown as a **crucial** one.

Owing to the high-volume and diverse data stored into the database, this component must provide load-balancing mechanisms to guarantee proper access and pay strong attention to security as well as provide per-user policies per data access.

As a design requirement, all data stored must follow a specific data model that must be used within the overall components of the architecture. This data model must be flexible enough to be ready to adapt easily to changes and integrate new elements in the future. In addition, it must be oriented to store and process data related to events and cyber security. Being sort of the de facto standards, the data model must be compatible with *Sigma* and YARA rules.

Sigma rules (Generic Signature Format for SIEM Systems) [75,76] is an open and generic signature format that allows specialists to describe log events. In addition, with *Sigma*, cyber security tools (such as SIEMs) are able to exchange information among them, with the evident benefits that this interoperability can provide. One of the best features of using *Sigma* rules is its *Sigma Converter*, which allows Threat Hunters to convert the rules in elements such as Elastic Search Queries, Splunk Searches, as well as their ability to be reused and integrated into many other systems.

The malware analysis technique YARA [77,78] is used to discover malware based on its static character strings (the ones allocated inside the program itself) and signatures. It helps, among other things, to identify and classify malware, find new samples based on family-specific patterns, and identify compromised devices.

When designing the data model and the database structure, it is compulsory to consider several elements among which aspects stand out, such as writing/reading priorities, data storing and indexing. This is a critical element as it is the cornerstone for fast and efficient future complex data searches [79], something mandatory from a *big data* perspective as the one stood for the proposed architecture.

All this work and effort is needed because of the wide variety of data sources and the diversity of nature and typologies of data (especially those collected from OSINT sources) to be gathered by a system which implements this architecture. Each data source will, potentially, have a different taxonomy and also heterogeneous data that must be processed and adapted to define the data model before storing it into the database. It is evident that having a common taxonomy will provide some sort of *quantization noise* and it could lead to some information loss; nevertheless, a trade-off will be taken with regards to this aspect.

Adding new data sources is as easy as implementing the matching interface and casting the received data attributes to their closest mapping in the data model.

Proposed Database and Data Model

After conducting the study of the existing data model solutions, it is proposed the usage of the Elastic Common Schema (ECS) [80] because it suits the previously stated necessities due to its wide and general definition of fields related to cyber-data and its extended usage, maturity, wide community of users and third-party tools ecosystem.

In Table 1, the most interesting ECS fields can be found in order to be used with the proposed architecture. Nevertheless, the data model is not limited to those fields, but it can be enlarged if any component of the architecture needs it.

Coupling Elastic Search (ES) as a data repository with ECS is a widely recommended approach due to several reasons. First and mainly, both products come from the same source, thus guaranteeing a long-standing alignment as ECS is defined and in continuous development by Elastic. In addition, Elastic Search is *big data* enabled by nature [81] and follows HA because it can be clustered.

Table 1. Data model highlighted ECS fields.

ECS Field	Description
event.dataset	Name of the dataset
event.id	Unique ID to describe the event
event.ingested	Timestamp when an event arrived to the central data store
event.created	The date/time when the event was first read by an agent
event.starts	The date when the event started
event.end	The date when the event ended
event.action	The action captured by the event
event.original	Raw text message of entire event
source.ip	IP address of the source (IPv4 or IPv6)
source.mac	MAC address of the source
source.port	Port of the source
source.hostname	Hostname of the source.
destination.ip	IP address of the destination (IPv4 or IPv6)
destination.mac	MAC address of the destination
destination.port	Port of the destination
destination.hostname	Hostname of the destination

4.3. Layer 2: Data Preprocessing Components

Raw data, despite being defined in a specific well-designed data model, is not usually suitable for being used, but, when required, it must be preprocessed. Provided that system defined preprocessing techniques are finite and they are not specific for one final element, they can be shared among them.

Regarding the previously set statements, it is considered interesting to have a pool of preprocessing components to perform the required preprocessing techniques. When an ML system is being defined, the ML expert will have the possibility of introducing one step between selecting data from the database and one step between executing the desired ML technique where the selected data will be preprocessed according to the chosen preprocessing techniques. Furthermore, there must be the possibility of adding, upgrading or removing those components according to the necessities of the system.

Some examples of preprocessing components are as follows:

- **Sigma Converters**: Sigma Converters components allows to convert *Sigma* rules [75,76] to Elastic Search Queries, Splunk Searches or any other supported output.
- **Number Normalization**: Number normalization components are in charge of modifying a dataset of numbers by generating a new dataset with standard deviation 1 and mean 0, by multiplying all values by a specific factor, setting all minimum values to a specific threshold, etc.
- **Text Normalization**: Text normalization components are in charge of modifying texts by removing all forbidden characters, by adapting sentences to a predefined structure, etc.
- **One-Hot Encoders**: One-Hot Encoders components convert a categorical classification to a numerical classification by assigning a number to each one of the possible values [82].

4.4. Layer 2: ML Components

Machine Learning has several techniques, algorithms, etc., and they are evolving day by day. Instead of having one big element which contains all the ML knowledge, it is proposed to split it into several small components, each one responsible for doing one specific task. In addition, the components can be added, upgraded or deleted according to the requirements.

It is important to highlight that some ML techniques such as neural networks [83–86] must have external data such as pre-trained models, etc. Those external files are also taken into account, providing an external repository of data that is ML specific and which can be accessed by every ML component.

Some of the proposed ML components are as follows:

- **APT Clustering**: Cluster tactics and techniques with their associated APTs. Thanks to its hierarchical method to cluster and reduce data, the Birch algorithm is proposed [87,88].
- **Anomaly Detection**: This detect anomalies at logs and network behavior. Several ML techniques such as DBSCAN [89,90], Isolation Forest [91,92] or One Class Vector Machine [93,94] can be used.
- **NLP**: Natural Language Processing is mainly used for generating intelligence from analysts reports [95–97].
- **Decision trees**: Decision trees is an ML technique based on a process to classify data through a series of rules. The final result is obtained after deriving some specific characteristics from a pre-defined structure of rules [2,98].
- **Neural networks**: Several Neural Networks techniques can be used, such as Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), among others [2,99].

4.5. Layer 3: Big Data, Exchangers, and Generators

4.5.1. Big Data Statistics

The overall system is collecting and generating huge amounts of data per second, which makes the work of Threat Hunters difficult because they are not able to process all the data at the proper pace; as a consequence, data is tagged by Threat Hunters manually depending on the level of criticality. In order to help Threat Hunters in tagging those vast amounts of data, this paper proposes the automatization of this process by means of ML.

After this previous stage of data tagging, one step further must be taken in terms of providing means to Threat Hunters to help them in constructing or elaborating Cyber Situational Awareness. To do so, the usage of visualization techniques must be taken to provide valuable insights not easily seen by the human eye [100].

This final step is where *big data* statistics components make the difference, generating on-demand and real-time specific datasets on what is considered relevant for Threat Hunters.

Some examples could be:

- Which are the types of attacks that have greater occurrence?
- Which are the types of attacks that have greater impact?
- Which are the devices usually attacked?
- Which are the devices not usually attacked but were attacked recently?

4.5.2. Data Exchangers

To speed up incident handling performance, it is mandatory to have proper and standardized interoperability mechanisms. Basically, the system must have the ability to request data from external sources and to send data to foreign sinks. This specific ability will be defined in the proposed architecture using data exchangers.

As defined previously, firstly, this component enables the system to request data from external sources of information using standardized protocols. Several specific components, per data originator system and per protocol, will be available in the architecture to request, on a periodic basis or at a one shot schema, remote data with the required authentication.

This will be left open for customization by administrator users to set up the data to the approach that fits best on each data source.

Secondly, this component also allows the system to provide stored data, potentially filtered following given requests, to any authorized external requester using one of the standards that best fits its query.

Standard approaches such as JSON data format [101] or XML [101] will be used and are recommended due to their widespread nature. However, proprietary schemas and methods will be used when no other approaches are left open, as happens to be with several proprietary products and systems.

One step further, cyber security standards will also be used in the architecture for data exchanges. For instance, STIX (Structured Threat Information eXchange) [102] is going to be used as it is the de facto standard for cyber threat intelligence nowadays [103]. Moreover, widely used existing standards for cyber intelligence, such as CVE (Common Vulnerability enumeration) [104] or the SCAP (Security Content Automation Protocol) [105] suite, are going to be enforced and less extended usage ones would also be considered.

All the previously related standard mechanisms will be implemented in the architecture for both data gathering and delivery, and one of the goals of the proposed approach is to avoid proprietary data exchange mechanisms at all levels, if possible, and enforce standards usage. The usage of standards is mandatory for the scalability and extendability of the platform. One example that is considered is the capability of connecting the system on demand to external sources such as Virustotal [106], URLHaus [107], among others, which also do provide their own APIs to request/provide data, mostly based on well-known standards such as API REST to enrich the data processed by the platform. External data is beneficial for aspects such as IP/URLs/fqdn, hashes/files, etc., regarding detected IoCs with relevant intelligence from those well-known and reputed internet repositories.

Regarding the communication mechanisms, other standards such as API REST [108] for one-shot requests or AMQP [109] to publish/subscribe messaging are to be used to exchange data.

4.5.3. Hypothesis Generators

In order to help Threat Hunters discriminate which are the most current critical threats and their likeliness, and as contribution to the current state-of-the-art, we propose a specific component in charge of generating hypotheses.

Humans follow patterns in every action they do in their life, and even further when they interact with IT systems. Some of these patterns can cause cyber security events recognizable by pattern detection tools as a cyber security threat, for example, trying to gain access to some resource without enough rights, requesting Virtual Private Network (VPN) access out of business hours, etc. After conducting deep research with cyber security analysts, it was discovered that the detection of these specific harmless human patterns can be automated as they have common traits such as a specific user always coming from the same IP address. In order to automate the detection of harmless human patterns, a hypothesis generator component must be able to reduce the likelihood of a specific cyber threat being harmful, following some rules or even with specific ML algorithms. As a consequence, this component is considered relevant due to the benefits that it provides to cyber security analysts by freeing them from attending repetitive and harmless threats and allowing them to focus on those which are harmful.

In order to use this component, Threat Hunters must create rules which will be used to process the data. A rule consists of one or more filters executed in a specific order set by Threat Hunters. Each filter returns a numeric value that can be added, subtracted, multiplied or divided between steps to generate a likelihood of being benign or malign. The available hypothesis generators filters are classified as follows:

- **Simple filters**: Basic filtering rules (e.g., if/else rules).
- **Complex filters**: These rules find context by selecting more data related to the analyzed one (e.g., find how many times this pattern has been repeated).

- **ML filters**: These apply ML techniques from ML components to generate hypotheses.

In addition, each rule has a frequency value used by the Hypothesis Generator component to automatically request data to the database, process it and generate a hypothesis.

Regarding the previously set statements, the hypothesis generator component will be able to reduce or increase the likeliness of a detected threat being harmful according to the established configuration.

4.5.4. ML Sequences Presets

As said in previous sections, Machine Learning systems are composed of several components and steps that can be ordered depending on given needs: firstly, collecting the data; next, preparing it to fit the requirements of each specific ML technique; third is to process it using Machine Learning techniques; and finally, storing the results that must be persistent at a data storage.

Therefore, the user will be given the possibility to choose which Machine Learning components they want to use, and in which order. To do so, the definition and the orchestration are proposed to be done by a specific component named ML sequences presets, which will also hold the responsibility of triggering them.

In Section 4.6.1, there will be a specific interface to create, update and delete definitions of ML systems.

When a specific system is launched, this component will request the required components to start at the required moment as well as to keep track of the status of the execution.

4.6. Layer 4: Interaction Components

4.6.1. HMI

Threat Hunters and Machine Learning experts should be able to interact with the overall system using a simple, well-designed and easy-to-use graphical interface where all the required tools and visualizations will be accessible. In the proposed architecture, this specific task is implemented at the Human–Machine Interface (HMI).

The HMI must be modular enough to allow the configuration of all fields required by the different components that compose the overall system. Furthermore, the HMI will represent the data considered as relevant by Threat Hunters in the most efficient way.

As well as with the other components of the system, the access to the HMI will also be restricted by a user/password combination. The Role-Based policy [110], where each user has assigned a specific role which defines the allowed permissions, will be enforced for use in the HMI.

A web-based approach is proposed for the HMI as it is OS-agnostic without losing usability in desktop environments [111].

4.6.2. External Access Gateway

As specified in Section 4.5.2, the system must be accessible by third-party elements to gather data in a standardized way. For security reasons, it is interesting to have a specific element to act as proxy or API Gateway [112,113]; in the proposed architecture, that specific element is the External Access Gateway.

The main functions of this element are as follows. First, providing the endpoint for external requests. Second, checking the authentication of the request to decide whether it must be processed or not. Third, verifying the format of the request to ensure it is valid. Fourth, checking the authorization of the request to ensure that the requester has the required permissions to obtain that specific set of data. Fifth, forwarding the request to Section 4.5.2. Sixth, forwarding the response from Section 4.5.2 to the requester.

4.7. Common Layer: Communications

Being a distributed system introduces several complexities and challenges in the overall architecture design. For instance, it is necessary to have a communications broker in charge of exchanging and forwarding messages between each component and guaranteeing their proper delivery. As a consequence, the communications broker is a **crucial** component.

As stated before, all components of the system must send their messages using the communications broker and, in order to avoid the possibility of any unauthorized agent sending or receiving messages, the access to the communications broker network will be restricted and can be considered the first authentication factor, enforcing messages integrity [63].

In addition, messages will be exchanged using the AMQP [109] protocol and using several communications patterns: namely, one-to-one, one-to-many, in a broadcast manner, etc. Not only that, components will be sending messages using a request-response or subscription-publishing mechanism.

The usage of a communications broker provides many benefits to any distributed architecture. First of all, there are several extended-usage platforms that are widely tested by huge communities ensuring minimal communication issues. Moreover, the new elements addition process is relayed in the broker procedures and usually consists in connecting the broker following its mechanisms. Not only that, but networking issues are reduced because each component only needs to obtain access to the communications broker endpoint, so network administrators do not need to take care of broadcasting issues or other related problems. In addition, most brokers, if not all of them, provide real-time broadcast queues and subscription-publishing mechanisms which allow for immediate data updates. As a side effect, one-to-many message exchange patterns, such as those provided by communication brokers, do yield significant bandwidth consumption reduction.

4.8. Common Layer: Authentication Management

In order to manage the authentication of the different components and also the users that could interact with the system, and the different roles defined in the overall system by the administrators, there must be a specific component in place, referred to in the proposed architecture as authentication management. As the first step to be taken by each component or user is to log into the system to verify the permissions of the assigned role to the user, this component is **crucial**.

There are several options, being most outstanding OTP (One-Time Passwords) and OAuth 2.0. Despite some efforts being done in order to authorize using OTP [114,115], the proposed protocol is OAuth 2.0 due to the reasons detailed hereunder.

Nowadays, OAuth 2.0 has become the standard authorization protocol for the industry [116]. It enables a third-party application to obtain limited access to a specific service [117]. In addition, it can be configured to send not only the username and assigned role but also metadata when needed. Moreover, there are many implementations which allow systems administrators to choose which one of them fits best the requirements of the deployment, and it could be deployed locally or remotely, allowing the use of the implemented application either in isolated or shared networks. To summarize, many OAuth 2.0 implementations offer High Availability, which is a positive reinforcement of other architecture's requirements.

5. System Prototype

In order to validate the proposed system architecture a prototype, has been implemented. A brief view of the different components developed are shown in Figure 2, including each component in their corresponding layer in Figure 1, regarding the group of components.

The prototype has been evaluated using synthetic data simulating real networks and hosts by means of a digital twin. A digital twin can be defined as a clone of physical assets and their data in a virtualized environment simulating the cloned one. Digital twins also allow to test the physical one at all stages of the life cycle with the associated benefits of bugs and vulnerabilities detection [118].

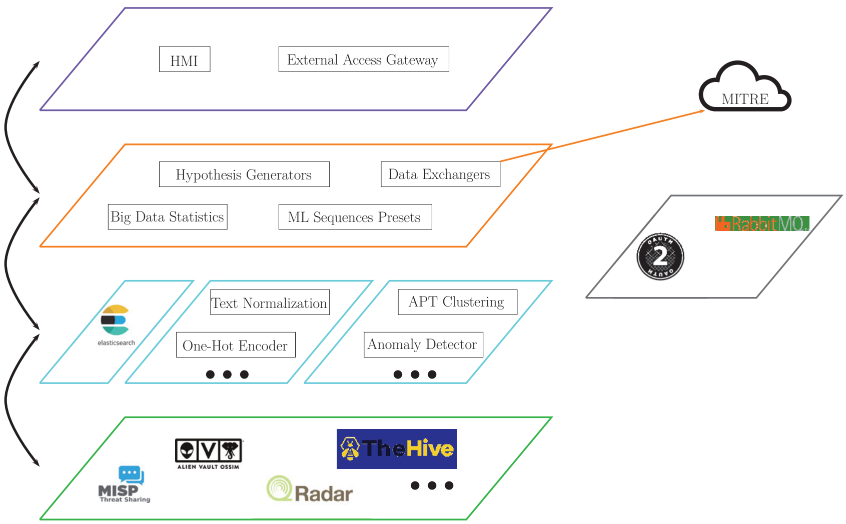


Figure 2. Prototype architecture.

In Figure 3 the implemented digital twin used to simulate a real Critical Infrastructure setup is detailed, including networks and assets (workstations, servers, network hardware, etc.) to verify the developed prototype that has been implemented using a virtualization platform. Three networks have been created. The first one contains all the monitored systems which will be attacked by an external actor in order to detect threats. The second one contains all the systems that the system prototype will collect data from. Lastly, the third network contains all the deployed components of the prototype.

5.1. Components

The components developed and deployed to verify the architecture will be described in this section. All of the developed components used Python [119–121] as the implementation language.

Following the same order as in previous sections, the data collectors were developed beforehand:

- MISP [70].
- OSSIM [67].
- QRadar [68].
- The Hive [71].
- PCAPs [69].
- Syslogs.
- Raw logs.

Regarding the database, Elastic Search was chosen along with Elastic Common Schema as the data model.

In addition, the data preprocessing components (Section 4.3) that were developed are the following:

- Sigma Converters.
- Number Normalization.
- Text Normalization.
- One-Hot Encoders.

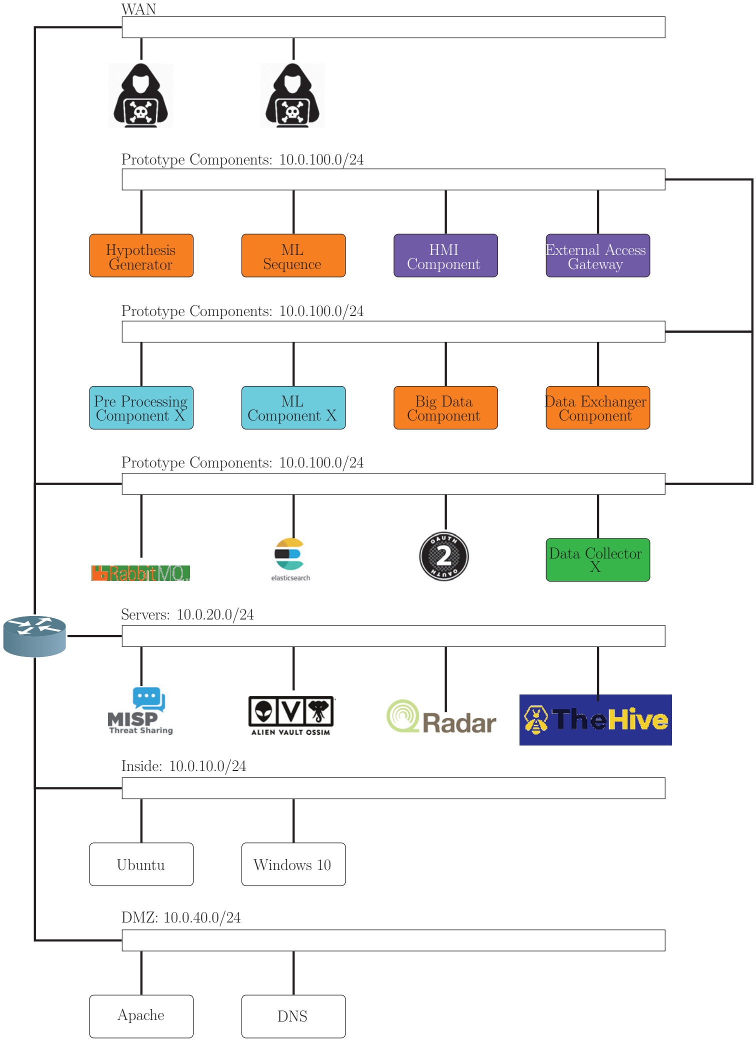


Figure 3. Digital twin.

Furthermore, the developed machine learning components (Section 4.4) used for verifying the architecture were the following:

- APT Clustering components.
- Anomaly detectors.
- NLP.
- Decision trees.
- Neural networks.

A model repository component was also used where pre-trained models were stored in order to feed the components which require them.

Big data statistics, the hypothesis generator, ML sequence presets and data exchangers components were also developed. It is considered interesting to highlight that data exchangers were able to query data from MITRE ATT&CK [122–124] as well as export data using STIX.

In order to interact with the system, an HMI and an External Access Gateway were also developed, acting as proxy to authenticate and authorize the requests before forwarding them to the available data exchangers.

Lastly, RabbitMQ [125–127] was used as a communications broker and a component which the OAuth 2.0 protocol implements was developed in order to manage the authentication.

5.2. Validation

The prototype has been validated layer by layer, following the same path that the data does, from the collection to the visualization.

The first step was to collect data from several sources. In order to do this, data collectors for MISP, OSSIM, QRadar and The Hive were deployed and properly configured, and, for each one of them, it was checked that the content was correctly collected and normalized following the proposed data model.

After that, the following step was to create Machine Learning systems using the ML Sequence Presets component. In the prototype, several ML Components along with Data Preprocessing Components were deployed in order to be used to generate sequences by concatenating all of those required in the order set by the ML expert. Those ML systems were executed either for one single shot or for recurrently generating valuable information about what is happening.

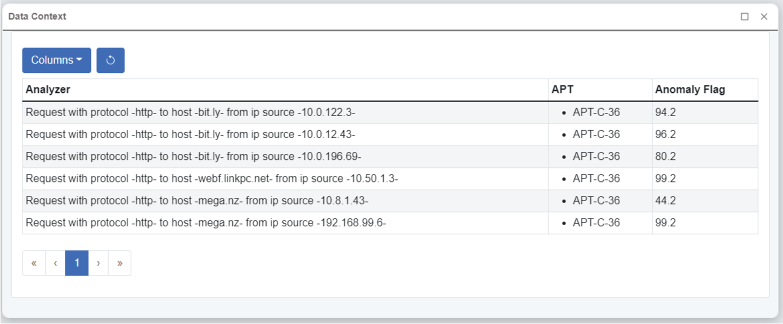
Having raw collected data and information generated by ML systems, the next step was to test the data exchangers in the two available ways: to export data to and import data from third parties. On one hand, using the External Access Gateway components, data was exported to an external system using STIX. On the other hand, data was imported from MITRE ATT&CK successfully.

As one key element of the proposed architecture, the Hypothesis Generator component was properly configured to process all the collected data and produce knowledge to generate valuable intelligence from those hypotheses previously checked and tuned by a Threat Hunter using the HMI.

The last step was to analyze and visualize all the gathered data, information and hypotheses to find threats in the monitored infrastructure. Some parts of the HMI regarding raw and chart data visualizations will be explained hereafter.

5.2.1. HMI: Raw Data Visualizations

The first highlighted generated data is used by Threat Hunters in order to conduct deep research about which actor is more likely to be targeting the monitored system. The information displayed relates actions detected by data collectors with some actors evaluating the relation with an anomaly flag. The data shown is generated using ML clustering and with data collected from external sources such as MITRE ATT&CK. The result is shown in Figure 4.

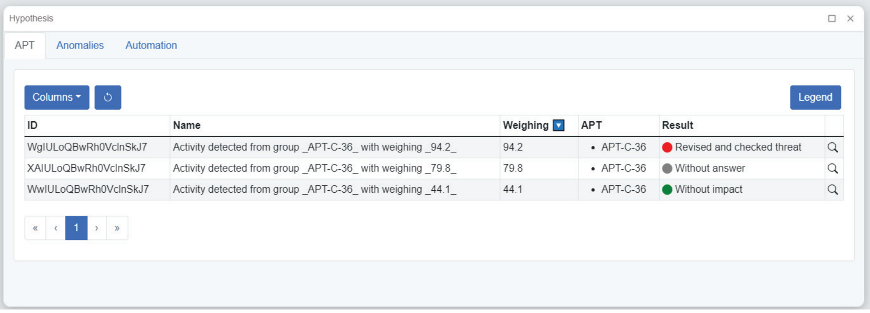


Analyzer	APT	Anomaly Flag
Request with protocol -http- to host -bit.ly- from ip source -10.0.122.3-	• APT-C-36	94.2
Request with protocol -http- to host -bit.ly- from ip source -10.0.12.43-	• APT-C-36	96.2
Request with protocol -http- to host -bit.ly- from ip source -10.0.196.69-	• APT-C-36	80.2
Request with protocol -http- to host -webf.linkpc.net- from ip source -10.50.1.3-	• APT-C-36	99.2
Request with protocol -http- to host -mega.nz- from ip source -10.8.1.43-	• APT-C-36	44.2
Request with protocol -http- to host -mega.nz- from ip source -192.168.99.6-	• APT-C-36	99.2

Figure 4. HMI: Data Context data.

A key of the proposed architecture is the ability of hypothesis generation, and, in order to do this, there is a specific component called Hypothesis Generator which is in charge of doing that specific task. The output of that component is listed at a specific visualization at the HMI which also enables to validate generated hypotheses.

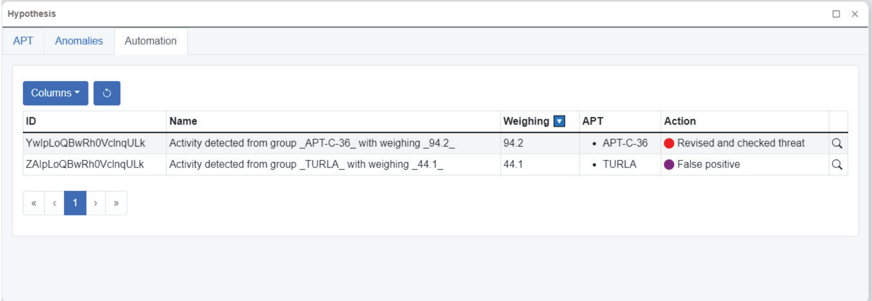
A hypothesis is a group of "Data Context" data which has been executed in a specific order and, optionally, can be associated to some APT. Once a hypothesis has been generated, it is shown to Threat Hunters with details containing the action chain to conduct a manual analysis in order to determine whether it is a threat or not. In Figure 5, there is an example of what would be seen by a Threat Hunter.



ID	Name	Weighing	APT	Result
WgIULoQBwRh0VclnSkJ7	Activity detected from group _APT-C-36_ with weighing _94.2_	94.2	• APT-C-36	● Revised and checked threat
XAIULoQBwRh0VclnSkJ7	Activity detected from group _APT-C-36_ with weighing _79.8_	79.8	• APT-C-36	● Without answer
WwIULoQBwRh0VclnSkJ7	Activity detected from group _APT-C-36_ with weighing _44.1_	44.1	• APT-C-36	● Without impact

Figure 5. HMI: Hypothesis: APT.

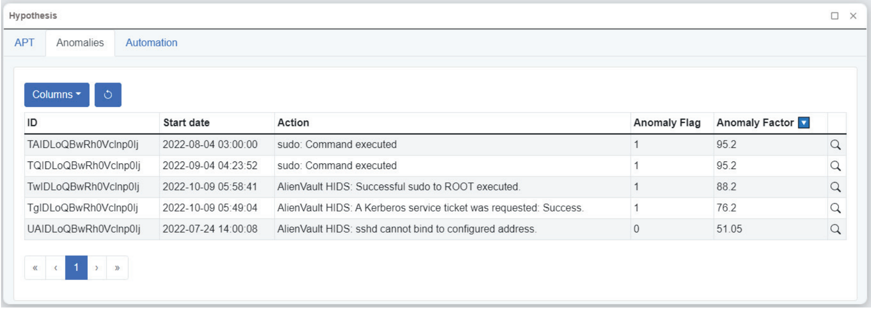
One outstanding feature of the proposed architecture is to provide ML capabilities to both Threat Hunting and hypothesis generation procedures. The Hypothesis Generators component is capable of continuously learning from Threat Hunters' hypothesis resolutions to distinguish between threats and benign behaviors, and, using the acquired intelligence, it is able to suggest to Threat Hunters the result of new hypotheses. The results proposed are shown in a view like the one in Figure 6.



ID	Name	Weighing	APT	Action
YwIpLoQBwRh0VclnqULk	Activity detected from group _APT-C-36_ with weighing _94.2_	94.2	• APT-C-36	● Revised and checked threat
ZAIpLoQBwRh0VclnqULk	Activity detected from group _TURLA_ with weighing _44.1_	44.1	• TURLA	● False positive

Figure 6. HMI: Hypothesis: Automation.

Another developed capability for the prototype is a hypothesis generator based on an anomaly detector, which creates results when some behavior deviates from the normal one of the system. It works by calculating an anomaly factor of the generated event and there is a configurable threshold which flags whether it is anomalous or not. One example can be shown in Figure 7.



ID	Start date	Action	Anomaly Flag	Anomaly Factor
TAIDLoQBwRh0Vclnp0lj	2022-08-04 03:00:00	sudo: Command executed	1	95.2
TQIDLoQBwRh0Vclnp0lj	2022-09-04 04:23:52	sudo: Command executed	1	95.2
TwIDLoQBwRh0Vclnp0lj	2022-10-09 05:58:41	AlienVault HIDS: Successful sudo to ROOT executed.	1	88.2
TgIDLoQBwRh0Vclnp0lj	2022-10-09 05:49:04	AlienVault HIDS: A Kerberos service ticket was requested: Success.	1	76.2
UAIDLoQBwRh0Vclnp0lj	2022-07-24 14:00:08	AlienVault HIDS: sshd cannot bind to configured address.	0	51.05

Figure 7. HMI: Hypothesis: Anomalies.

5.2.2. HMI: Chart Data Visualizations

As explained in [13,14], visual analysis can help Threat Hunters to solve difficult problems faster and ensure good results.

Regarding the importance of offering as many useful tools as possible for Threat Hunters, several configurable visualizations have been developed. It is considered important to highlight that color codes are enforced at any kind of visualization to obtain fast recognition about what is being visualized. Visualized data can also be filtered by Threat Hunters if they need it. In addition, all visualizations are interactive, offering zoom in, zoom out and pan capabilities to examine in detail those complex aspects.

Hereunder are some examples of implemented visualizations (Figures 8–11) in which all of them show the given assets with their existing services per asset and the vulnerabilities detected for that specific service but displayed using different visualization techniques.

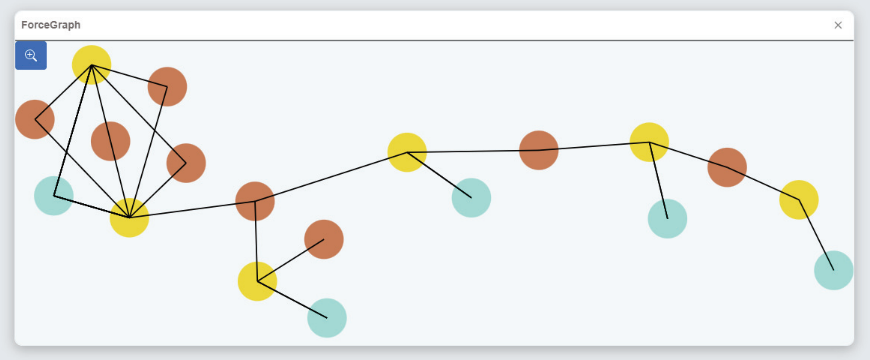


Figure 8. HMI: Chart Force Graph.

In the previous figure, we can find a graph showing the assets (brown color) connected to the services (yellow color) they have and the vulnerabilities (sky blue color) associated to them.

The same query to the data storage is shown in Figure 9 (i.e., assets per services per vulnerabilities) but with a different visualization technique, in this case, circle packing. The packing visualizations do lose the graph interconnection-display capability but provide means to see which element encircles another. Therefore, we can see here inside an asset (brown), its services (yellow circle), and inside each service its vulnerabilities (sky blue disc).

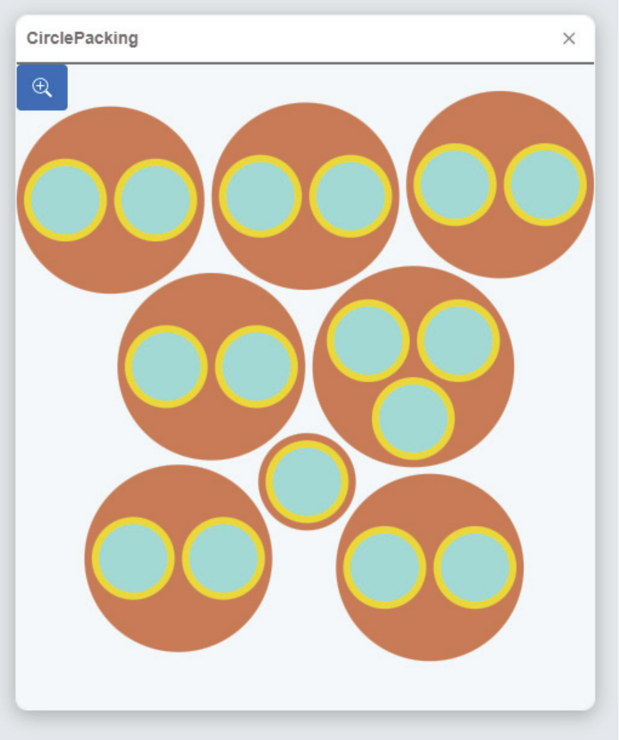


Figure 9. HMI: Chart Circle Packing.

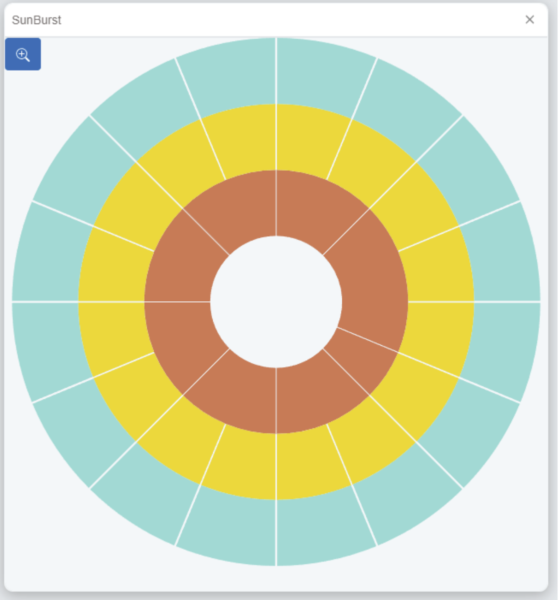


Figure 10. HMI: Chart Sun Burst.

In the above snapshot, the same query is shown (assets per services per vulnerabilities) with the same color schema (assets displayed with brown color, services with yellow color, and vulnerabilities with sky blue color) but, in this case, elements are not encircled but laid on a concentric set of discs, each one representing a layer.

It is remarkable to state that, in all the views, the user can interact at any time with what is currently displayed; if the users clicks on any figure, a new window with all the detailed information about the element is shown.

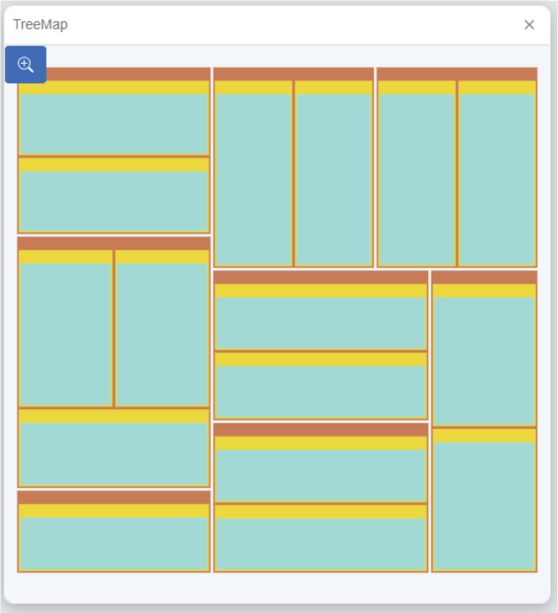


Figure 11. HMI: Chart Tree Map.

The tree map view is quite similar to the circle packing, but in this case it is representing a Hilbert space decomposition. Again, assets, their services and their associated vulnerabilities are shown with the same color code and grouped in the shown boxes. It is important to state that the user can interact with the visualization as they can do in all the other visualizations.

Implemented visualizations are not limited to these examples but they are composed of an extended range of techniques, all of them enforcing the capability of helping in detecting patterns in complex and multi-dimensional datasets. As relevant features, we can point out that they are graph-based and provide means to show multi-dimensional interrelated data in a few dimensions' graph.

5.3. Verification

After the validation process was successfully completed, a verification of the prototype was conducted with Threat Hunters (i) to ensure that the defined architecture copes with all the envisioned scenarios outlined in Section 2 and (ii) to validate the performance of the prototype against other solutions in the existing state-of-the-art.

Because there are no two identical people, it is difficult to ensure that a system is good enough for everyone, but with enough population, there can be a subjective approximation if it is fairly good or not. The subjective verification process was split into three stages: (i) Firstly, the implemented prototype was deployed in the networks monitored by the Threat Hunters in charge of evaluating it. (ii) After several months (time enough to have sufficient data in the prototype to obtain valid results through the ML components), the prototype was used by Threat Hunters in parallel with their own systems. (iii) Lastly, Threat Hunters were asked to answer specific surveys (some of whose questions are shown in Table 2) to determine how valid the system is.

Table 2. Sample of verification survey questions.

Question
Does the prototype give fast access to the information considered as relevant?
Does the prototype receive updated information from external sources?
Does the prototype send information to external sources?
Does the prototype provide tools to easily create/edit/delete preprocessing components?
Does the prototype provide tools to easily create/edit/delete ML components?
Does the prototype help at the decision making process?
Is the prototype easy to use?

The survey answers showed that, generally, the prototype was useful and the proposed architecture is strong enough to be used as a Threat Hunting tool for Critical Infrastructures.

Aside from the subjective evaluation of the prototype, some calculated metrics of the hypothesis generator component were also calculated, whose results are presented in Table 3.

Table 3. Metrics of the hypothesis generator component.

Metric	After 1 Month	After 6 Months
Percentage of benign events marked correctly by the prototype	31.56%	83.49%
Percentage of malign events marked correctly by the platform	23.16%	73.08%
Ratio of likeliness of the hypothesis	24.62%	89.24%
Percentage of attacks detected by the platform	26.74%	86.31%

6. Conclusions

In the previous sections, the architecture and all its features have been presented, followed by an exhaustive overall validation and verification. The results obtained can be used to compare given features to others from the tools and systems in the existing state-of-the-art. This comparison has drawn the following conclusions.

Firstly, it has been pointed out that there is a need to improve the tools used by Threat Hunters in Critical Infrastructures to improve their daily job. Among all the difficulties that Threat Hunters must face, a critical one is the vast amount of data that they must process with the consequent degradation in the process of situation understanding, decision making and the associated cognitive overwhelm.

This work, alongside others existing in the state-of-the-art, aims to solve that problem by proposing an architecture in order to help Threat Hunters by coping with the stated problem by means of a reduction of information presented to them using a Machine Learning approach that provides suggestions and hints about what is going on.

The current systems and tools stated in the state-of-the-art are mainly focused on the generation of IoCs, but none of them take into account tools to help Threat Hunters in the hypothesis generation process. As a consequence, there is gap in the generation of hypotheses using raw and/or ML processed data to know what is going on in the system monitored, which the proposed architecture tries to fill by enforcing hypothesis generation as a main aid to Threat Hunters. Consequently, one of the main contributions of the work described (and not fully found in similar solutions) is the provided capability to Threat Hunters to be helped by ML processes in generating complex and elaborated hypotheses about the current situation and what is more likely to happen in the near future. Furthermore, a key aspect of this kind of system, namely, visualization, is not fully exploited through the tools surveyed in the state-of-the-art, whereas in the proposed architecture, this element is enforced to help Threat Hunters in elaborating a proper understanding of the situation and the most likely evolution of events.

The proposed architecture takes into account several aspects. First of all, it is modular and upgradeable, as elements can be added or removed on demand dynamically, which gives it the capability of being ready for any kind of critical infrastructure. This is considered important from our point of view due to the fact that there are no two systems that are identical and this is not enforced in other papers and projects from the state-of-the-art. Secondly, it is asymmetrically scalable, so each resource assignment is orchestrated depending on the needs. Furthermore, it is *big data*-enabled, which means it can store and analyze vast amounts of data, and all the stored data is not only used for generating hypotheses, but Threat Hunters can also use it for conducting a deep study of potential malicious data or even for measuring the security levels of the Critical Infrastructure that is being monitored.

It is also able to exchange (request and response) data with external sources using standardized formats. This specific capability enables it to warn other Critical Infrastructures when there are common dependencies and when an attack with a similar entry vector is detected. In addition, as each component is stateless, the order of actions to perform a simple process is not relevant; therefore, processes can be parallelized to increase the performance of the overall system. Unlike the papers and projects in the current state-of-the-art, the proposed architecture follows High Availability enforcement schemas at all the essential components (database, communications broker and authentication management) to be confident about the uptime of the deployed system, which is crucial to be used in critical situations. Furthermore, this type of system is used in IT security departments to prevent and respond to cyber-attacks. Consequently, the data processed by the system are very sensitive, so being secure is a significant concern. To address this, the architecture allows several authentication methods to work safely with the data.

Lastly, the proposed architecture has been validated and verified implementing a prototype that was tested by Threat Hunters by answering specific surveys (Table 2) and by analyzing metrics of the hypothesis generator component (Table 3).

Author Contributions: Writing—original draft, M.A.L., I.P.L. and M.E.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the European Commission's Project PRAETORIAN (Protection of Critical Infrastructures from advanced combined cyber and physical threats) under the Horizon 2020 Framework (Grant Agreement No. 101021274).

Data Availability Statement: The data analyzed in this study was synthetically generated. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
APT	Advanced Persistent Threat
CI	Critical Infrastructures
CSA	Cyber Situational Awareness
ECS	Elastic Common Schema
ES	Elastic Search
HA	High Availability
HMI	Human-Machine Interface
IDS	Instrusion Detection System
IoC	Indicator of Compromise
IoT	Internet of Things
IP	Internet Protocol
IPS	Intrusion Prevention System

IT	Information Technology
ML	Machine Learning
OS	Operating System
OSINT	Open Source Intelligence
OTP	One Time Passwords
SDN	Software-Defined Networks
SIEM	Security Information and Event Management
SME	Small and Medium Enterprise
SSLA	Security Service Levels Agreements
TMP	Threat Management Platforms
VPN	Virtual Private Network
VR	Virtual-Reality

References

1. PRAETORIAN. D3.1 Transitioning Risk Management, 2021. *PRAETORIAN H2020 Project Deliverables*. Not yet published.
2. Li, J.H. Cyber security meets artificial intelligence: A survey. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 1462–1474. [CrossRef]
3. Falandays, J.B.; Nguyen, B.; Spivey, M.J. Is prediction nothing more than multi-scale pattern completion of the future? *Brain Res.* **2021**, *1768*, 147578. [CrossRef]
4. Federmeier, K.D. Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology* **2007**, *44*, 491–505. [CrossRef] [PubMed]
5. Riegler, A. The role of anticipation in cognition. In Proceedings of the AIP Conference Proceedings. *Am. Inst. Phys.* **2001**, *573*, 534–541.
6. Slattery, T.J.; Yates, M. Word skipping: Effects of word length, predictability, spelling and reading skill. *Q. J. Exp. Psychol.* **2018**, *71*, 250–259. [CrossRef] [PubMed]
7. Lehner, P.; Seyed-Solorforough, M.M.; O'Connor, M.F.; Sak, S.; Mullin, T. Cognitive biases and time stress in team decision making. *IEEE Trans. Syst. Man -Cybern.-Part Syst. Humans* **1997**, *27*, 698–703. [CrossRef]
8. Bilge, L.; Dumitraş, T. Before we knew it: An empirical study of zero-day attacks in the real world. In Proceedings of the 2012 ACM Conference on Computer and Communications Security, Raleigh North, CA, USA, 16–18 October 2012; pp. 833–844.
9. Markowsky, G.; Markowsky, L. Visualizing cybersecurity events. In Proceedings of the International Conference on Security and Management (SAM), Las Vegas, NV, USA, 22–25 July 2013; p. 1.
10. Young, C.S. Representing Cybersecurity Risk. In *Cybercomplexity*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 19–24.
11. Endsley, M.R. Measurement of situation awareness in dynamic systems. *Hum. Factors* **1995**, *37*, 65–84. [CrossRef]
12. Franke, U.; Brynielsson, J. Cyber situational awareness–a systematic review of the literature. *Comput. Secur.* **2014**, *46*, 18–31. [CrossRef]
13. Chen, S.; Guo, C.; Yuan, X.; Merkle, F.; Schaefer, H.; Ertl, T. Oceans: Online collaborative explorative analysis on network security. In Proceedings of Eleventh Workshop on Visualization for Cyber Security, Paris, France, 10 November 2014; pp. 1–8.
14. Choi, H.; Lee, H. PCAV: Internet attack visualization on parallel coordinates. In Proceedings of the International Conference on Information and Communications Security, Beijing, China, 10–13 December 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 454–466.
15. Jahromi, A.N.; Hashemi, S.; Dehghantanha, A.; Parizi, R.M.; Choo, K.K.R. An enhanced stacked LSTM method with no random initialization for malware threat hunting in safety and time-critical systems. *IEEE Trans. Emerg. Top. Comput. Intell.* **2020**, *4*, 630–640. [CrossRef]
16. Schmitt, S.; Kandah, F.I.; Brownell, D. Intelligent threat hunting in software-defined networking. In Proceedings of the 2019 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 11–13 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
17. Schmitt, S. *Advanced Threat Hunting over Software-Defined Networks in Smart Cities*; University of Tennessee at Chattanooga: Chattanooga, Tennessee, USA, 2018.
18. HaddadPajouh, H.; Dehghantanha, A.; Khayami, R.; Choo, K.K.R. A deep recurrent neural network based approach for internet of things malware threat hunting. *Future Gener. Comput. Syst.* **2018**, *85*, 88–96. [CrossRef]
19. Raju, A.D.; Abualhaol, I.Y.; Giagone, R.S.; Zhou, Y.; Huang, S. A survey on cross-architectural IoT malware threat hunting. *IEEE Access* **2021**, *9*, 91686–91709. [CrossRef]
20. Homayoun, S.; Dehghantanha, A.; Ahmadzadeh, M.; Hashemi, S.; Khayami, R. Know abnormal, find evil: Frequent pattern mining for ransomware threat hunting and intelligence. *IEEE Trans. Emerg. Top. Comput.* **2017**, *8*, 341–351. [CrossRef]
21. Neto, A.J.H.; dos Santos, A.F.P. Cyber threat hunting through automated hypothesis and multi-criteria decision making. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1823–1830.
22. Gonzalez-Granadillo, G.; Faiella, M.; Medeiros, I.; Azevedo, R.; Gonzalez-Zarzosa, S. ETIP: An Enriched Threat Intelligence Platform for improving OSINT correlation, analysis, visualization and sharing capabilities. *J. Inf. Secur. Appl.* **2021**, *58*, 102715. [CrossRef]

23. Azevedo, R.; Medeiros, I.; Bessani, A. PURE: Generating quality threat intelligence by clustering and correlating OSINT. In Proceedings of the 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications (TrustCom), Rotorua, New Zealand, 5–8 August 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 483–490.
24. Alves, F.; Ferreira, P.M.; Bessani, A. OSINT-based Data-driven Cybersecurity Discovery. In Proceedings of the 12th Eurosys Doctoral Conference, Porto, Portugal, 23 April 2018; pp. 1–5.
25. Kott, A.; Wang, C.; Erbacher, R.F. *Cyber Defense and Situational Awareness*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 62.
26. Greitzer, F.L.; Noonan, C.F.; Franklin, L. *Cognitive Foundations for Visual Analytics*; Technical Report; Pacific Northwest National Lab.(PNNL): Richland, WA, USA, 2011.
27. Eslami, M.; Zheng, G.; Eramian, H.; Levchuk, G. Deriving cyber use cases from graph projections of cyber data represented as bipartite graphs. In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 4658–4663.
28. Kotenko, I.; Novikova, E. Visualization of security metrics for cyber situation awareness. In Proceedings of the 2014 Ninth International Conference on Availability, Reliability and Security, Fribourg, Switzerland, 8–12 September 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 506–513.
29. Beaver, J.M.; Steed, C.A.; Patton, R.M.; Cui, X.; Schultz, M. Visualization techniques for computer network defense. In Proceedings of the Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense X. SPIE, Orlando, FL, USA, 25–28 April 2011; Volume 8019, pp. 18–26.
30. Goodall, J.R.; Ragan, E.D.; Steed, C.A.; Reed, J.W.; Richardson, G.D.; Huffer, K.M.; Bridges, R.A.; Laska, J.A. Situ: Identifying and explaining suspicious behavior in networks. *IEEE Trans. Vis. Comput. Graph.* **2018**, *25*, 204–214. [CrossRef] [PubMed]
31. Zhuo, Y.; Zhang, Q.; Gong, Z. Cyberspace situation representation based on niche theory. In Proceedings of the 2008 International Conference on Information and Automation, Zhangjiajie, China, 20–23 June 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1400–1405.
32. Pike, W.A.; Scherrer, C.; Zabriskie, S. Putting security in context: Visual correlation of network activity with real-world information. In *VizSEC 2007*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 203–220.
33. Abraham, S.; Nair, S. Comparative analysis and patch optimization using the cyber security analytics framework. *J. Def. Model. Simul.* **2018**, *15*, 161–180. [CrossRef]
34. Graf, R.; Gordea, S.; Ryan, H.M.; Houzanme, T. An Expert System for Facilitating an Institutional Risk Profile Definition for Cyber Situational Awareness. In Proceedings of the ICISSP, Rome, Italy, 19–21 February 2016; pp. 347–354.
35. Lohmann, S.; Heimerl, F.; Bopp, F.; Burch, M.; Ertl, T. Concentri cloud: Word cloud visualization for multiple text documents. In Proceedings of the 2015 19th International Conference on Information Visualisation, Barcelona, Spain, 22–24 July 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 114–120.
36. Xu, J.; Tao, Y.; Lin, H. Semantic word cloud generation based on word embeddings. In Proceedings of the 2016 IEEE Pacific Visualization Symposium (PacificVis), Taipei, Taiwan, 19–22 April 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 239–243.
37. De Ville, B. Decision trees. *Wiley Interdiscip. Rev. Comput. Stat.* **2013**, *5*, 448–455.
38. Tak, S.; Cockburn, A. Enhanced spatial stability with hilbert and moore treemaps. *IEEE Trans. Vis. Comput. Graph.* **2012**, *19*, 141–148. [CrossRef]
39. Angelini, M.; Bonomi, S.; Lenti, S.; Santucci, G.; Taggi, S. MAD: A visual analytics solution for Multi-step cyber Attacks Detection. *J. Comput. Lang.* **2019**, *52*, 10–24.
40. Zhong, C.; Alnusair, A.; Sayger, B.; Troxell, A.; Yao, J. AOH-map: A mind mapping system for supporting collaborative cyber security analysis. In Proceedings of the 2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), Las Vegas, NV, USA, 8–11 April 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 74–80.
41. Cho, S.; Han, I.; Jeong, H.; Kim, J.; Koo, S.; Oh, H.; Park, M. Cyber kill chain based threat taxonomy and its application on cyber common operational picture. In Proceedings of the 2018 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), Glasgow, Scotland, UK, 11–12 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–8.
42. Kabil, A.; Duval, T.; Cuppens, N.; Comte, G.L.; Halgand, Y.; Ponchel, C. From cyber security activities to collaborative virtual environments practices through the 3D cybercop platform. In Proceedings of the International Conference on Information Systems Security, Funchal, Madeira, Portugal, 22–24 January 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 272–287.
43. Kopylec, J.; D'Amico, A.; Goodall, J. Visualizing cascading failures in critical cyber infrastructures. In Proceedings of the International Conference on Critical Infrastructure Protection, Hanover, NH, USA, 18–21 March 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 351–364.
44. Llopis, S.; Hingant, J.; Pérez, I.; Esteve, M.; Carvajal, F.; Mees, W.; Debatty, T. A comparative analysis of visualisation techniques to achieve cyber situational awareness in the military. In Proceedings of the 2018 International Conference on Military Communications and Information Systems (ICMCIS), Varsoiva, Poland, 22–23 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–7.
45. Carvalho, V.S.; Polidoro, M.J.; Magalhaes, J.P. Owlsight: Platform for real-time detection and visualization of cyber threats. In Proceedings of the 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), New York, NY, USA, 8–10 April 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 61–66.
46. Pietrowicz, S.; Falchuk, B.; Kolarov, A.; Naidu, A. Web-Based Smart Grid Network Analytics Framework. In Proceedings of the 2015 IEEE International Conference on Information Reuse and Integration, San Francisco, CA, USA, 13–15 August 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 496–501.

47. Matuszak, W.J.; DiPippo, L.; Sun, Y.L. Cybersave: Situational awareness visualization for cyber security of smart grid systems. In Proceedings of the Tenth Workshop on Visualization for Cyber Security, Atlanta, GA, USA, 14 October 2013; pp. 25–32.
48. Kabil, A.; Duval, T.; Cuppens, N. Alert characterization by non-expert users in a cybersecurity virtual environment: A usability study. In Proceedings of the International Conference on Augmented Reality, Virtual Reality and Computer Graphics, Lecce, Italy, 7–10 September 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 82–101.
49. Kullman, K.; Cowley, J.; Ben-Asher, N. Enhancing cyber defense situational awareness using 3D visualizations. In Proceedings of the 13th International Conference on Cyber Warfare and Security ICCWS 2018, National Defense University, Washington, DC, USA, 8–9 March 2018; pp. 369–378.
50. Kullman, K.; Asher, N.B.; Sample, C. Operator impressions of 3D visualizations for cybersecurity analysts. In Proceedings of the ECCWS 2019 18th European Conference on Cyber Warfare and Security, Coimbra, Portugal, 4–5 July 2019; Academic Conferences and publishing limited: Red Hook, NY, USA, 2019; p. 257.
51. Reed, J. Threat Hunting with ML: Another Reason to SMLE. 17 February 2021. Available online: https://www.splunk.com/en_us/blog/platform/threat-research-at-splunk-using-smle.html (accessed on 28 March 2023).
52. Liang, J.; Kim, Y. Evolution of Firewalls: Toward Securer Network Using Next Generation Firewall. In Proceedings of the 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Virutal, 26–29 January 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 752–759.
53. IBM X-Force Exchange. Available online: https://exchange.xforce.ibmcloud.com/ (accessed on 3 March 2023).
54. The Security Immune System: An Integrated Approach to Protecting Your Organization. Available online: https://www.midlandinfosys.com/pdf/qradar-siem-cybersecurity-ai-products.pdf (accessed on 3 March 2023).
55. Anomali ThreatStream: Automated Threat Intelligence Management at Scale. Available online: https://www.anomali.com/products/threatstream (accessed on 3 March 2023).
56. Wang, B.; Najjar, L.; Xiong, N.N.; Chen, R.C. Stochastic optimization: Theory and applications. *J. Appl. Math.* **2013**, *2013*, 949131. [CrossRef]
57. McCall, J. Genetic algorithms for modelling and optimisation. *J. Comput. Appl. Math.* **2005**, *184*, 205–222. [CrossRef]
58. Jangla, K. Docker Compose. In *Accelerating Development Velocity Using Docker*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 77–98.
59. Li, Y.; Li, W.; Jiang, C. A survey of virtual machine system: Current technology and future trends. In Proceedings of the 2010 Third International Symposium on Electronic Commerce and Security, Guangzhou, China, 29–31 July 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 332–336.
60. Medel, V.; Rana, O.; Bañares, J.Á.; Arronategui, U. Modelling performance & resource management in kubernetes. In Proceedings of the 9th International Conference on Utility and Cloud Computing, Shanghai, Chine, 6–9 December 2016; pp. 257–262.
61. Kotas, C.; Naughton, T.; Imam, N. A comparison of Amazon Web Services and Microsoft Azure cloud platforms for high performance computing. In Proceedings of the 2018 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 12–14 January 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–4.
62. Gray, J.; Siewiorek, D.P. High-availability computer systems. *Computer* **1991**, *24*, 39–48. [CrossRef]
63. Wilson, K.S. Conflicts among the pillars of information assurance. *IT Prof.* **2012**, *15*, 44–49. [CrossRef]
64. Rinaldi, S.M.; Peerenboom, J.P.; Kelly, T.K. Identifying, understanding, and analyzing critical infrastructure interdependencies. *IEEE Control Syst. Mag.* **2001**, *21*, 11–25.
65. Fleissner, S.; Baniassad, E. A commensalistic software system. In Proceedings of the Companion to the 21st ACM SIGPLAN Symposium on Object-Oriented Programming Systems, Languages, and Applications, Portland, OR, USA, 22–26 October 2006; pp. 560–573.
66. Torchiano, M.; Jaccheri, L.; Sørensen, C.F.; Wang, A.I. COTS products characterization. In Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering, Ischia, Italy, 15–19 July 2002; pp. 335–338.
67. Coppolino, L.; D'Antonio, S.; Formicola, V.; Romano, L. Integration of a System for Critical Infrastructure Protection with the OSSIM SIEM Platform: A dam case study. In Proceedings of the International Conference on Computer Safety, Reliability, and Security, Naples, Italy, 19–22 September 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 199–212.
68. Cerullo, G.; Formicola, V.; Iamiglio, P.; Sgaglione, L. Critical Infrastructure Protection: Having SIEM technology cope with network heterogeneity. *arXiv* **2014**, arXiv:1404.7563.
69. Veselý, V. Extended Comparison Study on Merging PCAP Files. *ElectroScope* **2012**, *2012*, 1–6.
70. Wagner, C.; Dulaunoy, A.; Wagener, G.; Iklody, A. Misp: The design and implementation of a collaborative threat intelligence sharing platform. In Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security, Vienna, Austria, 24 October 2016; pp. 49–56.
71. Groenewegen, A.; Janssen, J.S. *TheHive Project: The Maturity of an Open-Source Security Incident Response Platform*; SNE/OS3; University of Amsterdam: Amsterdam, The Netherlands, 2021.
72. Gonashvili, M. *Knowledge Management for Incident Response Teams*; Masaryk University: Brno, Czech Republic, 2019.
73. Cole, E. *Advanced Persistent Threat: Understanding the Danger and How to Protect Your Organization*; Syngress: Oxford, UK, 2012.
74. Tabatabaei, F.; Wells, D. OSINT in the Context of Cyber-Security. *Open Source Intell. Investig.* **2016**, *1*, 213–231.
75. Verhoef, R. Sigma Rules! The Generic Signature Format for SIEM Systems. 19 June 2020. Available online: https://isc.sans.edu/diary/rss/26258 (accessed on 7 February 2023).

76. Ömer. What Is Sigma? Threat Hunting in Siem Products with Sigma Rules–Example Sigma Rules. 21 March 2021. Available online: https://www.systemconf.com/2021/03/21/what-is-sigma-threat-hunting-in-siem-products-with-sigma-rules-example-sigma-rules/ (accessed on 7 February 2023).
77. Naik, N.; Jenkins, P.; Savage, N.; Yang, L.; Boongoen, T.; Iam-On, N.; Naik, K.; Song, J. Embedded YARA rules: Strengthening YARA rules utilising fuzzy hashing and fuzzy rules for malware analysis. *Complex Intell. Syst.* **2021**, *7*, 687–702. [CrossRef]
78. Naik, N.; Jenkins, P.; Savage, N.; Yang, L. Cyberthreat Hunting-Part 1: Triaging ransomware using fuzzy hashing, import hashing and YARA rules. In Proceedings of the 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 23–26 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
79. Knuth, D.E. *The Art of Computer Programming*, 2nd ed.; Sorting and Searching; Addison Wesley Longman Publishing Co., Inc.: Boston, MA, USA, 1998; Volume 3.
80. Gianvecchio, S.; Burkhalter, C.; Lan, H.; Sillers, A.; Smith, K. Closing the Gap with APTs Through Semantic Clusters and Automated Cybergames. In Proceedings of the Security and Privacy in Communication Networks, Orlando, FL, USA, 23–25 October 2019; Chen, S., Choo, K.K.R., Fu, X., Lou, W., Mohaisen, A., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 235–254.
81. Divya, M.S.; Goyal, S.K. ElasticSearch: An advanced and quick search technique to handle voluminous data. *Compusoft* **2013**, *2*, 171.
82. Hancock, J.T.; Khoshgoftaar, T.M. Survey on categorical data for neural networks. *J. Big Data* **2020**, *7*, 28. [CrossRef]
83. Schetinin, V.; Schult, J. A neural-network technique to learn concepts from electroencephalograms. *Theory Biosci.* **2005**, *124*, 41–53. [CrossRef]
84. Gallant, S.I.; Gallant, S.I. *Neural Network Learning and Expert Systems*; MIT Press: Cambridge, MA, USA, 1993.
85. Murthy, S.K.; Kasif, S.; Salzberg, S. A system for induction of oblique decision trees. *J. Artif. Intell. Res.* **1994**, *2*, 1–32. [CrossRef]
86. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
87. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: A new data clustering algorithm and its applications. *Data Min. Knowl. Discov.* **1997**, *1*, 141–182. [CrossRef]
88. Zhang, T.; Ramakrishnan, R.; Livny, M. BIRCH: An efficient data clustering method for very large databases. *ACM Sigmod Rec.* **1996**, *25*, 103–114. [CrossRef]
89. Khan, K.; Rehman, S.U.; Aziz, K.; Fong, S.; Sarasvady, S. DBSCAN: Past, present and future. In Proceedings of the Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014), Chennai, India, 17–19 February 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 232–238.
90. Çelik, M.; Dadaşer-Çelik, F.; Dokuz, A.Ş. Anomaly detection in temperature data using DBSCAN algorithm. In Proceedings of the 2011 International Symposium on Innovations in Intelligent Systems and Applications, Istanbul, Turkey, 15–18 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 91–95.
91. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 413–422.
92. Ding, Z.; Fei, M. An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. *IFAC Proc. Vol.* **2013**, *46*, 12–17. [CrossRef]
93. Amer, M.; Goldstein, M.; Abdennadher, S. Enhancing one-class support vector machines for unsupervised anomaly detection. In Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, Chicago, Illinois, 11 August 2013; pp. 8–15.
94. Hejazi, M.; Singh, Y.P. One-class support vector machines approach to anomaly detection. *Appl. Artif. Intell.* **2013**, *27*, 351–366. [CrossRef]
95. Ukwen, D.O.; Karabatak, M. Review of NLP-based Systems in Digital Forensics and Cybersecurity. In Proceedings of the 2021 9th International Symposium on Digital Forensics and Security (ISDFS), Elazig, Turkey, 28–29 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–9.
96. Georgescu, T.M. Natural language processing model for automatic analysis of cybersecurity-related documents. *Symmetry* **2020**, *12*, 354. [CrossRef]
97. Mathews, S.M. Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review. In Proceedings of the Intelligent Computing-Proceedings of the Computing Conference, London, UK, 16–17 July 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 1269–1292.
98. Al-Omari, M.; Rawashdeh, M.; Qutaishat, F.; Alshira'H, M.; Ababneh, N. An intelligent tree-based intrusion detection model for cyber security. *J. Netw. Syst. Manag.* **2021**, *29*, 20. [CrossRef]
99. Sarker, I.H. Deep cybersecurity: A comprehensive overview from neural network and deep learning perspective. *SN Comput. Sci.* **2021**, *2*, 154.
100. Fang, H. Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In Proceedings of the 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), Shenyang, China, 8–12 June 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 820–824.
101. Goyal, G.; Singh, K.; Ramkumar, K. A detailed analysis of data consistency concepts in data exchange formats (JSON & XML). In Proceedings of the 2017 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, 5–6 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 72–77.

102. Barnum, S. Standardizing cyber threat intelligence information with the structured threat information expression (stix). *Mitre Corp.* **2012**, *11*, 1–22.
103. Riesco, R.; Villagrá, V.A. Leveraging cyber threat intelligence for a dynamic risk framework. *Int. J. Inf. Secur.* **2019**, *18*, 715–739. [CrossRef]
104. Na, S.; Kim, T.; Kim, H. A study on the classification of common vulnerabilities and exposures using naïve bayes. In Proceedings of the International Conference on Broadband and Wireless Computing, Communication and Applications, Asan, Republic of Korea, 5–7 November 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 657–662.
105. Radack, S.; Kuhn, R. Managing security: The security content automation protocol. *IT Prof.* **2011**, *13*, 9–11. [CrossRef]
106. VirusTotal: Analyse Suspicious Files, Domains, IPs and URLs to Detect Malware and Other Breaches, Automatically Share Them with the Security Community. Available online: https://www.virustotal.com (accessed on 3 March 2023).
107. URLhaus: Malware URL Exchange. Available online: https://urlhaus.abuse.ch/ (accessed on 3 March 2023).
108. Masse, M. *REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces*; O'Reilly Media, Inc.': Sebastopol, CA, USA, 2011.
109. Naik, N. Choice of effective messaging protocols for IoT systems: MQTT, CoAP, AMQP and HTTP. In Proceedings of the 2017 IEEE International Systems Engineering Symposium (ISSE), Vienna, Austria, 11–13 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–7.
110. Sandhu, R.S.; Coyne, E.J.; Feinstein, H.L.; Youman, C.E. Role-based access control models. *Computer* **1996**, *29*, 38–47. [CrossRef]
111. Tomasek, M.; Cerny, T. On web services ui in user interface generation in standalone applications. In Proceedings of the 2015 Conference on Research in Adaptive and Convergent Systems, Prague, Czech Republic, 9–12 October 2015; pp. 363–368.
112. Montesi, F.; Weber, J. Circuit breakers, discovery, and API gateways in microservices. *arXiv* **2016**, arXiv:1609.05830.
113. Xu, R.; Jin, W.; Kim, D. Microservice security agent based on API gateway in edge computing. *Sensors* **2019**, *19*, 4905. [CrossRef] [PubMed]
114. Jeong, J.; Chung, M.Y.; Choo, H. Integrated OTP-based user authentication scheme using smart cards in home networks. In Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), Big Island, HI, USA, 7–10 January 2008; IEEE: Piscataway, NJ, USA, 2008; p. 294.
115. Zhao, S.; Hu, W. Improvement on OTP authentication and a possession-based authentication framework. *Int. J. Multimed. Intell. Secur.* **2018**, *3*, 187–203. [CrossRef]
116. Bihis, C. *Mastering OAuth 2.0*; Packt Publishing Ltd.: Birmingham, UK, 2015.
117. Hardt, D. The OAuth 2.0 Authorization Framework. RFC 6749, RFC Editor, 2012. Available online: http://www.rfc-editor.org/rfc/rfc6749.txt (accessed on 28 March 2023).
118. Haag, S.; Anderl, R. Digital twin–Proof of concept. *Manuf. Lett.* **2018**, *15*, 64–66. [CrossRef]
119. Srinath, K. Python–the fastest growing programming language. *Int. Res. J. Eng. Technol.* **2017**, *4*, 354–357.
120. Nelli, F. *Python Data Analytics: Data Analysis and Science Using PANDAs, Matplotlib and the Python Programming Language*; Apress: Sebastopol, CA, USA, 2015.
121. Hao, J.; Ho, T.K. Machine learning made easy: A review of scikit-learn package in python programming language. *J. Educ. Behav. Stat.* **2019**, *44*, 348–361. [CrossRef]
122. Al-Shaer, R.; Spring, J.M.; Christou, E. Learning the associations of mitre att & ck adversarial techniques. In Proceedings of the 2020 IEEE Conference on Communications and Network Security (CNS), Virtual, 28–30 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–9.
123. Alexander, O.; Belisle, M.; Steele, J. *MITRE ATT&CK for Industrial Control Systems: Design and Philosophy*; The MITRE Corporation: Bedford, MA, USA, 2020.
124. Ahmed, M.; Panda, S.; Xenakis, C.; Panaousis, E. MITRE ATT&CK-driven cyber risk assessment. In Proceedings of the 17th International Conference on Availability, Reliability and Security, Vienna, Austria, 23–26 August 2022; pp. 1–10.
125. Roy, G.M. *RabbitMQ in Depth*; Simon and Schuster: New York, NY, USA, 2017.
126. Ionescu, V.M. The analysis of the performance of RabbitMQ and ActiveMQ. In Proceedings of the 2015 14th RoEduNet International Conference-Networking in Education and Research (RoEduNet NER), Craiova, Romania, 24–26 September 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 132–137.
127. Rostanski, M.; Grochla, K.; Seman, A. Evaluation of highly available and fault-tolerant middleware clustered architectures using RabbitMQ. In Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, Warsaw, Poland, 7–10 September 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 879–884.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

big data and
cognitive computing

MDPI

Article

PSO-Driven Feature Selection and Hybrid Ensemble for Network Anomaly Detection

Maya Hilda Lestari Louk [1] and Bayu Adhi Tama [2,*]

[1] Department of Informatics Engineering, University of Surabaya, Surabaya 60293, Indonesia
[2] Department of Information Systems, University of Maryland, Baltimore County (UMBC), Baltimore, MD 21250, USA
* Correspondence: bayu@umbc.edu

Abstract: As a system capable of monitoring and evaluating illegitimate network access, an intrusion detection system (IDS) profoundly impacts information security research. Since machine learning techniques constitute the backbone of IDS, it has been challenging to develop an accurate detection mechanism. This study aims to enhance the detection performance of IDS by using a particle swarm optimization (PSO)-driven feature selection approach and hybrid ensemble. Specifically, the final feature subsets derived from different IDS datasets, i.e., NSL-KDD, UNSW-NB15, and CICIDS-2017, are trained using a hybrid ensemble, comprising two well-known ensemble learners, i.e., gradient boosting machine (GBM) and bootstrap aggregation (bagging). Instead of training GBM with individual ensemble learning, we train GBM on a subsample of each intrusion dataset and combine the final class prediction using majority voting. Our proposed scheme led to pivotal refinements over existing baselines, such as TSE-IDS, voting ensembles, weighted majority voting, and other individual ensemble-based IDS such as LightGBM.

Keywords: multi-stage ensemble; particle swarm optimization; feature selection; anomaly detection; intrusion detection

Citation: Louk, M.H.L.; Tama, B.A. PSO-Driven Feature Selection and Hybrid Ensemble for Network Anomaly Detection. *Big Data Cogn. Comput.* **2022**, *6*, 137. https://doi.org/10.3390/bdcc6040137

Academic Editors: Yang-Im Lee and Peter R.J. Trim

Received: 3 October 2022
Accepted: 10 November 2022
Published: 13 November 2022

1. Introduction

An intrusion detection system, often known as an IDS, has the potential to make significant contributions to the field of information security research due to its capability to monitor and identify unauthorized access targeted at computing and network resources [1,2]. In conjunction with other mitigation techniques, such as access control and user authentication, an IDS is often utilized as a secondary line of defense in computer networks. In the past few decades, machine learning techniques have been applied to the network audit log to construct models for identifying attacks [3]. In this scenario, intrusion detection can be viewed as a data analytics process in which machine learning techniques are used to automatically uncover and model characteristics of a user's suspicious or normal behavior. Ensemble learning is a popular machine learning approach in which multiple distinct classifiers are weighted and combined to produce a classifier that outperforms each of them individually [4].

Tama and Lim [5] looked at how recent ensemble learning techniques have been exploited in IDS through a systematic mapping study. They argued that ensemble learning has made a significant difference over standalone classifiers, though this is sometimes the case, depending upon the voting schemes and base classifiers used to build the ensemble. This makes it challenging to design an accurate detection mechanism based on ensemble learning. Moreover, an IDS has to cope with an enormous amount of data that may contain unimportant features, resulting in poor performance. Consequently, selecting relevant features is considered a crucial criterion for IDS [6,7]. Feature selection minimizes redundant information, improves detection algorithm accuracy, and enhances generalization.

This article focuses on evaluating anomaly-based IDS by leveraging the combination of a feature selection technique and hybrid ensemble learning. More precisely, we adopt a particle swarm optimization (PSO) method as a search algorithm to traverse the whole feature space and assess potential feature subsets. Next, a hybrid ensemble learning approach, comprising two ensemble paradigms—gradient boosting machine (GBM) [8] and bootstrap aggregation (bagging) [9]—is utilized to improve the detection accuracy. Our proposed detector, combined with a feature selection technique, can substantially affect the performance accuracy of network anomaly detection with a comparable result over existing baselines. To put it in a nutshell, this article presents advancements to the existing IDS techniques.

(a) A simple yet accurate network anomaly detection using hybrid bagging and GBM ensemble is proposed. GBM is not trained independently as a classifier; rather, we use it as the base learning model for bagging in order to increase its detection performance.
(b) A PSO-guided feature selection is applied to choose the most optimal subset of features for the input of the hybrid ensemble model. The full feature set may not give substantial prediction accuracy; thus, we use an optimum feature subset derived from the PSO-based feature selection approach.
(c) Based on our experiment validation, our proposed model is superior compared to existing anomaly-based IDS methods presented in the current literature.

We break down the remaining parts of this article as follows. In Section 2, a brief survey of prior detection techniques is provided, followed by the description of the datasets and hybrid ensemble in Section 3. The experimental result is discussed in Section 4; lastly, some closing notes are given in Section 5.

2. Related Work

Ensemble learning approaches are not a novel IDS methodology. In IDS, combining multiple weak classifiers to generate a robust classifier has been discussed for a very significant period of time [5,10–15]. In this section, existing anomaly-based IDS methods employing feature selection and ensemble learning are explored briefly. It is worth mentioning that in order to give the most up-to-date literature on anomaly detectors, we have included publications published between 2020 and the present. Table 1 presents a summarization of each existing work published as an article, listed in chronological order.

Table 1. Summarization of prior anomaly-based intrusion detection techniques that employ feature selection and ensemble learning. The articles are chronologically ordered between 2020 and the present.

Author(s)	Ensemble Approach(es)	Base Learner(s)	Feature Selector	Validation Method(s)	Dataset(s)
[16]	Stacking	NN, NB, DL, SVM	IG	Hold-out	Private
[17]	AB, stacking	LR, RF	PCA	CV and hold-out	NSL-KDD, UNSW-NB15
[18]	RF, XGBoost, HGB, LightGBM	-	RF+PCA	CV	CICIDS-2018
[19]	XGBoost	-	GA	CV	CIRA-CIC-DoHBrw-2020, Bot-IoT, UNSW-NB15
[20]	RF	-	Gain ratio, Chi-squared, Pearson correlation	Hold-out	UNSW-NB15
[21]	Stacking	RF, LR	K-means	Hold-out	NSL-KDD, CIDDS-2017, Testbed

Table 1. *Cont.*

Author(s)	Ensemble Approach(es)	Base Learner(s)	Feature Selector	Validation Method(s)	Dataset(s)
[22]	Majority voting	SVM, NB, LR, DT	Filter and univariate ensemble	CV	Honeypot, NSL-KDD, Kyoto
[23]	LightGBM	-	-	CV	NSL-KDD, UNSW-NB15, CICIDS-2017
[24]	RF	-	-	Hold-out	CIDDS-001, UNSW-NB15
[25]	Weighted voting	C4.5, MLP, IBL	IFA	CV	NSL-KDD, UNSW-NB15
[26]	RF	-	-	CV	NSL-KDD, UNSW-NB15, CICIDS-2017
[27]	XGBoost, RF	-	-	Hold-out	NSL-KDD, CIDDS-001, CICIDS-2017
[28]	Weighted majority voting	SVM, LR, NB, DT	Gain-ratio, Chi-squared, Information gain	Hold-out	Honeypot, NSL-KDD, Kyoto
[29]	Stacking	DT, RF, XGBoost	*SelectKbest*	CV	NSL-KDD, UNSW-NB15
[30]	LightGBM	-	DNN	Hold-out	KDD-99, NSL-KDD, UNSW-NB15

Stacking [31] has been commonly mentioned as one of the ensemble procedures. It is a general method in which a classification algorithm is trained to integrate heterogeneous algorithms. Individual algorithms are referred to as first-level algorithms, while the combiner is referred to as a second-level algorithm or meta-classifier. Jafarian et al. [16], Kaur [17], Jain and Kaur [21], Rashid et al. [29], Wang et al. [30] demonstrate that stacking generates a promising intrusion detection capability; however, most of the proposed stacking procedures do not consider LR as a second-level algorithm, as suggested by [32]. Alternatively, combiner strategies, such as majority voting [22] and weighted majority voting [25,28] may be utilized as anomaly detectors. The most prevalent mode of voting is majority rule. In this context, each algorithm casts a vote for one class label, with the class label receiving more than fifty percent of the votes serving as the final output class label; if none of the class labels acquires more than fifty percent of the votes, a rejection choice will be given, and the blended algorithm will not make a prediction. On the other hand, if individual algorithms have inequitable performance, it seems reasonable to assign the more robust algorithms more significant influence during voting; this is achieved by weighted majority voting.

Furthermore, it is possible to construct homogeneous ensembles in which an ensemble procedure is built upon a single (e.g., the same type) algorithm. Kaur [17] compares three different adaptive boosting (AB) [33] families of algorithms for anomaly-based IDS, while the rest of proposed approaches utilize tree-based ensemble learning, such as RF [18,20,24,26,27], LightGBM [18,23,30], and XGBoost [18,19,27].

In the intrusion detection field, feature selection techniques have also been exploited [34,35]. Specifically, bio-inspired algorithms have gained popularity and evolved into an alternate method for finding the optimal feature subset from the feature space [19,25,36]. Other filter-based approaches such as IG, gain ratio, chi-squared, and Pearson correlation have been intensively utilized to remove unnecessary features [16,20,22,28,29]. The filter technique assesses feature subsets according to given criteria regardless of any grouping. Information gain, for example, utilizes a weighted feature scoring system to obtain the highest entropy value. In addition, previous research indicates that feature selectors using the wrapper technique are taken into account. A wrapper-based feature selector evaluates a specific machine learning algorithm to search optimal feature subset [17,18,21,30]. Examining the above-mentioned methods for anomaly detectors, our study fills a gap by examining hybrid ensemble and PSO-based feature selection, both of which are underexplored in the existing literature.

3. Materials and Methods

This seeks assess the performance of network anomaly detection using PSO-based feature selection and hybrid ensemble. Figure 1 denotes the phases of our detection framework.

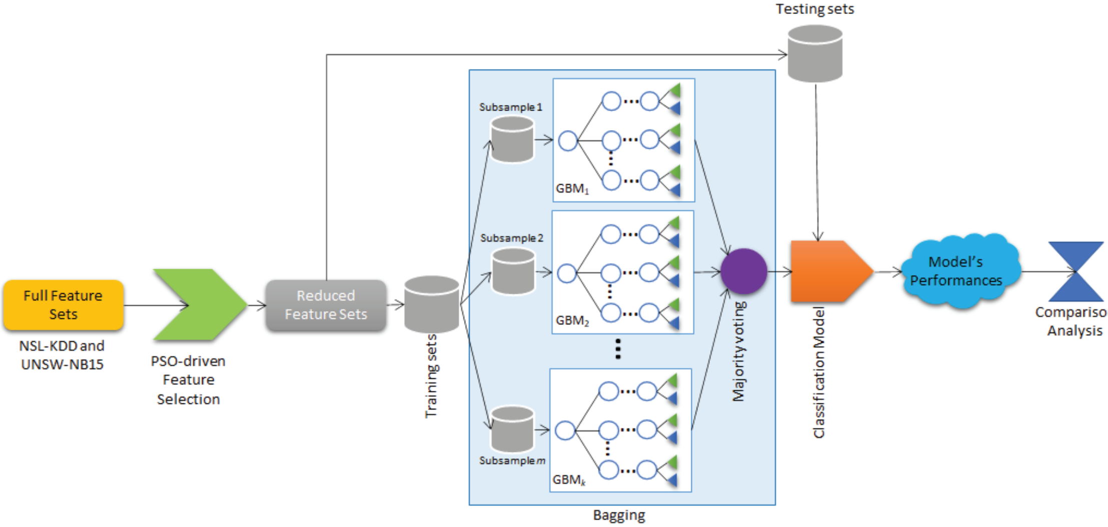


Figure 1. Proposed framework for intrusion detection based on PSO-driven feature selection and hybrid ensemble.

A PSO-driven feature selection technique is applied to identify the optimum feature subsets. Next, each dataset with an optimal feature subset is split into a training set and a testing set, where the training set is used to construct a classification model (e.g., a bagging–GBM model), and the testing set is used to validate the model's performance. Finally, different combinations of ensemble methods are statistically assessed and contrasted, along with a comparison study with prior works. In the following section, we break down the datasets used in our study, as well as the concept of our anomaly-based IDS.

3.1. Data Sets

In this study, we focus on using three distinct datasets, namely, NSL-KDD [37], UNSW-NB15 [38], and CICIDS-2017 [39]. Both datasets are extensively used for appraising IDS models and have been considered as standard benchmark datasets. The NSL-KDD dataset is an enhanced variant of its earlier versions, KDD Cup 99, which was the subject of widespread debate due to data redundancy, performance bias for machine learning algorithms, and unrealistic representation of attacks. We use an original training set of NSL-KDD (e.g., KDDTrain) that contains seven categorical input features and 34 numerical input features. There are a total of 25,192 samples, which are assigned as follows: 13,449 normal samples and 11,743 attack samples.

Furthermore, two independent testing sets (e.g., KDDTest-21 and KDDTest+) are used to appraise our proposed anomaly detector. KDDTest-21 and KDDTest+ consist of 11,850 samples and 22,544 samples, respectively. On the other hand, the UNSW-NB15 dataset also contains two primary sets, i.e., UNSW-NB15-Train and UNSW-NB15-Test, which are used for training and evaluating the model, respectively. The UNSW-NB15-Train includes six categorical input features and 38 numerical input features. There are a total of 82,332 samples, 45,332 of which are attack samples and 37,000 of which are normal samples. The UNSW-NB15-Test possesses a total of 175,341 samples, including 119,341 attack samples and 56,000 normal samples. The original version of the CICIDS-2017 dataset consists of 78 numerical input features and 170,366 samples, of which 168,186 are benign and 2180 are malicious. Given that the CICIDS-2017 does not provide predetermined training and testing sets, we employ holdout with a ratio of 80/20 for training and testing, respectively. Therefore, the CICIDS-2017 training set includes 136,293 instances that are proportionally sampled from the original dataset. The characteristics of the training datasets are outlined in Table 2.

Table 2. Description of training data sets.

Dataset	**#Total Samples**	**#Samples Labelled Normal**	**#Samples Labelled Anomaly**	**#Categorical Features**	**#Numerical Features**
NSL-KDD	25,192	13,449	11,743	7	34
UNSW-NB15	82,332	37,000	45,332	6	38
CICIDS-2017	136,292	134,548	1744	-	78

3.2. Methods

3.2.1. PSO-Based Feature Selection

A feature selection approach is a strategy for determining a granular, concise, and plausible subset of a particular set of features. In this work, we pick a correlation-based feature selection (CFS) method [40] that measures the significance of features using entropy and information gain. At the same time, a particle swarm optimization (PSO) algorithm [41] is taken into account as a search technique. A particle swarm optimization (PSO)-based feature selection approach models a feature set as a collection of particles that make up a swarm. A number of particles are scattered across a hyperspace and each of those particles is given a position ξ_n and velocity v_n, which are entirely random. Let $\mathbf{w}$ represents the inertia weight constant, and δ_1 and δ_2 represent the cognitive and social learning constants, respectively. Next, let σ_1 and σ_2 denote the random numbers, $\mathbf{l}_n$ denote the personal best location of particle n, and $\mathbf{g}$ denote the global location across the particles. The following are thus the basic rules for updating the position and velocity of each particle:

$$\xi_n(t+1) = \xi_n(t) + v_n(t+1) \tag{1}$$

$$v_n(t+1) = \mathbf{w}v_n(t) + \delta_1\sigma_1(\mathbf{l}_n - \xi_n(t)) + \delta_2\sigma_2(\mathbf{g} - \xi_n(t)) \tag{2}$$

3.2.2. Hybrid Ensemble Based on Bagging-GBM

The proposed hybrid ensemble is constructed based on a fusion of two individual ensemble learners, i.e., bagging [9] and gradient boosting machine (GBM) [8]. In lieu of training a bagging ensemble with a weak classifier, we employ another ensemble, e.g., GBM, as the base classifier of bagging. A bagging strategy is devised using $\mathcal{K}$ GBMs built from bootstrap replicates β of the training set. A training set containing π instances will be used to generate subsamples by sampling with replacement. Some peculiar instances appear several times in the subsamples, but others do not. Each individual GBM can then be trained on each subsample. Final class prediction is determined by the majority voting rule (e.g., each voter may only choose a single class label, and the class label prediction that gathers more than fifty percent of the most votes is chosen). We present a more formal way description of bagging–GBM in Algorithm 1.

Algorithm 1: A procedure to construct bagging–GBM for anomaly-based IDS.

```
Building classification model:
Require: Training set $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_\alpha, y_\alpha)\}$; base classifier (e.g.,
  GBM); number of GBMs $\mathcal{K}$; size of subsample $\gamma$.
1. $\kappa \leftarrow 1$
2. repeat
3.     $D_\kappa \leftarrow$ replacement-based subsample of $\gamma$ instances from $D$.
4.     Construct classifier $h_\kappa$ using GBM on $D_\kappa$.
5.     $\kappa \leftarrow \kappa + 1$
6. until $\kappa > \mathcal{K}$
Evaluating classification model:
Require: An object deserving of a classification $\mathbf{x}$.
Output: Final class label prediction $\tau$
1. $Counter_1, \ldots, Counter_y \leftarrow 0$
2. for $i = 1$ to $\mathcal{K}$ do
3.     $vote_i \leftarrow h_i(\mathbf{x})$
4.     $Counter_{vote_i} \leftarrow Counter_{vote_i} + 1$
5. end for
6. $\tau \leftarrow$ the most prevalent class label chosen by constituents.
7. Return $\tau$
```

3.2.3. Evaluation Criteria

3.3. Metrics

The objective of a performance evaluation is to ensure that the proposed model works correctly with the IDS datasets. In addition, such an assessment seeks specific criteria so that the effectiveness of the proposed model can be better justified. As an anomaly-based IDS is a binary classification problem, we utilize various performance indicators that are relevant to the task, such as accuracy (Acc), precision, recall, balanced accuracy (BAcc), AUC, F1, and MCC. It is important to note that various metrics have been applied in prior research, except for BAcc and MCC, which have not been widely utilized. Balanced accuracy shows benefits over general accuracy as a metric [42], while MCC is a reliable measure that describes the classification algorithm in a single value, assuming that anomalous and normal samples are of equal merit [43]. More precisely, BAcc is specified as the arithmetic mean of the true positive rate (TPR) and true negative rate (TNR) as follows.

$$\text{BAcc} = \frac{1}{2} \times (TPR + TNR) \tag{3}$$

MCC assesses the strength of the relationship between the actual classes a and predicted labels p:

$$\text{MCC} = \frac{\text{Cov}(a, p)}{\sigma_a \times \sigma_p} \tag{4}$$

where $\text{Cov}(a, p)$ is the covariance between the actual classes a and predicted labels p, while σ_a and σ_p are the standard deviations of the actual classes a and predicted labels p, respectively.

3.4. Validation Procedure

As stated in Section 3.1, except for the CICIDS-2017 dataset, each intrusion dataset was built with a predefined split between training and testing sets. As a result, we utilized such a training/testing split (e.g., hold-out) as a validation strategy in the experiment. The hold-out procedure was repeated five times for each classification algorithm to verify that the performance results were not achieved by chance. The final performance value was calculated by averaging the five performance values.

4. Results and Discussion

The experimental assessment of the proposed framework is presented and discussed in this section. The final subsets of the NSL-KDD and UNSW-NB15 derived by PSO-based feature selection are taken from our earlier solutions reported in [6,7]. Here, 38 optimal features from the NSL-KDD and 20 optimal features from the UNSW-NB15 were employed, respectively. In contrast, the proposed feature selection identifies 17 optimal features from the original CICIDS-2017 dataset.

Furthermore, we appraised the potency of the proposed model under several ensemble strategies corresponding to different ensemble sizes. The size of the ensemble was determined by the number of base classifiers (e.g., GBM in our example) used to train the ensemble (e.g., bagging in our case). For instance, GBM-2 indicates that two GBMs were included when training the bagging ensemble, and so on. The experiment was conducted on a Linux operating system, 32 GB, and Intel Core i5 using the *R* program. Figure 2 shows the performance average with five times of hold-out for each ensemble strategy. The plot also depicts the performance of the base classifier as a standalone classifier. Taking AUC, F1, and MCC metrics as examples, the proposed model surpasses the individual classifier in all datasets considered by a substantial margin.

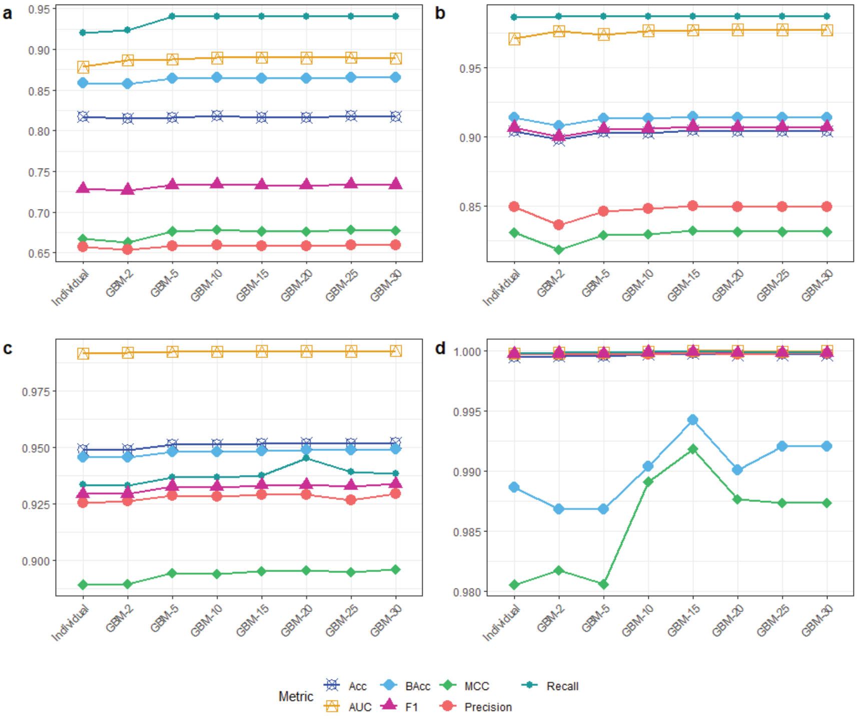


Figure 2. Performance average of all classification algorithms on KDDTest-21 (**a**), KDDTest+ (**b**), UNSW-NB15-Test (**c**), and CICIDS-2017 (**d**).

We next analyzed the performance difference of all algorithms using statistical significance tests. Here, we adopted two statistical omnibus tests, namely the Friedman test and the Nemenyi posthoc test [44]. Performance differences across classification algorithms were calculated by Friedman rank, as illustrated in Table 3. Each algorithm was given a rank for each dataset based on the MCC score, and the average rank of each algorithm was then determined. Table 3 demonstrates that bagging with 30 GBMs (e.g., GBM-30) was the top-performing algorithm, followed by GBM-15. Interestingly, GBM-2 was the weakest performer, failing to outperform a standalone GBM model.

Table 3. Friedman rank matrix of all classifiers relative to each dataset with respect to MCC metric. Bold indicates the best rank, while the second best is underlined. The Friedman test indicates that performance differences across algorithms are significant (p-value < 0.05).

Dataset	**GBM-10**	**GBM-15**	**GBM-2**	**GBM-20**	**GBM-25**	**GBM-30**	**GBM-5**	**Individual**
CICIDS-2017	2	1	6	3	4	5	7	8
KDDTest-21	2	4	8	6	1	3	5	7
KDDTest+	6	1	8	3	4	2	7	5
UNSW-NB15-Test	6	3	7	2	4	1	5	8
Average rank	4.00	<u>2.25</u>	7.25	3.50	3.25	**2.75**	6.00	7.00
p-value					0.01197			

The Nemenyi test employs the Friedman rank; if such average differences are more than or equal to a critical difference (CD), then the performances of such algorithms are substantially different. Figure 3 illustrates that there are no significant performance differences across the benchmarked algorithms, as no average rank exceeds the critical difference (CD) of the Nemenyi test. As shown by a horizontal line, all algorithms are linked. As a final comparison, our best-proposed model (e.g., GBM-30) is compared against existing solutions for anomaly-based IDS. We contrast the efficacy of our proposed scheme to those with a comparative validation approach (e.g., hold-out using predetermined training/test sets).

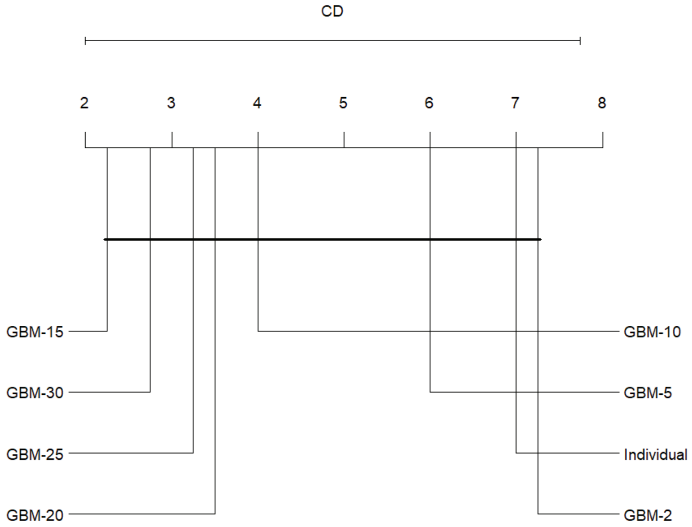


Figure 3. Critical difference plot based on Nemenyi test with respect to MCC metric. Critical difference (CD) is at 5.74, which exceeds the average rank, while all classifiers are tied altogether.

Table 4 compares the performance of our proposed model (e.g., GBM-30) against that of a variety of existing studies published in the latest scientific literature. The proposed model achieves the highest FPR, recall, AUC, and F1 metrics on KDDTest+. Nonetheless, compared to [45], there are minor variations in accuracy and precision measures. Except for the precision metric, our proposed model is the best performer on the KDDTest-21 across all performance criteria. Similarly, on UNSW-NB15-Test and CICIDS-2017, our proposed model outperforms all other models in all performance measures except the FPR metric.

In general, our proposed model is shown to be a feasible solution for anomaly-based IDS, at least for the public datasets addressed in this study. Specifically, with respect to the lowering of FPR and increasing recall, AUC, and F1 scores, our suggested model has shown a significant improvement over the existing studies. In addition, we show the computational time required for individual GBM as well as GBM-15 on the reduced and full feature sets for each dataset in Figure 4. Our feature selection technique significantly lessens the training and testing complexity by roughly one-third compared to the complete feature set, particularly when large datasets such as CICIDS-2017 and UNSW-NB15 are employed.

Table 4. Comparison of the proposed model's outcomes to that of previous network anomaly detectors. Bold indicates the best values.

Ref.	Method	Feature Selection	Acc (%)	FPR (%)	Precision (%)	Recall (%)	AUC	F1
KDDTest+								
[45]	Stacking	-	**92.17**	2.52	-	-	-	-
[46]	Autoencoder	-	84.21	-	-	87.00	-	-
[23]	LightGBM	-	89.79	9.13	-	-	-	-
[26]	MFFSEM	RF	84.33	24.82	74.61	97.15	-	0.841
[28]	Weighted majority voting	GR, IG, and χ^2	85.23	12.8	**90.3**	-	-	0.855
This study	Hybrid ensemble	PSO	90.39	**1.59**	84.94	**98.68**	**0.9767**	**0.907**
KDDTest-21								
[47]	Voting ensemble	CFS-BA	73.57	12.92	73.6	-	-	-
This study	Hybrid ensemble	PSO	**81.72**	**2.1**	65.87	**94.00**	**0.8886**	**0.7332**
UNSW-NB15-Test								
[45]	Stacking	-	92.45	11.3	-	-	-	-
[26]	MFFSEM	RF	88.85	**2.27**	-	80.44	-	-
[20]	RF	GR, χ^2, and PC	83.12	3.7	-	-	-	-
[23]	LightGBM	-	85.89	14.79	-	-	-	-
[30]	LightGBM	DNN	88.34	12.46	-	-	-	0.881
This study	Hybrid ensemble	PSO	**95.20**	4.03	**92.93**	**93.84**	**0.9925**	**0.9338**
CICIDS-2017								
[48]	Rough set theory + Bayes	FPE	97.95	-	-	96.37	-	0.9637
[21]	Stacking	*K*-Means	98.0	**0.2**	97.0	98.0	-	0.98
[49]	ICVAE-BSM	-	99.86	-	99.68	99.68	-	0.9968
This study	Hybrid ensemble	PSO	**99.98**	2.6	**99.99**	**99.99**	**1.00**	**0.9998**

Lastly, we discuss two main implications of our study as follows. First, most previous comparisons were made on particular performance metrics. Our work, however, aims to examine a more trustworthy metric (e.g., MCC) that creates more accurate estimates for the proposed model [43]. The MCC measure could be used to judge future work, especially for detecting network anomalies. Second, a strategy for detecting intrusions should ideally have a low proportion of false positives. Unfortunately, it is nearly impossible to prevent false positives in network anomaly detection. Our work, however, produces the lowest false positive rate on the NSL-KDD dataset and fair results on the UNSW-NB15 and CICIDS-2017.

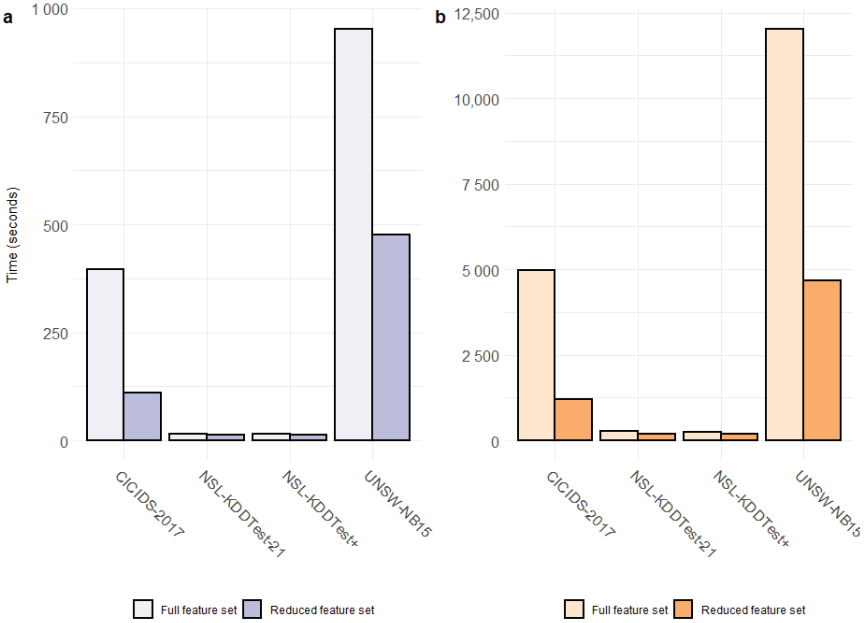


Figure 4. Training and testing complexity for individual GBM (**a**) and GBM-15 (**b**) on reduced and complete feature sets for each data set.

5. Conclusions

An anomaly-based intrusion detection system (IDS) was proposed to thwart any malicious attack and was recognized as a viable method for detecting novel attacks. This work investigated a novel anomaly-based intrusion detection system (IDS) strategy that combines particle swarm optimization (PSO)-guided feature selection with a hybrid ensemble approach. The reduced feature subset was utilized as input for the hybrid ensemble, which was a combination of two well-known ensemble paradigms, including bootstrap aggregation (Bagging) and gradient boosting machine (GBM). The proposed model revealed a substantial performance gain compared to existing studies using the NSL-KDD, UNSW-NB15, and CICIDS-2017 datasets. More specifically, our anomaly detector achieved the lowest FPR at 1.59% and 2.1% on KDDTest+ and KDDTest-21, respectively. With respect to the accuracy, recall, AUC, and F1 metrics, our proposed model consistently surpassed previous research across all datasets considered.

Author Contributions: Conceptualization, M.H.L.L. and B.A.T.; methodology, B.A.T.; validation, M.H.L.L.; investigation, M.H.L.L.; writing—original draft preparation, M.H.L.L.; writing—review and editing, M.H.L.L. and B.A.T.; visualization, B.A.T.; supervision, B.A.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

List of Acronyms

AB	Adaboost
AUC	Area Under ROC Curve
BA	Bat Algorithm
CFS	Correlation-based Feature Selection
CV	Cross Validation
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
FPE	Feature Probability Estimation
GA	Genetic Algorithm
GR	Gain Ratio
HGB	Histogram-based Gradient Boosting
IBL	Instance-based Learning
IFA	Improved Firefly Algorithm
IG	Information Gain
LR	Logistic Regression
MCC	Matthew Correlation Coefficient
MLP	Multilayer Perceptron
NB	Naive Bayes
NN	Neural Network
PC	Pearson Correlation
PCA	Principle Component Analysis
RF	Random Forest
SVM	Support Vector Machine

References

1. Ghorbani, A.A.; Lu, W.; Tavallaee, M. *Network Intrusion Detection and Prevention: Concepts and Techniques*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2009; Volume 47.
2. Bhattacharyya, D.K.; Kalita, J.K. *Network Anomaly Detection: A Machine Learning Perspective*; CRC Press: Boca Raton, FL, USA, 2013.
3. Thakkar, A.; Lohiya, R. A review on machine learning and deep learning perspectives of IDS for IoT: Recent updates, security issues, and challenges. *Arch. Comput. Methods Eng.* **2021**, *28*, 3211–3243. [CrossRef]
4. Rokach, L. *Pattern Classification Using Ensemble Methods*; World Scientific: Singapore, 2010; Volume 75.
5. Tama, B.A.; Lim, S. Ensemble learning for intrusion detection systems: A systematic mapping study and cross-benchmark evaluation. *Comput. Sci. Rev.* **2021**, *39*, 100357. [CrossRef]
6. Tama, B.A.; Rhee, K.H. HFSTE: Hybrid Feature Selections and Tree-Based Classifiers Ensemble for Intrusion Detection System. *IEICE Trans. Inf. Syst.* **2017**, *100D*, 1729–1737. [CrossRef]
7. Tama, B.A.; Comuzzi, M.; Rhee, K.H. TSE-IDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system. *IEEE Access* **2019**, *7*, 94497–94507. [CrossRef]
8. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
9. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
10. Khraisat, A.; Gondal, I.; Vamplew, P.; Kamruzzaman, J. Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity* **2019**, *2*, 20. [CrossRef]
11. Resende, P.A.A.; Drummond, A.C. A Survey of Random Forest Based Methods for Intrusion Detection Systems. *ACM Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]
12. Aburomman, A.A.; Reaz, M.B.I. A survey of intrusion detection systems based on ensemble and hybrid classifiers. *Comput. Secur.* **2017**, *65*, 135–152. [CrossRef]
13. Thakkar, A.; Lohiya, R. A survey on intrusion detection system: Feature selection, model, performance measures, application perspective, challenges, and future research directions. *Artif. Intell. Rev.* **2022**, *55*, 453–563. [CrossRef]
14. Lohiya, R.; Thakkar, A. Application domains, evaluation data sets, and research challenges of IoT: A Systematic Review. *IEEE Internet Things J.* **2020**, *8*, 8774–8798. [CrossRef]
15. Thakkar, A.; Lohiya, R. A review of the advancement in intrusion detection datasets. *Procedia Comput. Sci.* **2020**, *167*, 636–645. [CrossRef]
16. Jafarian, T.; Masdari, M.; Ghaffari, A.; Majidzadeh, K. Security anomaly detection in software-defined networking based on a prediction technique. *Int. J. Commun. Syst.* **2020**, *33*, e4524. [CrossRef]
17. Kaur, G. A comparison of two hybrid ensemble techniques for network anomaly detection in spark distributed environment. *J. Inf. Secur. Appl.* **2020**, *55*, 102601. [CrossRef]

18. Seth, S.; Chahal, K.K.; Singh, G. A novel ensemble framework for an intelligent intrusion detection system. *IEEE Access* **2021**, *9*, 138451–138467. [CrossRef]
19. Halim, Z.; Yousaf, M.N.; Waqas, M.; Sulaiman, M.; Abbas, G.; Hussain, M.; Ahmad, I.; Hanif, M. An effective genetic algorithm-based feature selection method for intrusion detection systems. *Comput. Secur.* **2021**, *110*, 102448. [CrossRef]
20. Nazir, A.; Khan, R.A. A novel combinatorial optimization based feature selection method for network intrusion detection. *Comput. Secur.* **2021**, *102*, 102164. [CrossRef]
21. Jain, M.; Kaur, G. Distributed anomaly detection using concept drift detection based hybrid ensemble techniques in streamed network data. *Clust. Comput.* **2021**, *24*, 2099–2114. [CrossRef]
22. Krishnaveni, S.; Sivamohan, S.; Sridhar, S.; Prabakaran, S. Efficient feature selection and classification through ensemble method for network intrusion detection on cloud computing. *Clust. Comput.* **2021**, *24*, 1761–1779. [CrossRef]
23. Liu, J.; Gao, Y.; Hu, F. A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM. *Comput. Secur.* **2021**, *106*, 102289. [CrossRef]
24. Al, S.; Dener, M. STL-HDL: A new hybrid network intrusion detection system for imbalanced dataset on big data environment. *Comput. Secur.* **2021**, *110*, 102435. [CrossRef]
25. Tian, Q.; Han, D.; Hsieh, M.Y.; Li, K.C.; Castiglione, A. A two-stage intrusion detection approach for software-defined IoT networks. *Soft Comput.* **2021**, *25*, 10935–10951. [CrossRef]
26. Zhang, H.; Li, J.L.; Liu, X.M.; Dong, C. Multi-dimensional feature fusion and stacking ensemble mechanism for network intrusion detection. *Future Gener. Comput. Syst.* **2021**, *122*, 130–143. [CrossRef]
27. Gupta, N.; Jindal, V.; Bedi, P. CSE-IDS: Using cost-sensitive deep learning and ensemble algorithms to handle class imbalance in network-based intrusion detection systems. *Comput. Secur.* **2022**, *112*, 102499. [CrossRef]
28. Krishnaveni, S.; Sivamohan, S.; Sridhar, S.; Prabhakaran, S. Network intrusion detection based on ensemble classification and feature selection method for cloud computing. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e6838. [CrossRef]
29. Rashid, M.; Kamruzzaman, J.; Imam, T.; Wibowo, S.; Gordon, S. A tree-based stacking ensemble technique with feature selection for network intrusion detection. *Appl. Intell.* **2022**, *52*, 9768–9781. [CrossRef]
30. Wang, Z.; Liu, J.; Sun, L. EFS-DNN: An Ensemble Feature Selection-Based Deep Learning Approach to Network Intrusion Detection System. *Secur. Commun. Netw.* **2022**, *2022*, 2693948. [CrossRef]
31. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
32. Ting, K.M.; Witten, I.H. Issues in stacked generalization. *J. Artif. Intell. Res.* **1999**, *10*, 271–289. [CrossRef]
33. Schapire, R.E.; Freund, Y. Boosting: Foundations and algorithms. *Kybernetes* **2013**, *42*, 164–166. [CrossRef]
34. Thakkar, A.; Lohiya, R. Fusion of statistical importance for feature selection in Deep Neural Network-based Intrusion Detection System. *Inf. Fusion* **2022**, *90*, 353–363. [CrossRef]
35. Thakkar, A.; Lohiya, R. Attack classification using feature selection techniques: A comparative study. *J. Ambient Intell. Humaniz. Comput.* **2021**, *12*, 1249–1266. [CrossRef]
36. Thakkar, A.; Lohiya, R. Role of swarm and evolutionary algorithms for intrusion detection system: A survey. *Swarm Evol. Comput.* **2020**, *53*, 100631. [CrossRef]
37. Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDD CUP 99 data set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 8–10 July 2009; pp. 1–6.
38. Moustafa, N.; Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, Australia, 10–12 November 2015; pp. 1–6.
39. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp* **2018**, *1*, 108–116.
40. Hall, M.A. Correlation-Based Feature Selection for Machine Learning. Ph.D. Thesis, The University of Waikato, Hamilton, New Zealand, 1999.
41. Kennedy, J.; Eberhart, R.C. A discrete binary version of the particle swarm algorithm. In Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics, Computational Cybernetics and Simulation, Orlando, FL, USA, 12–15 October 1997; Volume 5, pp. 4104–4108.
42. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. The balanced accuracy and its posterior distribution. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 3121–3124.
43. Chicco, D.; Tötsch, N.; Jurman, G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min.* **2021**, *14*, 13. [CrossRef] [PubMed]
44. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
45. Tama, B.A.; Nkenyereye, L.; Islam, S.R.; Kwak, K.S. An Enhanced Anomaly Detection in Web Traffic Using a Stack of Classifier Ensemble. *IEEE Access* **2020**, *8*, 24120–24134. [CrossRef]
46. Ieracitano, C.; Adeel, A.; Morabito, F.C.; Hussain, A. A novel statistical analysis and autoencoder driven intelligent intrusion detection approach. *Neurocomputing* **2020**, *387*, 51–62. [CrossRef]
47. Zhou, Y.; Cheng, G.; Jiang, S.; Dai, M. Building an efficient intrusion detection system based on feature selection and ensemble classifier. *Comput. Netw.* **2020**, *174*, 107247. [CrossRef]

48. Prasad, M.; Tripathi, S.; Dahal, K. An efficient feature selection based Bayesian and Rough set approach for intrusion detection. *Appl. Soft Comput.* **2020**, *87*, 105980. [CrossRef]
49. Zhang, Y.; Liu, Q. On IoT intrusion detection based on data augmentation for enhancing learning on unbalanced samples. *Future Gener. Comput. Syst.* **2022**, *133*, 213–227. [CrossRef]

big data and
cognitive computing

MDPI

Article

Botnet Detection Employing a Dilated Convolutional Autoencoder Classifier with the Aid of Hybrid Shark and Bear Smell Optimization Algorithm-Based Feature Selection in FANETs

Nejood Faisal Abdulsattar [1], Firas Abedi [2], Hayder M. A. Ghanimi [3], Sachin Kumar [4,*], Ali Hashim Abbas [1,*], Ali S. Abosinnee [5], Ahmed Alkhayyat [6], Mustafa Hamid Hassan [1] and Fatima Hashim Abbas [7]

[1] Department of Computer Technical engineering, College of Information Technology, Imam Ja'afar Al-Sadiq University, Al Muthanna 66002, Iraq
[2] Department of Mathematics, College of Education, Al-Zahraa University for Women, Karbala 56001, Iraq
[3] Biomedical Engineering Department, College of Engineering, University of Warith Al-Anbiyaa, Karbala 56001, Iraq
[4] Big Data and Machine Learning Lab, South Ural State University, 454080 Chelyabinsk, Russia
[5] World Rankings Unit, Altoosi University College, Najaf 54001, Iraq
[6] College of technical engineering, The Islamic University, Najaf 54001, Iraq
[7] Medical Laboratories Techniques Department, Al-Mustaqbal University College, Hillah 51001, Iraq
* Correspondence: kumars@susu.ru (S.K.); alsalamy1987@gmail.com (A.H.A.)

Citation: Abdulsattar, N.F.; Abedi, F.; Ghanimi, H.M.A.; Kumar, S.; Abbas, A.H.; Abosinnee, A.S.; Alkhayyat, A.; Hassan, M.H.; Abbas, F.H. Botnet Detection Employing a Dilated Convolutional Autoencoder Classifier with the Aid of Hybrid Shark and Bear Smell Optimization Algorithm-Based Feature Selection in FANETs. *Big Data Cogn. Comput.* **2022**, *6*, 112. https://doi.org/10.3390/bdcc6040112

Academic Editors: Yang-Im Lee and Peter R.J. Trim

Received: 9 September 2022
Accepted: 27 September 2022
Published: 11 October 2022

Abstract: Flying ad hoc networks (FANETs) or drone technologies have attracted great focus recently because of their crucial implementations. Hence, diverse research has been performed on establishing FANET implementations in disparate disciplines. Indeed, civil airspaces have progressively embraced FANET technology in their systems. Nevertheless, the FANETs' distinct characteristics can be tuned and reinforced for evolving security threats (STs), specifically for intrusion detection (ID). In this study, we introduce a deep learning approach to detect botnet threats in FANET. The proposed approach uses a hybrid shark and bear smell optimization algorithm (HSBSOA) to extract the essential features. This hybrid algorithm allows for searching different feature solutions within the search space regions to guarantee a superior solution. Then, a dilated convolutional autoencoder classifier is used to detect and classify the security threats. Some of the most common botnet attacks use the N-BaIoT dataset, which automatically learns features from raw data to capture a malicious file. The proposed framework is named the hybrid shark and bear smell optimized dilated convolutional autoencoder (HSBSOpt_DCA). The experiments show that the proposed approach outperforms existing models such as CNN-SSDI, BI-LSTM, ODNN, and RPCO-BCNN. The proposed HSBSOpt_DCA can achieve improvements of 97% accuracy, 89% precision, 98% recall, and 98% F1-score as compared with those existing models.

Keywords: FANETs; intrusion detection; botnet attack; deep neural network; feature selection; optimization

1. Introduction

In recent years, unmanned aerial vehicles (UAVs) have attracted additional focus. The use of UAVs provides several distinct benefits over standard human-crewed airplanes, particularly concerning the operative charge, the operator's protection, the UAVs' functionality in arduous or risky settings, and their availability for civil implementations [1]. The latest technological developments have made it easy to set up an unmanned aerial system with a complex topology for crucial operations [2]. Their swift development and intense involvement in intelligent transportation (IT) has significantly affected the path that drone societies have attempted to establish for the prospective UAV systems. The present decentralized technology advances allow for diverse operations and the correlation of resources [3]. This technique permits unnecessary the use of crucial elements and enhances the system's comprehensive strength [4]. Nevertheless, many contemporary developments in the network-attached UAV fleet domain concentrate on the path to attaining a drone network (DN) [5]. Low regard is given to the DN systems' cyber security, resulting in the very advanced DN systems being defenseless against diverse STs [6,7].

This assures the data's secrecy, attainability, and unity while transmitting during UAV-to-UAV transmission, and the safety of UAV-to-ground-node transmission remains a major problem experienced by FANETs. In FANETs, UAVs transfer data that encompass audio, video, image, text, GPS position, and other formats. In transmitting these data, they must possess a fine QoS, having low delay and error rates [8]. For dependable data delivery, FANETs send the most significant data in disparate deployments that must be dispatched in a time-bound way. Hence, the networks' dependability remains excellent [9].

The compromised FANET-IoT devices (IoTD) in no way exhibit signs of being hacked and function as zombies for the botmaster (BM) when initiating the attacks [10]. A BN's dimensions may remain small, comprising hundreds of bots, while a bigger BN can have thousands of bots. A few bots will be present on the dark web very inexpensively, while enormous BNs have heavy costs [11].

There are two kinds of BNs: (i) BNs accepting commands and in consistent interaction with the BM within a client–server framework; (ii) peer-to-peer bots that communicate independently with one another and initiate the attacks after obtaining the BM's commands. BMs interact with bots by employing the aid of a command-and-control (CnC) server; the bots remain concealed until the BM gives commands. The concealed bots' conduct creates infested bots and a botnet attack (BA), which is an intricate job [12]. The BAs include the following: (i) scan commands employed in discovering the defenseless IoTD; (ii) ACK, SYN, UDP, and TCP flooding; (iii) combination attacks employed in starting a link and transferring the spam into this [13]. The current drawback in UAV-assisted FANETs is the effective detection of security threats. For that purpose, the feature selection and classification methods need improvement. The contributions of this study are described below:

- A new technique is proposed that utilizes the hybrid shark and bear smell optimization algorithm (HSBSOA) for FS and the deep neural classifiers to enhance the efficient and precise BN identification approach in FANETs;
- The aim of this study remains in identifying and classifying the implementation-specified threats, such as scan attacks, DDoS, TCP, UDP, and sync flooding, which are a few of the typical attacks.

The proposed hybrid HSBSOpt_DCA approach allows for more precise multiclass classification, including various types of attacks and non-attacks (NAs), and has shown encouraging results. The organization of remainder of this paper is as follows. Section 2 provides a state-of-the-art literature review. Section 3, the Materials and Methods, discusses the dataset used and the proposed methods. Section 4 provides a detailed analysis and the results. Section 5 concludes the article.

2. Related Works

In [14], Fried and Last proposed a novel and optimistic technique of employing wide-range and publicly accessible flight records for training in machine learning (ML) paradigms, which could identify anomalous flight designs and was proven to be a coherent counteractant for many ADS-B attacks. This novel technique varies from the formerly proffered methodologies, incorporating elementariness with the present ADS-B system. In [15], Mall et al. discussed unsupervised settings with sensors fixed in specific regions where the data can be gathered via mobile gadgets that remain attached to a UAV or drone. The authors initially modeled an appropriate framework and a lightweight convention for initiating safe transmission amongst the gadgets and the cloud through a portable drone. This convention also employs the physically unclonable function's (PUF) advantages for creation, which is employed to encrypt the messages in transmission. The familiar Scyther simulator is employed to stimulate the convention, and the outcomes show that this convention remains fully secured, preventing confidential data seepage.

In [16], Mairaj et al. attempted to learn the benefits of game-theoretic (GT) implementations for the avoidance of DDoSAs upon a drone emanating data out of standard game solutions, and optimized this with an encompassed authenticity concept named the quantal response equilibrium (QRE). The authors detected possible schemes for every player via simulations and devised five non-collaborative game scenarios for the DDoSAs' two versions. In such games, the conventional GT resolution or Nash equilibrium (NashE) gives data regarding the drone's suggested modes, the hacker's favored scheme, and the GT threshold (TH), presuming that the participants remain exceptionally brilliant.

In [17], Popoola et al. suggested the federated DL (FDL) methodology for zero-day BA identification to prevent data secrecy seepage in IoT-edge gadgets (IoTEG). This study utilizes an optimal deep neural network (ODNN) framework for NT classification. A model parameter server (MPS) distantly organizes the DNN paradigms' training in several IoTEGs when the federated averaging algorithm is employed to sum up the local paradigm updates. A global DNN paradigm is generated after many transmission rounds between the MPS and the IoTEG.

In [18], Hatzivasilis et al. introduced WARDOG, an awareness and digital forensic system, which notifies the end-user of the BN's contamination, reveals the BN framework, and catches confirmable data, which is then employed in a law court. The accountable administration system collects the data and automatically creates documentation for each instance. The document comprises authentic forensic data tracing entire engaged bodies and their parts in the attack.

In [19], Xi et al. proposed convolutional neural networks (CNNs) with a new deep learning framework that consists of dilated convolutional neural networks and recurrent neural networks. These stacked dilated convolutional networks perform effective feature selection, and the softmax classifier is used to recognize activities, which increases the accuracy of the classification performance. In [20], Alharbi and Alsubhi proposed a graph-based machine learning (ML) technique for botnet detection. For feature evaluation, filter-based theories are used, which exhibit robustness to zero-day attacks. This method achieved high precision, but its accuracy was moderate. In [21], Sung et al. presented a new methodology for discovering the malware in GCSs, which employed a fastText paradigm to generate low-size vectors when compared with the vectors from one-hot encoding (OhE) and a bidirectional LSTM paradigm for a comparison alongside sequential opcodes (SO). Furthermore, the API function names were employed to enhance the classification precision of the SO. In the experimentation, the Microsoft malware classification competency database was employed, and the family types classified the malware within the database. This proffered methodology exhibited an execution enhancement of 1.8%, correlating with the execution of the OhE-related technique.

In [22], Shitharth and Prasad proposed the supervisory control and data acquisition (SCADA) systems with the Markov chain clustering (MCC) technique, rapid probabilistic correlated optimization (RPCO) approach, and block-correlated neural network (BCNN) method to improve the accuracy of the network. However, it failed to reduce the cost-effectiveness of the process. Several studies have executed intrusion and malware identification processes. Nevertheless, there is a deficit of research discussing the problems concerning BN detection and feature extraction, magnitude reductions to repress counterfeit data, overfitting, and meticulous criteria calibration. Many research studies have employed actual BA databases in actual settings.

Furthermore, studies have analyzed ML paradigms for synthetic BN data devoid of apportions for feature engineering and an exhaustive overfitting analysis. Many studies have employed unbalanced live databases for learning and BN identification. The research studies chiefly concentrate on achieving greater precision, without discussing the constraints of greatly unbalanced databases or acquiring ostensive precision. In Table 1, a summary is provided with the limitations of the earlier research studies.

Table 1. Summary and limitations of some existing studies.

Ref.	Method Name	Outcome	Limitation	Advantage
[14]	Recurrent autoencoder classifier	Better classification rate	Quality predictions need large amount of data	Able to manage abundant amounts of data and input variables
[15]	Physically Unclonable Function *(PUF)*	Lower packet delivery ratio	Lots of labelled data are required for classification	Great capacity in predicting models
[16]	Quantal response equilibrium (QRE).	More throughput	Computational process is expensive during initialization	More flexible
[17]	Federated Averaging Algorithm	Less accuracy	Vanishing gradient problem is there while training network	More efficient
[18]	WARDOG	Less speed	Computationally expensive—data splitting is complicated and it maintains unbalanced database	Easy to deploy
[19]	Dilated Convolutional Neural Network	High accuracy	Computationally expensive process	Higher classification performance
[20]	Graph-based Machine learning for botnet detection	High precision	Accuracy is moderate, needs to be improved	Easy to deploy
[21]	Bidirectional LSTM	Less complexity	Takes long time to process large neural network	Appealing attributes of non-linear identification and control
[22]	RPCO-BCNN	High accuracy	Computationally expensive process	More flexible

3. Proposed HSBSOpt_DCA

UAV sets can be linked with one another to function as a relay to transfer the data out of a remote area (RA) network. Generally, the UAVs possess a mission for a surveillance operation and an operation to create a relay network for gathering data from RAs, such as in a desert or jungle. The UAVs' motility and versatility make it effortless to arrive at these RAs and give connectivity to the network. Nevertheless, with minor exertion, the attacker could effortlessly hijack the system. As a result, the deficit of a firm framework and the vulnerable wireless medium within FANETs make the nodes liable to attackers.

The N-BaIoT database comprises traffic data for pre-processing using the one-hot encoding method. The pre-processed data are then input in the feature selection step using the hybrid shark and bear smell optimization algorithm, after which the classification is performed using a dilated convolutional autoencoder. The proposed HSBSOpt_DCA (Figure 1) consists of several segments, including the dataset description, pre-processing employing OhE, FS employing HSBSOA, optimization initialization, odor absorption, frontward motion (FtM) toward the target, rotatory motion, updating the particle location, attaining the GS and LS, and classification employing DCAE.

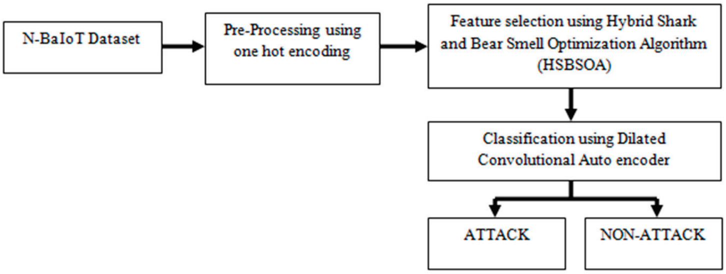

N-BaIoT Dataset
Pre-Processing using one hot encoding
Feature selection using Hybrid Shark and Bear Smell Optimization Algorithm (HSBSOA)
Classification using Dilated Convolutional Auto encoder
ATTACK
NON-ATTACK

Figure 1. Block schematic illustration for attack classification.

3.1. Dataset Description

The N-BaIoT database [23] comprises traffic data out of nine Industrial IoTD, whereby seven gadgets gather instances for eleven classes, and the other two gather data for six classes (Ennio_doorbell and Samsung_SNH_1011_N_Webcam). The data consist of harmless traffic and diverse malevolent attacks such as scan, TCP, UDP, and SYN attacks. There remains a sum of eighty-nine csv files in the current database's variant, having sum dimensions of 7.58 GB and 1,486,418 instances for ordinary and attack happenings. The 2 Bas—MIRAI and BASHLITE—have been classified into ten attack classes (AC) and NA. The AC includes:

- Scan commands for finding the defenseless IoTD;
- ACK, SYN, UDP, and TCP flooding;
- Combo or combination attacks employed to open a link and transmit the spam into this.

3.2. Pre-Processing Employing OhE

A categorical column (CC) is a column containing classes, where the cardinality remains minimum in nature. In the N-BaIoT database, four columns are detected as CCs, specifically 'Dir', 'Proto', 'sTos', and 'dTos'. The first column comprises seven classes, the second one comprises fifteen classes, the third one comprises six classes, and the fourth one comprises five classes. OhE indicates the procedure of transforming CCs into vectors of zeros and ones. A column with two and three classes has vector lengths of two and three, respectively. Transforming a five-class CC into a vector of zeros and ones with a length of five produces multicollinearity problems (MP).

The MP leads to unnecessary data and associated anticipators. The MP could be resolved by dropping a column's OhE classes. Thus, a column having five classes possesses a vector length of four rather than five. Relating to N-BaIoT, the OhE columns' quantity for four CCs would be twenty-nine columns currently. Every categorical feature (CF) exhibiting m feasible categorical values will be converted into a value in Rm employing a function *e*, which maps the feature's *j*th value into the m-dimensional vector's *j*th element.

$$e(xi) = (0, \ldots 1, \ldots, 0)\ if\ xi = j \tag{1}$$

The two arithmetical CFs will be scaled concerning every feature's average π and standard deviation β:

$$n(xi) = \frac{x1 - \pi}{\beta} \tag{2}$$

Pre-processing transforms NT into an observance sequence in which every observance will be portrayed as a feature vector (FV). The observances will be selectively labelled by their class as 'normal' or 'anomalous'. Such FVs will later be appropriate as inputs for data mining or ML algorithms.

3.3. FS Employing HSBSOA

The motivation behind the shark smell optimization (SSO) algorithm is the shark's capability and supremacy in capturing prey by employing a strong sense of smell (SoS) in a short time. A bear's olfactory bulb remains many times bigger than the rest of the beasts when its top job is to forward smell data from the nose toward the brain. In the bear smell optimization (BSO) methodology, the bear's SoS is exemplary in seeking foodstuffs at 1000 miles and beyond (known as the global solution (GS)) in optimization). As bears cannot see foodstuffs that far away, the statistical paradigm centered upon the SoS proposes a decisive manner for seeking such goals. By merging these two algorithms, a better fitness value (FtV) could be acquired for the FS procedure.

3.4. Initialization Procedure

The initial solution (IS) for the SSO algorithm's (SSOA) populace should be produced haphazardly inside the search space (SSp). Every IS portrays an odor particle (OP) that exhibits a feasible shark location at the start of the search procedure. The IS vector will be illustrated in Equations (3) and (4), accordingly to which $X_i^1 = i$th refers to the populace vector's starting location and NP = population size refers to the populace's dimensions:

$$X^1 = \left[x_1^1, x_2^1, \ldots x_{NP}^1\right] \tag{3}$$

The concerned optimization issue could be conveyed by:

$$x_i^1 = \left[x_{i,1}^1, x_{i,2}^1, \ldots x_{i,NP}^1\right] \tag{4}$$

where $x_{i,j}^1$ represents the jth size of the shark's ith location and ND represents the decision variables' numeral. By employing the BSO methodology, the bear's nose absorbs disparate smells; every one exhibits a location for movement, since all things possess a distinct odor in the ecosystem. Notice that several of these are called local solutions (LS). The desirable foodstuff's specific smell remains the final solution and is regarded as the GS. Consider $F_i = \left[fc_i^1, fc_i^2, \ldots fc_i^j, \ldots fc_i^k\right]$ being the ith obtained smell having k elements or particles, which is designed to solve the optimization issue $x_i^1 = \left[x_{i,1}^1, x_{i,2}^1, \ldots x_{i,NP}^1\right]$. As the bear obtains n smells during the breathing duration, the IS remains a matrix $FM = [fc_i^j]\ N * k$. Presently, as per the glomerular layer procedure and breathing action in a sniff sequence, DS_i^j indicates the jth smell element within ith. Centered upon statistical formulas, we obtain two conditions, which are $t\_\ inhale \leq t \leq t\_\ exhale$ and $t\_\ exhale \leq t$ with the presence of fairness, which includes the balanced energy to maintain the traffic in the transmission line:

$$DS_i^j = MG_i\ \ (t - t_{inhale}) + DS_i^{t\_\ inhale} + BE_i\ \ (t - t_{inhale}) \tag{5}$$

Equation (5) works for the condition $t\_\ inhale \leq t \leq t\_\ exhale$, where $t\_\ inhale$ represents the inhalation time (IT) and $BE_i\ (t - t_{inhale})$ denotes the balanced energy required during the inhalation process:

$$DS_i^j = DS_i^{t\_\ exhale} * BE_i^{t\_\ exhale}\ \exp\left(\frac{t\_\ exhale - t}{\varepsilon\ exhale}\right) \tag{6}$$

Equation (6) works for the condition $t\_\ exhale \leq tt\_\ inhale$, where $t\_\ exhale$ represents the exhalation time (ET) and $BE_i^{t\_\ exhale}$ denotes the balanced energy required during the process of exhalation. In the optimization procedure, the comprehensive duration of a breathing cycle remains identical to k or the ith smell's length, and as per the ET and IT the smell elements are split into 2 sets.

The total balanced energy is the summation of the energy required for the processes of vital energy (VE) and energy loss (EL) and is mathematically expressed below:

$$BE_{total} = BE_{vital} + BE_{loss} \tag{7}$$

where BE_{vital} denotes the dissipated energy during the process of inhalation and exhalation and BE_{loss} denotes the transmission loss that occurs.

3.5. Odor Absorption (OA)

For the process of odor absorption, mitral and granular parts are used to contain the receptor sensitivity, OA, as well as the input data, which are presented as $OB_{MG} = \left(OB_{MG}^1, OB_{MG}^2, \ldots, OB_{MG}^i, \ldots, OB_{MG}^N\right)$. Presently in this condition, $DS_i^j = 0$ exhibits that there is no smell in the olfactory epithelium prior to the subsequent inhalation. The non-negative array could be computed as:

$$OB_{MG}^i(F_i) = \frac{1}{k}\sum_{j=1}^{k} f\left(fc_i^j\right), f\left(fc_i^j\right) * S_{factor} \tag{8}$$

where k indicates the odor's extent in ith odor, while Equation (7) works for two conditions, which are the threshold values $V_t \leq fc_i^j$ and $V_t \geq fc_i^j$, where the arrays centered upon the odors data's represent the mean value. Here, S_{factor} denotes the satisfaction factor, whereby the mathematical expression for this factor is expressed as:

$$S_{factor} = W * \sum_{i=1}^{N} (1 - W) \tag{9}$$

where N denotes the total number of odor absorption mitral and W denotes the weight factor. The neural dynamics evolving out of the granular and mitral (GM) layers are calculated as:

$$\begin{gathered} X = -H_0\omega_y(Y) - \propto_x X + \sum L_0\omega_y\,(X) + DS + (E_{initial} - E_{least}) \\ Y = W_0\omega_x\,(x) - \propto_y Y + DS_c + +(E_{initial} - E_{least}) \end{gathered} \tag{10}$$

where $X = \{x_1, x_2, \ldots x_n\}$ and $Y = \{y_1, y_2, \ldots y_n\}$ represent the G-M cell (GMC) actions accordingly; $DS = \{ds_1, ds_2, \ldots ds_n\}$ and $DS_c = \{ds_{c1}, ds_{c2}, \ldots ds_{cn}\}$ represent the outward inputs to the mitral and middle of the granule cells, respectively; $E_{initial}$ denotes the initial energy and E_{least} denotes the lowest energy unit.

3.6. Frontward Motion (FtM) toward the Target

If the blood is discharged into the water, a shark possessing a velocity V goes towards the powerful OPs in every position to move nearer to the prey (target). Thus, the velocity within each size will be computed as:

$$v_{i,1}^k = \mu k.R1.\frac{\partial(OF)}{\partial xj} \tag{11}$$

where k = 1, 2, … $k_{max}\frac{\partial(OF)}{\partial xj}$, which would be the objective function (OF) at location $x_{i,1}^k$; k_{max} indicates the phases' maximal quantity for the forward motion of the shark, k indicates the phases' quantity, μk indicates a value within the interval [0, 1], and $R1$ is a haphazard number in the interval [0, 1]. The rise in the odor intensity decides the increase in the shark's velocity. Owing to inertia, the shark's acceleration remains a constraint. Thus, the present shark's velocity depends upon its former velocity, which can be utilized by altering (9), as exhibited in the following expression:

$$v_{i,1}^k = \mu k.R1.\frac{\partial(OF)}{\partial xj} + \propto k.R2v_{i,1}^{k-1} \tag{12}$$

where $\propto k$ portrays the inertia coefficient within the interval [0, 1], $v_{i,1}^{k-1}$ portrays the shark's former velocity, and $R2$, like $R1$, remains a haphazard number in the interval [0, 1]. Because of the shark's FtM, its novel location remains $Y_{i,1}^{k+1}$, which is decided depending upon its former location (x_i^k) and velocity (v_i^k). Hence, the shark's novel location can be described as:

$$Y_{i,1}^{k+1} = x_i^k + v_i^k.\Delta t_k \tag{13}$$

where Δt_ki denotes a time interval that can be presumed to be one for simplicity:

```
Pseudocode for frontward motion begins
Calculate velocity V
    Update the position of target prey
    Velocity of each shark ($v_{i,1}^{k}$)
      $v_{i,1}^{k} = \mu k.R1.\frac{\partial(OF)}{\partial xj}$
Find maximal quantity for forward motion
Release the odor and find its intensity
Update the shark's novel location
End
```

3.7. Rotatory Motion (RM)

The shark also possesses an RM that will be employed to discover powerful OPs. The SSOA procedure can be named the local search (LcS), which can be defined as:

$$Z_{i,1}^{k+1,m} = Y_i^{k+1} + R3.Y_{i,}^{k+1} \tag{14}$$

in which $m = 1,\ 2, \ldots,\ M$, and $R3$ denotes a haphazard number in the interval $[-1, 1]$. In the LcS, several points (M) will be linked to create closed contour lines and to design the shark's RM within the SSp.

3.8. Updating the Particle Location

The shark's search path will carry on with the RM, since this is nearer to the point of having a powerful SoS. This feature within the SSOA could be described by:

$$x_i^{k+1} = \mathrm{argmax}\left\{OF(Y_i^{k+1}), OF\left(Z_i^{k+1,i}\right), \ldots OF\left(Z_i^{k+1,M}\right)\right\} \tag{15}$$

in which x_i^{k+1} portrays the shark's subsequent location with the greatest OF value.

3.9. Attaining GS and LS

In the process of attaining GS and LS at the initial stage, two values are determined, which are $\omega_x\ (X)$ and $\omega_y\ (Y)$, and the expression for this is given below:

$$\omega_x\ (X) = \{f_x(x_1), f_x(x_2) \ldots f_x(x_n)\} \tag{16}$$

$$\omega_y\ (Y) = \{f_y(y_1), f_y(y_2) \ldots f_y(y_n)\} \tag{17}$$

The expressions $\omega_x\ (X) = \{f_x(x_1), f_x(x_2) \ldots f_x(x_n)\}$ and $\omega_y\ (Y) = \{f_y(y_1), f_y(y_2) \ldots f_y(y_n)\}$ indicate the GMC accordingly; $\propto_x$ and $\propto_y$ portray the GMC's time constants, and their values remain as 0.14; f_x and f_y simulate the cell output actions for the GMCs. Thus, we can obtain:

$$f_x(X) = \begin{cases} \propto_x + \propto_x \tanh\left(\frac{x-\varphi}{\propto_x}\right) \\ \propto_x + \propto_x \tanh\left(\frac{x-\varphi}{\propto_x}\right) \end{cases} \tag{18}$$

$$f_y(Y) = \begin{cases} \propto_y + \propto_y \tanh\left(\frac{x-\varphi}{\propto_y}\right) \\ \propto_y + \propto_y \tanh\left(\frac{x-\varphi}{\propto_y}\right) \end{cases} \tag{19}$$

In both Equations (18) and (19), the term φ represents the threshold value, and the values of $\propto_x$ and $\propto_y$ are 0.14 and 0.29, respectively. Here, the synaptic-strength connection matrices are calculated, which are represented as H_0, W_0, and L_0, which indicate the association between the GMCs and the mitral cells. This is computed as:

$$H_{0i}^{\ j} = \frac{\mathrm{rand}()}{T_h},\ \ W_{0i}^{\ j} = \frac{\mathrm{rand}()}{T_w},\ \ L_{0i}^{\ j} = \frac{\mathrm{rand}()}{T_l} \tag{20}$$

T_h, T_w, and T_l indicate the connection constants, rand() indicates a haphazard value, d_i^j indicates the space between the ith and gth odors based on their data, and the gth odor indicates the desirable odor for the bear; that is to say, this distance can be described between every odor (LS) and the intended odor (GS). This exhibits that the supervised operation centered upon the GS will be utilized while performing the optimization procedure to enhance the exploitation. As per the above-mentioned explanations, if the brain acquires all data from the neural action, the disjoining procedure is centered upon the discrepancy analysis. This procedure will be simulated while centered upon the Pearson correlation. Hence, this point assists the bear in choosing the finest manner for the subsequent location. The probability odor components (*POC*), probability odor fitness (*POF*), and odor fitness (*OF*) are described by:

$$POC_i = \frac{F}{\max(F_i)} * mid_{scale} \tag{21}$$

$$POF_i = \frac{OF_i}{\max(OF)} * mid_{scale} \tag{22}$$

where mid_{scale} denotes the lower and upper limits of the odor components. The mathematical expression for the calculation of mid_{scale} is described as:

$$mid_{scale} = \frac{\left({}^{OC_{ul}}/_{OC_{il}} \right) * OC_{il}}{2} \tag{23}$$

where OC_{ul} and OC_{il} are the lower and upper limits of the odor components, respectively. The discrepancy between 2 odors can be computed using the expected odor fitness (*EOF*) and distance odor component (*DOC*) formulas as:

$$DOC_i = 1 - \frac{\sum_{i=1}^{k}\left(POC_j^1 - POC_j^2\right)}{\sqrt{\sum_{i=1}^{k}\left(POC_j^1 - POC_j^2\right)^2}} * d(POC_i) \tag{24}$$

$$EOF_i = (POF_i - POF^g) * d(POF_i) \tag{25}$$

where g denotes the GS. The values of the odor fitness (*EOF*) and distance odor components (DOCs) are measured according to Equations (19) and (20), where the distances of the probability odor components (*POC*) and probability odor fitness (*POF*) are considered. The mathematical expressions for the calculation of $d(POC_i)$ and $d(POF_i)$ are given below:

$$d(POC_i) = \sqrt{\sum_{k=1}^{N}(x_i - y_i)} \tag{26}$$

$$d(POF_i) = \sqrt{\sum_{k=1}^{M}(x_j - y_j)} \tag{27}$$

where the distances between the source and destination coordinates are used for the calculation of the distances of *POC* and *POF*; x_i, x_j denotes the source coordinates and y_i, y_j denotes the destination coordinates.

These expressions denote the feasible manner shift. Indeed, these indices describe the association between the odors that have been reached at the desirable location. It is legibly exhibited that the brain's output determines an appropriate manner for the subsequent location. In the mesh grid region, the distance between entire odors can be centered upon 2 THs.

In this phase, the HSBSOA can be employed to extract the finest features. Initially, the shark and bear's beginning locations will be located to be in the middle of the data. Next, the fitness or finesse is noted for every position surrounding the shark and bear by employing the fitness function. Then, the HSBSOA will be implemented to extract the finest features. In this study we extracted twenty-one features via the HSOSOA out of every datapoint by implementing twenty-one repetitions. Every repetition possesses just one feature extracted with the greatest FtV. While performing each repetition, the shark and bear's positions will be updated to be frontward or rotatory-centered upon the FtV. When the position's FtV in the shark position's FM remains above the shark RM's FtV, the shark's location will be updated. The shark's trajectory will move frontward or rotatory depending upon the position's FtV; additionally, the positions that will be viewed using the HSBSOA can be reviewed.

```
Pseudocode: HSBSOA Algorithm
Begin: Initialize search space
          Indicate the total number of populations
          Compute the optimization issue
              $x_i^1 = \left[x_{i,1}^1, x_{i,2}^1, \ldots x_{i,NP}^1\right]$
          Compute decision variables numeral
           Compute local solution (LS) from decision variable
              Update the inhale and exhale parameter
              Update exhalation time (ET), inhalation time (IT)
          Initiate Odor absorption
              $MG = \{MG_1, MG_2, \ldots MG_i, \ldots MG_n\}$
          Compute non-negative array $MG_i(O_i)$
           Compute granular and mitral (G-M) layers
          Initiate Frontward motionCompute velocity V for each shark
              Update $k_{max}$ for all location
          Find shark's acceleration
          Initiate Rotatory motion
          Compute local search (LcS)
              Updating the particle location
          Compute probability odor components
          Compute probability odor fitness ($POF$)
          Find the fitness parameter
End
```

3.10. Classification Employing DCAE

Before introducing DCAE, for detailed comprehension, it remains notable that the notation 'dilated convolution' (DC) portrays a convolution procedure with a dilated filter (DlF). Generally, the DC is implemented in the wavelet decomposition discipline. As the DC operant solely employs a similar filter at disparate scales having disparate dilation factors (DtF), its application in no way encompasses the DlF's formation. In addition, the dilated convolutional network can extend the receptive field (RF) dimension, which depends upon enhancing the DtF instead of expanding the network's field map (FMp) dimensions.

The layers involved in the process of the ACAE framework are the input layer, convolutional layer, DC layer, flatten and reshape layer, recurrent layer, and then finally the output layer as shown in Figure 2. The dilated convolutional layer is incorporated with a filter size of (3, 3) and with a dilation size of (1, 2, 4). In order to process the dilation in the mathematical order, the discrete function is given as $D_c = {}^\circ\text{F} \rightarrow S$, while the size of the discrete filter is mentioned as $\frac{(2r+1)\times(2r+1)}{(2r-1)}$. The math expression for the calculation of the DC operator © is given below:

$$(F©k)_{(x,y)} = \sum_{g=1}^{r}\sum_{h=1}^{r} F(X,Y) * (v_{ci}^{i}(X,Y)) * (L_{ce}(X,Y)) * (k(g-h)) \quad (28)$$

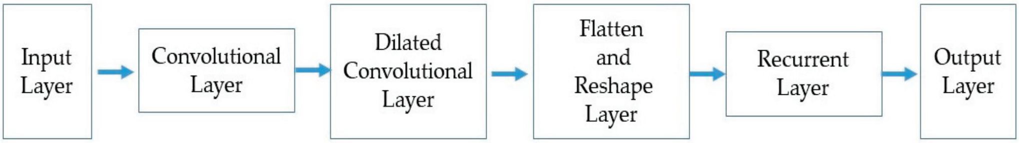


Figure 2. ACAE framework.

In Equation (28), the term X represents $(x-g)$, Y represents $(y-h)$, and $k:\rho_r \rightarrow R$, which is the discrete filter with a size of $\frac{(2r+1)\times(2r+1)}{(2r-1)}$. Here, $v_{ci}^{i}(X,Y)$ denotes the corresponding integer index value, which lies between (0 to 5) and $L_{ce}(X,Y)$, denoted as the entropy loss calculation, which lies between 0 and 10.

Secondly, an improved dilation convolution is developed with the variants X_I and Y_I. The math expression for the calculation of the improved DC operator $©_I$ is given below.

$$(F©_I k)_{(x,y)} = \sum_{g=1}^{r}\sum_{h=1}^{r} F(X_I\,,Y_I) * (v_{ci}^{i}(X_I,Y_I)) * (L_{ce}(X_I,Y_I)) * (k(g-h)) \tag{29}$$

Thus, the convolutions © and $©_I$ are called one-DC. Here, we presume that $F_0, F_1, \ldots F_{n-1} : {}^{\circ}\mathrm{F}^2 \rightarrow S$ for the remaining DFs and $k_0, k_1, \ldots k_{n-2} : \rho_1 \rightarrow R$ for the remaining 3 × 3 DsFs. Furthermore, the filters are implemented by aggressively enhancing DtFs such as $2^0, 2^1, \ldots 2^{n-2}$. Next, the DF ${}^{\circ}\mathrm{F}_{i+1}$ could be conveyed as:

$${}^{\circ}\mathrm{F}_{i+1} = \alpha{}^{\circ}\mathrm{F}_i \times \beta k_i \text{ for } i = 0,1,2,\ldots g-2 \tag{30}$$

Similarly:

$${}^{\circ}\mathrm{F}_{j+1} = \alpha{}^{\circ}\mathrm{F}_j \times \beta k_j \text{ for } j = 0,1,2,\ldots h-2 \tag{31}$$

As per the RF description, two sections are present for every component, which are ${}^{\circ}\mathrm{F}_{i+1}$ and ${}^{\circ}\mathrm{F}_{j+1}$. The terms α and β are the constant values that are used for experimental purposes and which satisfy the condition ($\alpha + \beta = 1$). The math expression for the combined detection methodology is given below:

$$M^{\circ}\mathrm{F} = ({}^{\circ}\mathrm{F}_{i+1}) \times ({}^{\circ}\mathrm{F}_{j+1}) = \left(\left(2^{i+2}-1\right) * \left(2^{i+2}-1\right)\right) \times \left(\left(2^{j+2}-1\right) * \left(2^{j+2}-1\right)\right) \tag{32}$$

Thus, RF remains a square of aggressively enhanced dimensions. In the convolutional layers (CvLs), the former layer's FMs will be convolved with multiple convolutional kernels (CKs), especially FMp. Next, the independent layer's outcomes added with a bias will be supplied to an activation function (AF) to create an FM. Presuming that $v_{i,j}^{x,d}$ remains a value at the xth row for channel d within the jth FM of the ith layer, the value of $v_{i,j}^{x,d}$ could be acquired as:

$$v_i^{x,d} = \tan^{\circ}\mathrm{B}_1(b_i + \textstyle\sum_g \displaystyle\sum_{p=1}^{p_{i-1}} \omega_{ig}^{p} * (v_{(i-1)g}^{x+p,d}) * (of_{(i-1)g}^{x+p,d}))\;\; d = 1,2,3\ldots D \tag{33}$$

$$v_j^{x,d} = \tan^{\circ}\mathrm{B}_2(b_j + \textstyle\sum_h \displaystyle\sum_{p=1}^{p_{j-1}} \omega_{jh}^{p} * (v_{(j-1)h}^{x+p,d}) * (of_{(j-1)h}^{x+p,d}))\;\; d = 1,2,3\ldots D \tag{34}$$

where $\tanh(\cdot)$ refers to a hyperbolic tangent function for $v_i^{x,d}$ and $v_j^{x,d}$; specifically, b_i and b_j are the biases for the FM (i, j), g refers to the present FM linked to the $(i-1)$th layer, and ω_{ig}^{p} and ω_{jh}^{p} refer to a value at location p within CK to which the dimensions are p_i and p_j, while the terms $of_i^{x,d}$ and $of_j^{x,d}$ are the objective functions.

For the initial block, every CvL layer will be incorporated by (1) a CL that convolves its inputs with an array of kernels to be learnt in the training stage, (2) a rectified linear unit (ReLU) layer that maps convolved outcomes by the function $relu(v) = max(v,0)$;, and (3) a normalization layer that normalizes values of disparate FMs in the former layer. The math expression for v_i and v_j is given below.

$$v_i = v_{(i-1)}(k+\alpha) \sum_{t\epsilon G(i)} v^2(i-1)t \tag{35}$$

$$v_j = v_{(j-1)}(k+\beta) \sum_{t\epsilon G(j)} v^2(j-1)t \tag{36}$$

In Equations (35) and (36), the terms k, α, and β remain the hyper-criteria, and G(i) and G(j) remain the FMs' array-incorporated terms during normalization. The ensuing 3 layers remain DC layers, having disparate dilated factors. For example, in this study we consecutively selected one, two, and four.

For the next block, centered upon the former exposure, the depth of a minimum of 2 recurrent layers remains advantageous for processing the concatenative data. This study utilizes a 2-layer stacked LSTM. Moreover, a ReLU will be used as the AF. The dropout layer is implemented in the LSTM layer's input for regularization. Furthermore, recurrent batch normalization is employed to lessen the internal covariance shift amidst the time phases.

The next block remains a completely linked network layer. This remains akin to a conventional multilayer perceptron neural network (NN), which maps the latent features into the output classes (*OC*). In this layer, the softmax function is described below:

$$v_{i,j} = \frac{\exp(v_{(i-1)j})}{\sum_{j=1}^{c} \exp(v_{(i-1)j})} \tag{37}$$

Next, an entropy cost function will be incorporated, centered upon the probabilistic outcomes and the training instances' actual labels. In the course of the training stage, all the criteria will be modified to search for the minimal cost. Additionally, a sliding window (SW) scheme will be utilized to segment the time sequence signal into signals' small pieces. In particular, an instance employed by the CNN remains a 2D matrix comprising r unprocessed samples (with every sample having D features). In this way, r will be selected to remain as the sampling rate or the finite duration, and the SW's phase dimension will be selected to retain a fifty percent overlap between the nearby windows. Hence, the shorter phase dimension remains the instances' bigger quantity that experiences greater calculative workloads. Furthermore, the signals' small portion will be generally very frequently labelled.

```
Pseudocode: Proposed Approach
Begin
    five classes = CC
    categorical feature (CF) = $R^m$, e
                    $e(xi) = (0, \ldots 1, \ldots, 0)\ if\ xi = j$
                    Compute average $\pi$
                    Compute standard deviation $\beta$
     Find $n(xi) = \frac{x1-\pi}{\beta}$
     Check the shark's capability
          Capture the prey
              EmploySoS
              Initiate the smelling process
              Achieve global solution
              Find the fitness value
     Indicate the total number of population
     Compute the optimization issue
        $x_i^1 = \left[x_{i,1}^1, x_{i,2}^1, \ldots x_{i,NP}^1\right]$
     Compute decision variable numeral
     Compute local solution (LS) from decision variable
        Update the inhale and exhale parameter
        Update exhalation time (ET), inhalation time (IT)
```

```
          Initiate Odor absorption
              $MG = \{MG_1, MG_2, \ldots MG_i, \ \ldots MG_n\}$
          Compute Compute non-negative array $MG_i(O_i)$
          Compute granular and mitral (G-M) layers
              Calculate velocity V
                Update the position of target prey
                Velocity of each shark ($v^k_{i,1}$)
                                 $v^k_{i,1} = \mu k.R1.\frac{\partial(OF)}{\partial xj}$
              Find maximal quantity for forwarding motion
              Release the odor and find its intensity
              Update the shark's novel location
          Initiate Frontward motion
              Compute velocity V for each shark
              Update $k_{max}$ for all locations
              Find shark's acceleration
          Initiate Rotatory motion
          Compute local search (LcS)
              Update the particle location
          Compute probability odor components
          Compute probability odor fitness ($POF$)
          Find the fitness parameter
     Stop
```

4. Performance Analysis

The dilated convolutional classifier-based botnet detection method (HSBSOpt_DCA) is implemented in Python 3.7 using the Ubuntu 16.04 operating system with 8 GB of RAM. The database chosen for the feature selection is the N-BaIoT database, which includes the traffic data for nine industrial IoTD. Seven databases are the gadgets' gathered instances for eleven classes, and two are the gadgets' gathered data for six classes (Ennio_doorbell and Samsung_SNH_1011_N_Webcam). The experimental outcome will be assessed by measuring the performance matrices, such as the accuracy, precision, recall, and F1-score. Such criteria will be correlated with four advanced methodologies: CNN-related SS for DD and its identification (CNN-SSDI), the bidirectional LSTM model (BI_LSTM), ODNN, and RPCO_BCNN with the proffered HSBSOpt_DCA.

4.1. Performance Matrices

- Accuracy: This provides the capability for comprehensive anticipation generated by the paradigm. The true positive (*TP*) and true negative (*TN*) give the ability to anticipate the intrusion's existence or non-existence. The false positive (*FP*) and false negative (FN) provide the false anticipation given by the employed paradigm. The mathematical expression for the calculation of the accuracy is described as [15]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{38}$$

- Precision: Precision is defined as the positive output achieved by the algorithm used in the proposed model, which lies in the range of (0 to 1). It computes the intrusion classification paradigm's victory. It defines the classifier's probability for anticipating the outcome as positive if the intrusion exists. It is as called the *TP* rate. It can be measured as:

$$\text{Precision}(P) = \frac{TP}{TP + FP} \tag{39}$$

- Recall: This is the classifier's probability of anticipating the outcome as negative if the intrusion does not exist. It is also known as the *TN* rate, as mentioned below:

$$\text{Recall}(R) = \frac{TP}{TP + FN} \tag{40}$$

- F1-Score: This is used to measure the anticipation execution. It is defined as the weighted mean calculation of the precision and recall. The F1-score lies between 0 and 1. If the score is 1, it is considered the most acceptable value; if it is 0, it is regarded as weak. The mathematical expression for the calculation of the F1-score [15] is given below:

$$F1\text{-Score} = \frac{2 * P * R}{P + R} \tag{41}$$

4.2. Results and Discussion

In this section, the metrics such as the accuracy, precision, recall, and F1-score are measured with respect to 50 and 100 epochs. Each metric calculation on the various epochs is evaluated. The accuracy calculations with variable epochs numbering 50 and 100 are demonstrated in Figures 3 and 4.

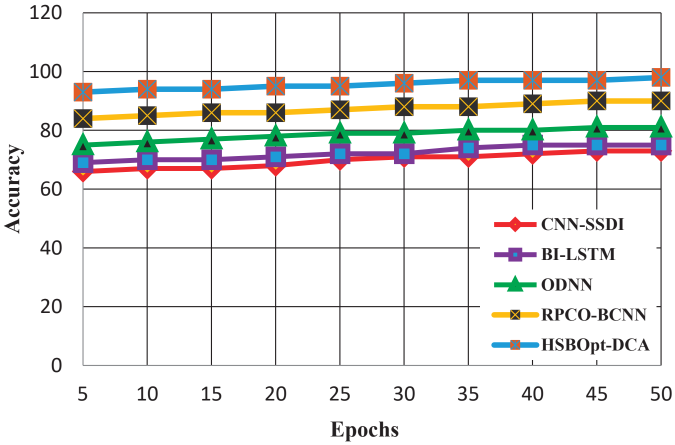


Figure 3. Accuracy calculation with 50 epochs.

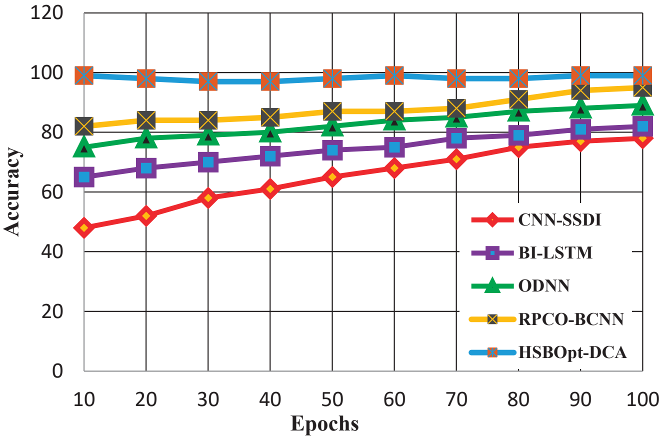


Figure 4. Accuracy calculation with 100 epochs.

Figure 3 shows the accuracy calculation for methods such as the CNN-SSDI, BI_LSTM, ODNN, RPCO_BCNN, and HSBSOpt_DCA. It can be understood from Figure 3 that the proposed HSBSOpt_DCA produces better accuracy when compared with other methods with respect to the 50 epochs. Various levels of accuracy are achieved by the CNN-SSDI (73%), BI_LSTM (75%), ODNN (81%), RPCO_BCNN (90%), and HSBSOpt_DCA (98%) methods. The accuracy achieved by the proffered HSBSOpt_DCA method is high and is achieved by using the hybrid optimization and dilated convolution process.

Figure 4 shows the accuracy calculation for methods such as CNN-SSDI, BI_LSTM, ODNN, RPCO_BCNN, and HSBSOpt_DCA. The figure proves that the proffered HSBSOpt_DCA method produces better accuracy than the other methods for 100 epochs. The accuracy scores achieved by the methods vary for CNN-SSDI (78%), BI_LSTM (82%), ODNN (89%), RPCO_BCNN (95%), and HSBSOpt_DCA (99%). The accuracy achieved by the proffered HSBSOpt_DCA method is high using the hybrid shark and bear smell optimization algorithm.

The precision calculations with 50 and 100 epochs are demonstrated in Figures 5 and 6. Figure 5 shows the precision calculation for methods such as CNN-SSDI, BI_LSTM, ODNN, RPCO_BCNN, and HSBSOpt_DCA. The figure proves that the proffered HSBSOpt_DCA method produces better precision when compared with the other methods for 50 epochs. The precision scores achieved by the methods vary for CNN-SSDI (58%), BI_LSTM (69%), ODNN (75%), RPCO_BCNN (93%), and HSBSOpt_DCA (99%). The precision achieved by the proffered HSBSOpt_DCA method is high using the hybrid shark and bear smell optimization algorithm.

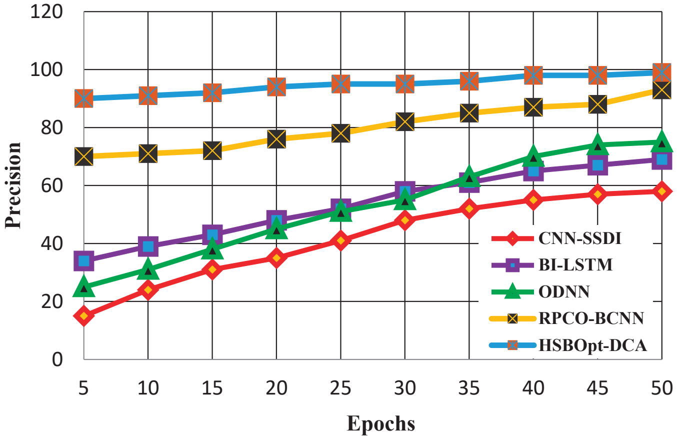


Figure 5. Precision calculation with 50 epochs.

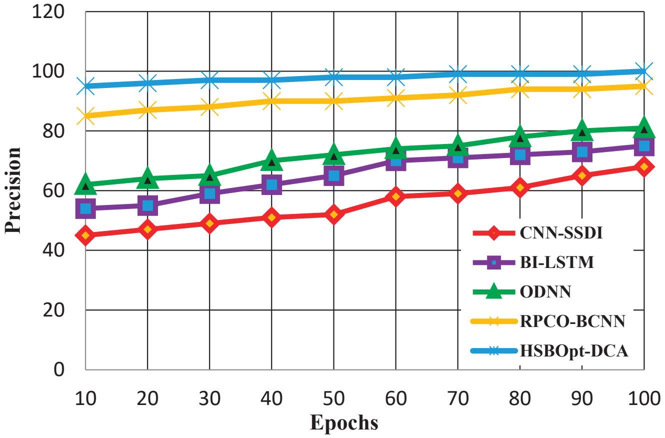


Figure 6. Precision calculation with 100 epochs.

Figure 6 shows the precision calculation for methods such as CNN-SSDI, BI_LSTM, ODNN, RPCO_BCNN, and HSBSOpt_DCA. The figure proves that the proffered HSBSOpt_DCA method produces better precision when compared with the other methods for 100 epochs. The precision scores achieved by the methods vary for CNN-SSDI (68%), BI_LSTM (75%), ODNN (81%), RPCO_BCNN (95%), and HSBSOpt_DCA (99.9%). The precision achieved by the proffered HSBSOpt_DCA method is high using the hybrid shark and bear smell optimization algorithm.

The recall calculations with 50 and 100 epochs are demonstrated in Figures 7 and 8. Figure 7 shows the recall calculation for methods such as CNN-SSDI, BI_LSTM, ODNN, RPCO_BCNN, and HSBSOpt_DCA. The figure proves that the proffered HSBSOpt_DCA method produces better recall than the other methods for 50 epochs. The recall scores achieved by the methods vary for CNN-SSDI (85%), BI_LSTM (81%), ODNN (85%), RPCO_BCNN (85%), and HSBSOpt_DCA (91%). The recall achieved by the proffered HSBSOpt_DCA method is high and is achieved by using the hybrid optimization and dilated convolution process.

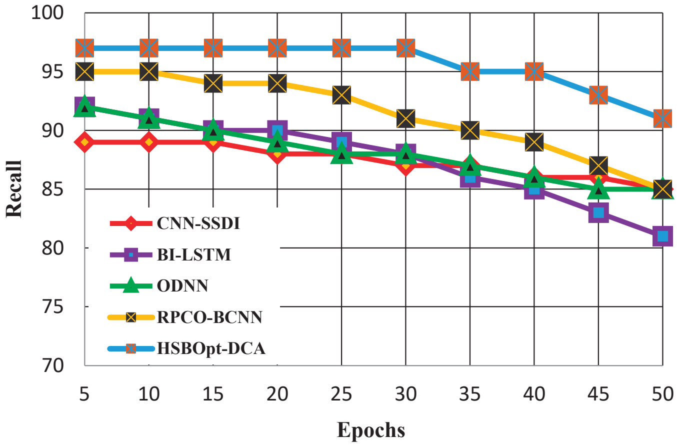


Figure 7. Recall calculation with 50 epochs.

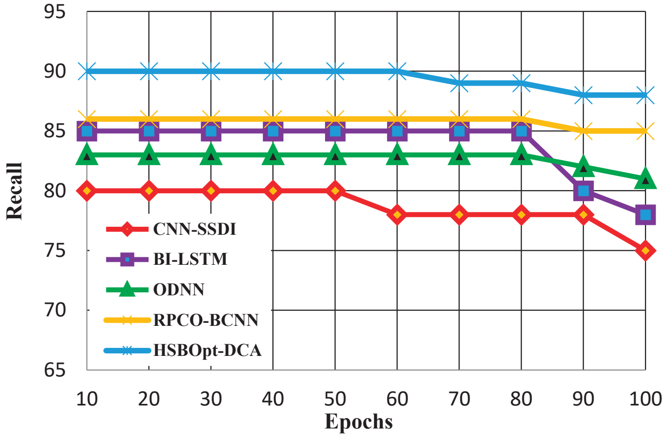


Figure 8. Recall calculation with 100 epochs.

Figure 8 shows the recall calculation for methods such as CNN-SSDI, BI_LSTM, ODNN, RPCO_BCNN, and HSBSOpt_DCA. The figure proves that the proffered HSBSOpt_DCA method produces better recall than the other methods for 100 epochs. The recall scores achieved by the methods vary for CNN-SSDI (75%), BI_LSTM (78%), ODNN (81%), RPCO_BCNN (85%), and HSBSOpt_DCA (88%). The recall achieved by the proffered HSBSOpt_DCA method is high and is achieved by using the improved dilated convolution process.

The F1-score evaluations for 50 and 100 epochs are demonstrated in Figures 9 and 10. Figure 9 shows the calculation of the F1-scores for the proposed and existing methods. The figure proves that the proffered HSBSOpt_DCA method produces a better F1-score than the other methods for 50 epochs. The F1-scores achieved by the methods vary for CNN-SSDI (73%), BI_LSTM (75%), ODNN (81%), RPCO_BCNN (94%), and HSBSOpt_DCA (98%). The F1-score achieved by the proffered HSBSOpt_DCA method is high and is achieved by using the improved dilated convolution process.

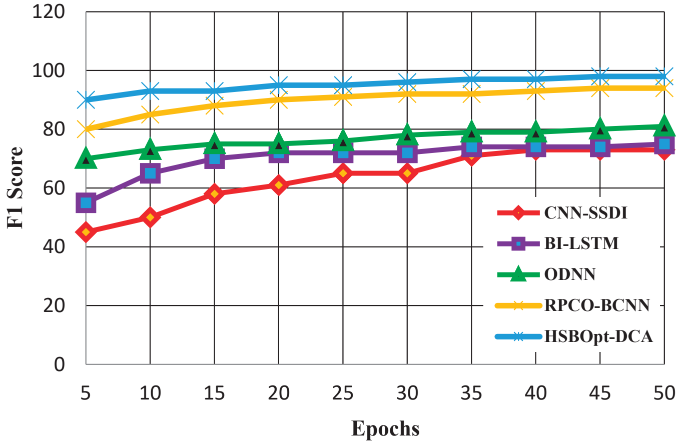


Figure 9. F1-score calculation with 50 epochs.

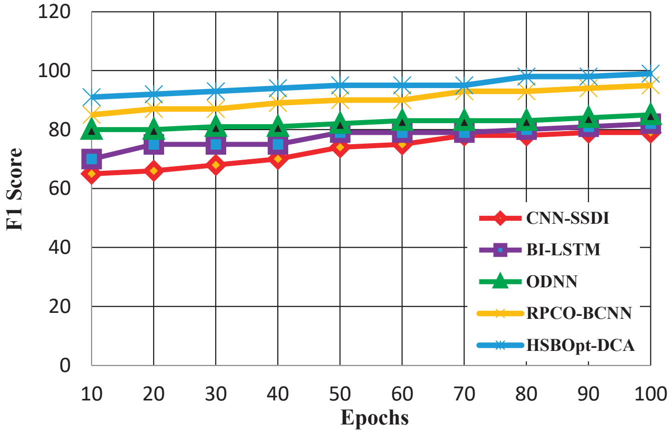


Figure 10. F1-score calculation with 100 epochs.

Figure 10 shows the calculation of the F1-scores for the proposed and existing methods. The figure proves that the proffered HSBSOpt_DCA method produces a better F1-score compared with other methods for 100 epochs. The F1-scores achieved by the methods vary for CNN-SSDI (79%), BI_LSTM (82%), ODNN (85%), RPCO_BCNN (95%), and HSBSOpt_DCA (99%). The F1-score achieved by the proffered HSBSOpt_DCA method is high and is achieved by using the improved dilated convolution process. Therefore, it is evident from the experiments that the proposed approach outperforms other existing methods, and it can be concluded that the feature extraction using the optimization algorithms definitely increases the performance of the classification model; therefore, the model can be used to detect and classify security threats in FANET.

5. Conclusions

In this study, we proposed an effective model combining hybrid shark and bear smell optimization (HSBSOA) to secure the FANET from security threats. It provides a solution to investigate the FANET botnet detection threat and to solve the combinational optimization problem. Then, a dilated convolution autoencoder classifier is employed to detect and classify the security threats in the network. The parameters considered for the performance analysis of the proffered HSBOpt_DCA are the accuracy, precision, recall, and F1-score. Moreover, the performance of the proposed approach was compared with CNN-SSDI, bi_LSTM, ODNN, and RPCO-BCNN. The performance of the proposed HSBOpt_DCA network was evaluated with different epochs. The proposed model with 50 epochs achieved 98% accuracy, 99% precision, 91% recall, and a 98% F1-score. For 100 epochs, it achieved 99% accuracy, 99.9% precision, 88% recall, and a 99% F1-score. The comparison showed that the proposed HSBOpt_DCA achieved 33% better accuracy, 30% better precision, 13% better recall, and a 20% better F1-score than the existing methods. The proposed method provides a global security solution to the security issues in the UAV-FANET framework. The proposed hybrid-optimization-based feature selection process reduced the computational time. It achieved higher accuracy, precision, recall, and F1-scores than the existing approaches. However, the classification tasks still require improvement, which can be considered in the future.

Author Contributions: Conceptualization, N.F.A., F.A., H.M.A.G., S.K. and A.A.; methodology, N.F.A. and A.A.; software, N.F.A. and A.H.A.; validation, A.H.A., A.S.A. and A.A.; formal analysis, N.F.A.; investigation, A.S.A.; resources, A.H.A.; data curation, M.H.H. and F.H.A.; writing—original draft preparation, A.A. and A.H.A.; writing—review and editing, S.K.; supervision, A.H.A. and S.K. All authors have read and agreed to the published version of the manuscript.

Funding: The research received no external funding.

Data Availability Statement: Not Applicable.

Acknowledgments: This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Government Order FENU-2020-0022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gupta, S.; Sharma, N.; Rathi, R.; Gupta, D. *Dual Detection Procedure to Secure Flying Ad Hoc Networks: A Trust-Based Framework*; Springer: Singapore, 2021; Volume 210, pp. 83–95. [CrossRef]
2. Jasim, K.S.; Alheeti, K.M.A.; Alaloosy, A.K.A.N. A Review Paper on Secure Communications in FANET. In Proceedings of the 2021 International Conference of Modern Trends in Information and Communication Technology Industry (MTICTI), Sana'a, Yemen, 4–6 December 2021; 146, pp. 1–7. [CrossRef]
3. Bekmezci, İ.; Şentürk, E.; Türker, T. Security issues in flying ad-hoc networks (FANETS). *J. Aeronaut. Space Technol.* **2016**, *9*, 13–21.
4. Dadi, S.; Abid, M. *Enhanced Intrusion Detection System Based on AutoEncoder Network and Support Vector Machine*; Springer: Singapore, 2021; Volume 466, pp. 327–341. [CrossRef]
5. Rodrigues, M.; Pigatto, F.D.; Fontes, J.V.C.; Pinto, A.S.R.; Diguet, J.-P.; Branco, C.K. UAV Integration IntoIoIT: Opportunities and Challenges. In Proceedings of the 13th International Conference on Autonomic and Autonomous Systems (ICAS 2017), Barcelona, Spain, 21–25 May 2017; p. 95.
6. Bekmezci, I.; Sahingoz, O.K.; Temel, Ş. Flying Ad-Hoc Networks (FANETs): A survey. *Ad Hoc Netw.* **2013**, *11*, 1254–1270. [CrossRef]
7. Sang, Q.; Wu, H.; Xing, L.; Xie, P. Review and Comparison of Emerging Routing Protocols in Flying Ad Hoc Networks. *Symmetry* **2020**, *12*, 971. [CrossRef]
8. Hussain, A. A Hybrid and Robust Delay and Link Stability Aware (DLSA) Routing Protocol for Unmanned Aerial Ad-Hoc Networks (UAANETs). *Res. Sq.* 2021; *Preprint (Version 1)*. [CrossRef]
9. Zafar, W.; Khan, B.M. A reliable, delay bounded and less complex communication protocol for multicluster FANETs. *Digit. Commun. Netw.* **2017**, *3*, 30–38. [CrossRef]
10. Walia, E.; Bhatia, V.; Kaur, G. Detection Of Malicious Nodes in Flying Ad-HOC Networks (FANET). *Int. J. Electron. Commun. Eng.* **2018**, *5*, 6–12. [CrossRef]
11. Yanmaz, E.; Costanzo, C.; Bettstetter, C.; Elmenreich, W. A discrete stochastic process for coverage analysis of autonomous UAV networks. In Proceedings of the 2010 IEEE Globecom Workshops, Miami, FL, USA, 6–10 December 2010; 40, pp. 1777–1782. [CrossRef]
12. Ahamed S, M.J.; Krishnamoorthy, J. Cyber threats based on botnet and its detection mechanisms. In Proceedings of the 8th Annual International Research Conference, Oluvil, Sri Lanka, 25 November 2019.
13. Verma, S.; Sharma, N.; Singh, A.; Alharbi, A.; Alosaimi, W.; Alyami, H.; Gupta, D.; Goyal, N. DNNBoT: Deep Neural Network-Based Botnet Detection and Classification. *Comput. Mater. Contin.* **2022**, *71*, 1729–1750. [CrossRef]
14. Fried, A.; Last, M. Facing airborne attacks on ADS-B data with autoencoders. *Comput. Secur.* **2021**, *109*, 102405. [CrossRef]
15. Mall, P.; Amin, R.; Obaidat, M.S.; Hsiao, K.-F. CoMSeC++: PUF-based secured light-weight mutual authentication protocol for Drone-enabled WSN. *Comput. Netw.* **2021**, *199*, 108476. [CrossRef]
16. Mairaj, A.; Javaid, A.Y. Game theoretic solution for an Unmanned Aerial Vehicle network host under DDoS attack. *Comput. Netw.* **2022**, *211*, 108962. [CrossRef]
17. Popoola, S.I.; Ande, R.; Adebisi, B.; Gui, G.; Hammoudeh, M.; Jogunola, O. Federated Deep Learning for Zero-Day Botnet Attack Detection in IoT-Edge Devices. *IEEE Internet Things J.* **2021**, *9*, 3930–3944. [CrossRef]

18. Hatzivasilis, G.; Soultatos, O.; Chatziadam, P.; Fysarakis, K.; Askoxylakis, I.; Ioannidis, S.; Alexandris, G.; Katos, V.; Spanoudakis, G. WARDOG: Awareness Detection Watchdog for Botnet Infection on the Host Device. *IEEE Trans. Sustain. Comput.* **2019**, *6*, 4–18. [CrossRef]
19. Xi, R.; Hou, M.; Fu, M.; Qu, H.; Liu, D. Deep Dilated Convolution on Multimodality Time Series for Human Activity Recognition. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–8. [CrossRef]
20. Alharbi, A.; Alsubhi, K. Botnet Detection Approach Using Graph-Based Machine Learning. *IEEE Access* **2021**, *9*, 99166–99180. [CrossRef]
21. Sung, Y.; Jang, S.; Jeong, Y.-S.; Park, J.H. Malware classification algorithm using advanced Word2vec-based Bi-LSTM for ground control stations. *Comput. Commun.* **2020**, *153*, 342–348. [CrossRef]
22. Shitharth, S.; Prasad, K.M.; Sangeetha, K.; Kshirsagar, P.R.; Babu, T.S.; Alhelou, H.H. An Enriched RPCO-BCNN Mechanisms for Attack Detection and Classification in SCADA Systems. *IEEE Access* **2021**, *9*, 156297–156312. [CrossRef]
23. Dua, D.; Graff, C. UCI Machine Learning Repository. 2019. Available online: https://archive.ics.uci.edu/ml/datasets/detection_of_IoT_botnet_attacks_N_BaIoT (accessed on 5 June 2022).

big data and cognitive computing

Article

Proposal of Decentralized P2P Service Model for Transfer between Blockchain-Based Heterogeneous Cryptocurrencies and CBDCs

Keundug Park [1] and Heung-Youl Youm [2,*]

[1] AI&Blockchain Research Center, Seoul University of Foreign Studies, Seoul 60745, Republic of Korea
[2] Department of Information Security Engineering, Soonchunhyang University, Asan 31538, Republic of Korea
* Correspondence: hyyoum@sch.ac.kr

Abstract: This paper proposes a solution to the transfer problem between blockchain-based heterogeneous cryptocurrencies and CBDCs, with research derived from an analysis of the existing literature. Interoperability between heterogeneous blockchains has been an obstacle to service diversity and user convenience. Many types of cryptocurrencies are currently trading on the market, and many countries are researching and testing central bank digital currencies (CBDCs). In this paper, existing interoperability studies and solutions between heterogeneous blockchains and differences from the proposed service model are described. To enhance digital financial services and improve user convenience, transfer between heterogeneous cryptocurrencies, transfer between heterogeneous CBDCs, and transfer between cryptocurrency and CBDC should be required. This paper proposes an interoperable architecture between heterogeneous blockchains, and a decentralized peer-to-peer (P2P) service model based on the interoperable architecture for transferring between blockchain-based heterogeneous cryptocurrencies and CBDCs. Security threats to the proposed service model are identified and security requirements to prevent the identified security threats are specified. The mentioned security threats and security requirements should be considered when implementing the proposed service model.

Keywords: blockchain; cryptocurrency; central bank digital currency; virtual asset; transfer; payment; blockchain interoperability; decentralized finance

Citation: Park, K.; Youm, H.-Y. Proposal of Decentralized P2P Service Model for Transfer between Blockchain-Based Heterogeneous Cryptocurrencies and CBDCs. *Big Data Cogn. Comput.* **2022**, *6*, 159. https://doi.org/10.3390/bdcc6040159

Academic Editors: Peter R.J. Trim, Yang-Im Lee and Min Chen

Received: 7 November 2022
Accepted: 15 December 2022
Published: 19 December 2022

1. Introduction

About ten thousand cryptocurrencies are being traded on cryptocurrency exchanges [1], and about one hundred countries are exploring central bank digital currencies (CBDCs) in one form or another. For example, some countries are researching, some are testing, and some have already distributed CBDCs to the public [2–4]. To enhance digital financial services and improve user convenience, transfer between heterogeneous cryptocurrencies, transfer between heterogeneous CBDCs, and further transfer between cryptocurrency and CBDC should be required.

However, due to the lack of interoperability between heterogeneous blockchains, there is a problem related to the transfer between blockchain-based heterogeneous cryptocurrencies (e.g., Bitcoin [5], Ether [6], etc.) and CBDCs (e.g., US CBDC, UK CDBC, Korean CBDC, Chinese CBDC, etc.). For example, it is difficult to transfer between a Bitcoin wallet and an Ether wallet, between a US CBDC wallet and a Korean CBDC wallet, or between a Bitcoin wallet and a US CBDC wallet. Existing studies to address the lack of interoperability between heterogeneous blockchains have progressed towards centralized architectures where intermediaries handle ledger data sharing between blockchains. The sharing of ledger data that records the transaction history of cryptocurrencies and CBDCs is an essential operation for the interoperability between heterogeneous blockchains.

This paper proposes a decentralized peer-to-peer (P2P) service model for transferring between blockchain-based heterogeneous cryptocurrencies and CBDCs. The proposed service model provides a solution for transferring between blockchain-based heterogeneous cryptocurrencies and CBDCs without centralized intermediaries, such as cryptocurrency exchanges, banks, transfer service providers, and so on.

The contribution of this paper is as follows: to the best of our knowledge, there has been no previous study on a decentralized P2P service model for transferring between blockchain-based heterogeneous cryptocurrencies and CBDCs, and the proposed service model, based on an interoperable architecture that shares ledger data without intermediaries between heterogeneous blockchains, provides a solution for transferring between blockchain-based heterogeneous cryptocurrencies and CBDCs. The proposed decentralized P2P service model improves user convenience and ledger data security compared to the existing centralized service model.

This paper is organized into the following sections. Section 1 introduces cryptocurrency market trends and CBDC-related activities. Section 2 proposes an interoperable architecture to share ledger data without intermediaries between heterogeneous blockchains. Section 3 describes related studies including a problem with the transfer between blockchain-based heterogeneous cryptocurrencies and CBDCs. Section 4 proposes a decentralized P2P transfer service model to solve the problem identified in Section 3. Section 5 identifies security threats to the proposed service model and specifies security requirements to counter those security threats. Section 6 discusses the results and concludes the paper.

2. Interoperable Architecture between Heterogeneous Blockchains

This section proposes an interoperable architecture to share ledger data without intermediaries between heterogeneous blockchains. Interoperability between heterogeneous blockchains should be required to transfer between blockchain-based heterogeneous cryptocurrencies and CBDCs.

The proposed interoperable architecture is based on the proposed service model in Section 4 for sharing ledger data between heterogeneous blockchains. The proposed interoperable architecture is a decentralized architecture without intermediaries, whereas existing interoperable architectures, such as the inter-blockchain communication (IBC) protocol and the heterogeneous multi-chain framework described in Section 3.2, are centralized architectures with intermediaries.

In Figure 1, the blockchain-based interoperable management system (BIMS) maintains the registered information of blockchains and distributes common operations (COPs) to the contact nodes running on the blockchains. The BIMS does not store and maintain the ledger data from blockchain-1 and blockchain-2. Blockchain-1 and blockchain-2 can directly share ledger structure and ledger data through contact node-1 and contact node-2. The registered information of the blockchains includes the names of the blockchains, names of the consensus algorithms, names of the cryptocurrencies, IP addresses of the contact nodes, and more. The contact nodes running on the heterogeneous blockchains share data between the heterogeneous blockchains by common operations.

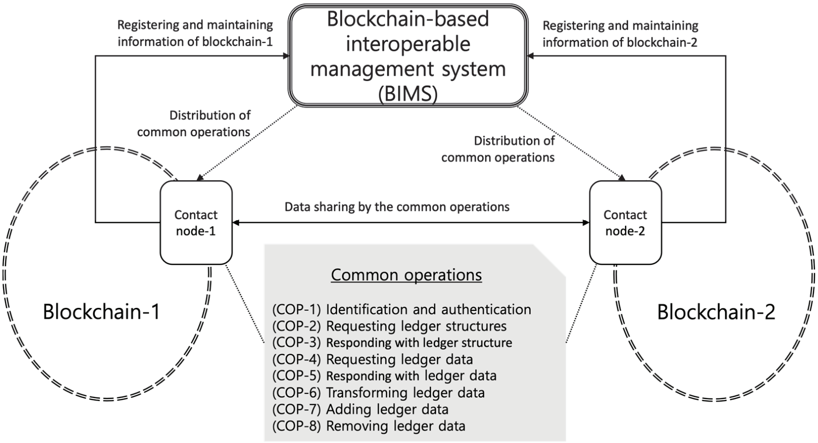

Registering and maintaining information of blockchain-1
Blockchain-based interoperable management system (BIMS)
Registering and maintaining information of blockchain-2
Distribution of common operations
Distribution of common operations
Contact node-1
Data sharing by the common operations
Contact node-2
Blockchain-1
Blockchain-2
Common operations
(COP-1) Identification and authentication
(COP-2) Requesting ledger structures
(COP-3) Responding with ledger structure
(COP-4) Requesting ledger data
(COP-5) Responding with ledger data
(COP-6) Transforming ledger data
(COP-7) Adding ledger data
(COP-8) Removing ledger data

Figure 1. The interoperable architecture between heterogeneous blockchains.

The common operations are described as follows:

- (COP-1) Identification and authentication: operation for mutual identification and authentication between heterogeneous blockchains;
- (COP-2) Requesting ledger structures: operation to request the ledger structures of other blockchains;
- (COP-3) Responding with ledger structure: operation to provide the ledger structure of one's own blockchain in response to operation '(COP-2) Requesting ledger structures';
- (COP-4) Requesting ledger data: operation to request ledger data from other blockchains;
- (COP-5) Responding with ledger data: operation to provide the ledger data of one's own blockchain in response to operation '(COP-4) Requesting ledger data';
- (COP-6) Transforming ledger data: operation of converting (e.g., processing, combining, etc.) ledger data provided from other blockchains according to the ledger structure and data format of one's own blockchain;
- (COP-7) Adding ledger data: operation to add the data converted (e.g., processed, combined, etc.) by operation '(COP-6) Transforming ledger data' to the ledger of one's own blockchain;
- (COP-8) Removing ledger data: operation to delete ledger data provided from other blockchains.

In Figure 2, the ledger data sharing process based on the interoperable architecture between heterogeneous blockchains is described as follows:

1. Contact node-1 and contact node-2 register the information (e.g., the names of blockchains, names of the consensus algorithms, names of the cryptocurrencies, the IP addresses of the contact nodes, etc.) of blockchain-1 and blockchain-2 with the BIMS;
2. BIMS distributes the common operations to contact node-1 and contact node-2;
3. Contact node-1 and contact node-2 identify and authenticate each other by COP-1;
4. Contact node-1 requests contact node-2 for the ledger structure of blockchain-2 by the COP-2;
5. Blockchain-2 provides its own ledger structure to contact node-2;
6. Contact node-2 responds to contact node-1 with the ledger structure of blockchain-2 by COP-3;
7. Contact node-1 requests contact node-2 for the ledger data of blockchain-2 by COP-4;
8. Blockchain-2 provides its own ledger data to contact node-2;
9. Contact node-2 responds to contact node-1 with the ledger data of blockchain-2 by COP-5;

10. Contact node-1 transforms the ledger data of blockchain-2 by COP-6, and then contact node-1 stores the transformed ledger data to blockchain-1 by COP-7. Contact node-1 removes the transformed ledger data and the ledger data of blockchain-2 by COP-8.

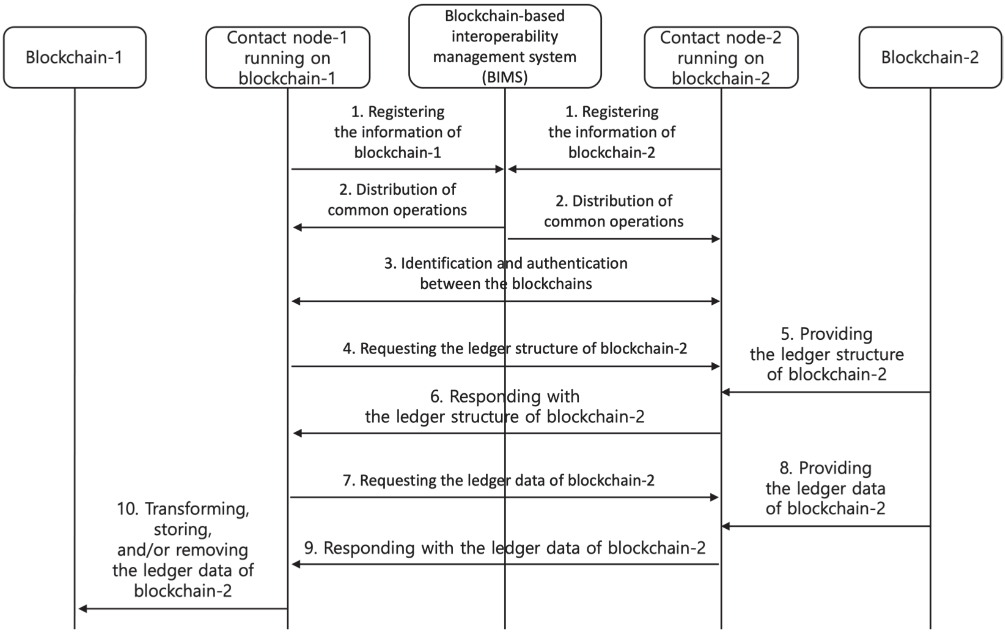

Blockchain-1
Contact node-1 running on blockchain-1
Blockchain-based interoperability management system (BIMS)
Contact node-2 running on blockchain-2
Blockchain-2
1. Registering the information of blockchain-1
1. Registering the information of blockchain-2
2. Distribution of common operations
2. Distribution of common operations
3. Identification and authentication between the blockchains
4. Requesting the ledger structure of blockchain-2
5. Providing the ledger structure of blockchain-2
6. Responding with the ledger structure of blockchain-2
7. Requesting the ledger data of blockchain-2
8. Providing the ledger data of blockchain-2
9. Responding with the ledger data of blockchain-2
10. Transforming, storing, and/or removing the ledger data of blockchain-2

Figure 2. The ledger data sharing process based on the interoperable architecture between heterogeneous blockchains.

3. Related Studies

This section describes the problem with the transfer between blockchain-based heterogeneous cryptocurrencies and CBDCs and examines other studies related to the problem.

3.1. Problem with the Transfer between Blockchain-Based Heterogeneous Cryptocurrencies and CBDCs

It is easy for users to transfer cryptocurrencies within the same blockchain (e.g., Bitcoin blockchain [7], etc.). For example, when an originator with a Bitcoin wallet wants to transfer to a beneficiary with a Bitcoin wallet, the originator can easily transfer Bitcoins to the beneficiary using the beneficiary's wallet addresses within the Bitcoin blockchain.

However, it is difficult for users to transfer cryptocurrencies between heterogeneous blockchains (e.g., a transfer between Bitcoin blockchain and Ethereum, etc.). For example, when an originator with a Bitcoin wallet wants to transfer to a beneficiary with an Ether wallet, the originator cannot transfer Bitcoins to the beneficiary using the beneficiary's wallet addresses within the Ethereum. This problem is due to the lack of interoperability between heterogeneous blockchains. Due to the nature of blockchain, transfer between blockchain-based heterogeneous CBDCs has the same problem as cryptocurrency. Additionally, transfer between blockchain-based cryptocurrencies and CBDCs encounters the same problem.

3.2. Other Approaches for the Transfer between Blockchain-Based Heterogeneous Cryptocurrencies and CBDCs

Several organizations and studies have made proposals to solve the problem mentioned in Section 3.1, but their proposals differ from the proposed service model in terms of concept and concreteness.

The inter-blockchain communication (IBC) protocol is proposed in [8]. The Cosmos is a network of independent parallel blockchains with a Tendermint [9] consensus algorithm, such as the practical byzantine fault tolerance (PBFT [10]) consensus algorithm. The Cosmos Hub will be the first blockchain in the Cosmos network. Many other blockchains are connected by the Cosmos Hub using the IBC protocol. The Cosmos Hub can track many token types and record the total number of tokens for each connected blockchain. All inter-blockchain coin transfers go through the Cosmos Hub, allowing tokens to be transferred from one blockchain to another without a liquid exchange between blockchains. The Cosmos Hub is an intermediary that connects heterogeneous blockchains.

The heterogeneous multi-chain framework Polkadot is proposed in [11]. Polkadot is a sharded blockchain, meaning it connects several blockchains together in a single network, allowing them to process transactions in parallel and exchange data between blockchains [12]. Polkadot allows any type of data to be sent between any type of blockchains [12]. Polkadot is an intermediary connecting heterogeneous blockchains, which is very similar to the Cosmos Hub.

The hub-and-spoke payment route called universal payment channels (UPC) is proposed in [13]. UPC can be used to support digital currency transfers of funds across different networks through payment channels. UPC hub can be useful in the context of CBDCs to support cross-border payment flows between CBDCs that may run on different blockchains [13]. UPC hub can also play an important role between private stablecoins [14] and public CBDCs by providing permissioned access for whitelisted stablecoins to be interoperable with CBDCs. The UPC hub concept that emerged would connect different blockchains by establishing dedicated payment channels between them—whether that means connecting CBDC blockchains between countries or connecting CBDC blockchains with vetted private stablecoin blockchains [15]. UPC hub is an intermediary that connects heterogeneous blockchains for CBDCs and stablecoins.

The blockchain implementation method for interoperability between CBDCs is proposed in [16]. This paper focuses on a blockchain system and management method, based on the ISO/IEC 11179 metadata registries (MDR) [17], for exchanges between CBDCs that records transactions between registered CBDCs. Furthermore, this paper describes implementing the blockchain system and experiment with the operation method, measuring the block generation time of blockchains using the proposed method.

The blockchain interoperability towards a sustainable payment system is proposed in [18]. This paper investigates different blockchain interoperability approaches, including industrial solutions, categorizing them, identifying the key mechanisms used, and listing several example projects for each category. As examples of the underlying technologies for cross-blockchain transactions, this paper describes the notary schemes such as centralized cryptocurrency exchanges (e.g., Coinbase [19], Binance [20], etc.), the sidechain-based solutions, the blockchain routers, the hashed time locks, and the industrial solutions (e.g., Cosmos Hub [8], Polkadot [11], etc.).

The formation and development of Von Hayek's theory of private money is analyzed in [21]. This paper concludes that when the national currency is replaced by digital currency, due to the international nature of digital currencies, both developing and developed economies will be vulnerable to 'digital dollarisation'. Moreover, this paper describes how governments can ask central banks to use a CBDC, which is preferable to a national currency for forecasting, computation, and accounting.

The main objective of this paper is to propose an interoperable architecture between heterogeneous blockchains without intermediaries, and a new decentralized P2P transfer service model based on the proposed interoperable architecture between blockchain-based heterogeneous cryptocurrencies and CBDCs.

4. Decentralized P2P Transfer Service Model and the Service Scenarios

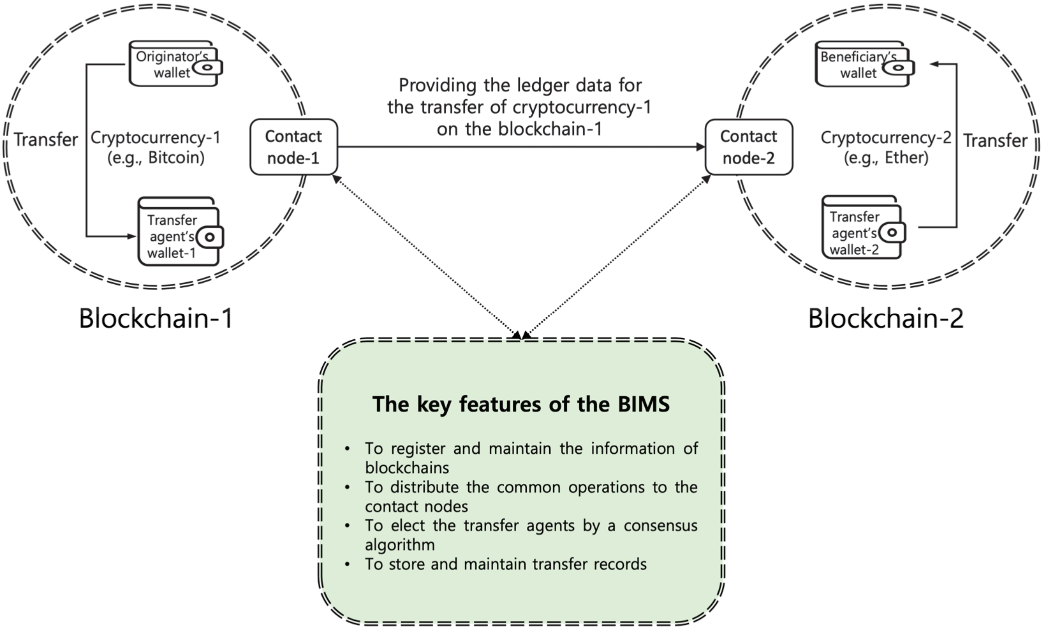
This section proposes a decentralized P2P service model based on an interoperable architecture for transferring between blockchain-based heterogeneous cryptocurrencies and CBDCs to solve the transfer problem mentioned in Section 3.1.

4.1. Service Model

The decentralized P2P service model, based on the interoperable architecture for transferring between blockchain-based heterogeneous cryptocurrencies and CBDCs, includes transfer between cryptocurrencies, transfer between cryptocurrency and CBDC, and transfer between CBDCs. In the proposed service model, the transfer agent is an entity that receives cryptocurrency and CBDC from the originator and sends another cryptocurrency and CBDC to the beneficiary. Any entity can be a candidate for the transfer agent.

In Figure 3, cryptocurrency-1 (e.g., Bitcoin) is transferred from the originator's wallet to the transfer agent's wallet-1 on blockchain-1. Contact node-1, running on blockchain-1, directly provides the ledger data for the transfer of cryptocurrency-1 to contact node-2, running on blockchain-2, without any intermediaries (see Figure 2). Cryptocurrency-2 (e.g., Ether) is transferred from the transfer agent's wallet-2 to the beneficiary's wallet on blockchain-2.

In Figure 4, cryptocurrency-1 (e.g., Bitcoin) is transferred from the originator's wallet to the transfer agent's wallet-1 on blockchain-1. Contact node-1, running on blockchain-1, directly provides the ledger data for the transfer of cryptocurrency-1 to contact node-2, running on blockchain-2, without any intermediaries (see Figure 2). CBDC-1 (e.g., Korean CBDC) is transferred from the transfer agent's wallet-2 to the beneficiary's wallet on blockchain-2.


Originator's wallet
Transfer
Cryptocurrency-1 (e.g., Bitcoin)
Transfer agent's wallet-1
Contact node-1
Providing the ledger data for the transfer of cryptocurrency-1 on the blockchain-1
Contact node-2
Beneficiary's wallet
Cryptocurrency-2 (e.g., Ether)
Transfer
Transfer agent's wallet-2
Blockchain-1
Blockchain-2

The key features of the BIMS

- To register and maintain the information of blockchains
- To distribute the common operations to the contact nodes
- To elect the transfer agents by a consensus algorithm
- To store and maintain transfer records

Figure 3. The service model for the transfer between cryptocurrency-1 and cryptocurrency-2.

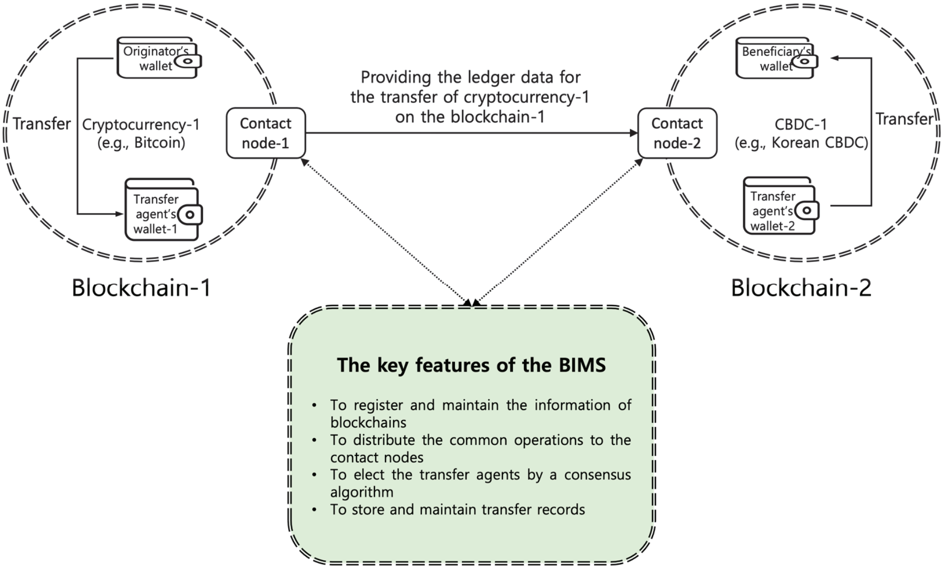

Originator's wallet
Transfer
Cryptocurrency-1 (e.g., Bitcoin)
Transfer agent's wallet-1
Contact node-1
Providing the ledger data for the transfer of cryptocurrency-1 on the blockchain-1
Contact node-2
Beneficiary's wallet
CBDC-1 (e.g., Korean CBDC)
Transfer
Transfer agent's wallet-2
Blockchain-1
Blockchain-2
The key features of the BIMS
• To register and maintain the information of blockchains
• To distribute the common operations to the contact nodes
• To elect the transfer agents by a consensus algorithm
• To store and maintain transfer records

Figure 4. The service model for the transfer between cryptocurrency-1 and CBDC-1.

In Figure 5, CBDC-1 (e.g., Korean CBDC) is transferred from the originator's wallet to the transfer agent's wallet-1 on blockchain-1. Contact node-1, running on blockchain-1, directly provides the ledger data for the transfer of CBDC-1 to contact node-2, running on blockchain-2, without any intermediaries (see Figure 2). CBDC-2 (e.g., US CBDC) is transferred from the transfer agent's wallet-2 to the beneficiary's wallet on blockchain-2.

The key features of BIMS are included in Figures 3–5. BIMS registers and maintains the information of the blockchains (e.g., the names of the blockchains, names of the consensus algorithms, names of the cryptocurrencies, IP addresses of the contact nodes, etc.), and distributes common operations (COPs) to the contact nodes running on the registered blockchains. The transfer agents are elected by a consensus algorithm. The transfer records are stored and maintained.

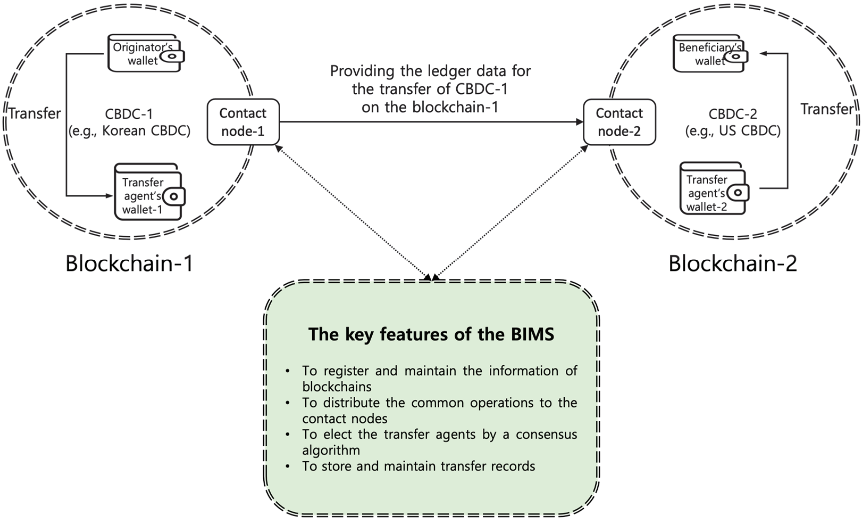

Originator's wallet
Transfer
CBDC-1 (e.g., Korean CBDC)
Transfer agent's wallet-1
Contact node-1
Providing the ledger data for the transfer of CBDC-1 on the blockchain-1
Contact node-2
Beneficiary's wallet
CBDC-2 (e.g., US CBDC)
Transfer
Transfer agent's wallet-2
Blockchain-1
Blockchain-2
The key features of the BIMS
• To register and maintain the information of blockchains
• To distribute the common operations to the contact nodes
• To elect the transfer agents by a consensus algorithm
• To store and maintain transfer records

Figure 5. The service model for the transfer between CBDC-1 and CBDC-2.

4.2. Service Scenarios and Data Flow

The service scenarios include the transfer between cryptocurrency-1 and cryptocurrency-2, the transfer between cryptocurrency-1 and CBDC-1, and the transfer between CBDC-1 and CBDC-2. The service scenario for the transfer between cryptocurrency-1 and CBDC-1 and the service scenario for the transfer between CBDC-1 and CBDC-2 are very similar to the service scenario for the transfer between cryptocurrency-1 and cryptocurrency-2.

In Figure 6, the service scenario for the transfer between the cryptocurrency-1 and cryptocurrency-2 is describes as follows:

1. BIMS elects transfer agents with wallets on blockchain-1 and blockchain-2 by a consensus algorithm;
2. Cryptocurrency-1 is transferred from the originator's wallet to the transfer agent's wallet on blockchain-1. In the transfer between CBDC-1 and CBDC-2, CBDC-1 is transferred from the originator's wallet to the transfer agent's wallet on blockchain-1. This process is performed by the originator;
3. The ledger for the cryptocurrency-1 transfer from the originator's wallet to the transfer agent's wallet is stored in the blockchain-1. The ledger data include the transfer date, the originator's wallet address, the transfer agent's wallet address, cryptocurrency-1 amount, and the fee amount for blockchain-1. In the transfer between CBDC-1 and CBDC-2, the ledger for the CBDC-1 transfer from the originator's wallet to the transfer agent's wallet is stored in blockchain-1. The ledger data include the transfer date, the originator's wallet address, the transfer agent's wallet address, CBDC-1 amount, and the fee amount for blockchain-1;
4. The record for the cryptocurrency-1 transfer from the originator's wallet to the transfer agent's wallet is stored in BIMS. Examples of the record data include the transfer date, the originator's wallet address, the beneficiary's wallet address, the amount of cryptocurrency-1, the fee amount for blockchain-1, and the fee amount for the transfer agent of blockchain-1. In the transfer between CBDC-1 and CBDC-2, the record for the CBDC-1 transfer from the originator's wallet to the transfer agent's wallet is stored in BIMS. Examples of the record data include the transfer date, the originator's wallet address, the beneficiary's wallet address, the amount of CBDC-1, the fee amount for blockchain-1, and the fee amount for the transfer agent of blockchain-1;
5. Contact node-1, running on blockchain-1, directly provides the ledger data to contact node-2, running on blockchain-2, without any intermediaries. The ledger data are for the cryptocurrency-1 transfer from the originator's wallet to the transfer agent's wallet on blockchain-1. In the transfer between CBDC-1 and CBDC-2, the ledger data are for the CBDC-1 transfer from the originator's wallet to the transfer agent's wallet on blockchain-1;
6. Cryptocurrency-2 equal to the amount of cryptocurrency-1 is transferred from the transfer agent's wallet to the beneficiary's wallet on blockchain-2. In the transfer between CBDC-1 and CBDC-2, CBDC-2 equal to the amount of CBDC-1 is transferred from the transfer agent's wallet to the beneficiary's wallet on blockchain-2. In the transfer between cryptocurrency-1 and CBDC-1, CBDC-1 equal to the amount of cryptocurrency-1 is transferred from the transfer agent's wallet to the beneficiary's wallet on blockchain-2. This process is performed by the transfer agent or an application that can use the transfer agent's private key;
7. The ledger for the cryptocurrency-2 transfer from the transfer agent's wallet to the beneficiary's wallet is stored in blockchain-2. For example, the ledger data include the transfer date, the transfer agent's wallet address, the beneficiary's wallet address, the cryptocurrency-2 amount, and the fee amount for blockchain-2. In the transfer between CBDC-1 and CBDC-2, the ledger for the CBDC-2 transfer from the transfer agent's wallet to the beneficiary's wallet is stored in blockchain-2. Examples of ledger data include the transfer date, the transfer agent's wallet address, the beneficiary's wallet address, the CBDC-2 amount, and the fee amount for the blockchain-2. In the transfer between cryptocurrency-1 and CBDC-1, the ledger for the CBDC-1 transfer

from the transfer agent's wallet to the beneficiary's wallet is stored in blockchain-2. For example, the ledger data include the transfer date, the transfer agent's wallet address, the beneficiary's wallet address, the CBDC-1 amount, and the fee amount for the blockchain-2;

8. The record for the cryptocurrency-2 transfer from the transfer agent's wallet to the beneficiary's wallet is stored in BIMS. Examples of the record data include the transfer date, the transfer agent's wallet address, the beneficiary's wallet address, the cryptocurrency-2 amount, the fee amount for the blockchain-2, and the fee amount for the transfer agent of blockchain-2. In the transfer between CBDC-1 and CBDC-2, the record for the CBDC-2 transfer from the transfer agent's wallet to the beneficiary's wallet is stored in BIMS. Examples of the record data include the transfer date, the transfer agent's wallet address, the beneficiary's wallet address, the CBDC-2 amount, the fee amount for the blockchain-2, and the fee amount for the transfer agent of the blockchain-2. In the transfer between cryptocurrency-1 and CBDC-1, the record for the CBDC-1 transfer from the transfer agent's wallet to the beneficiary's wallet is stored in BIMS. Examples of the record data include the transfer date, the transfer agent's wallet address, the beneficiary's wallet address, the CBDC-1 amount, the fee amount for the blockchain-2, and the fee amount for the transfer agent of blockchain-2.

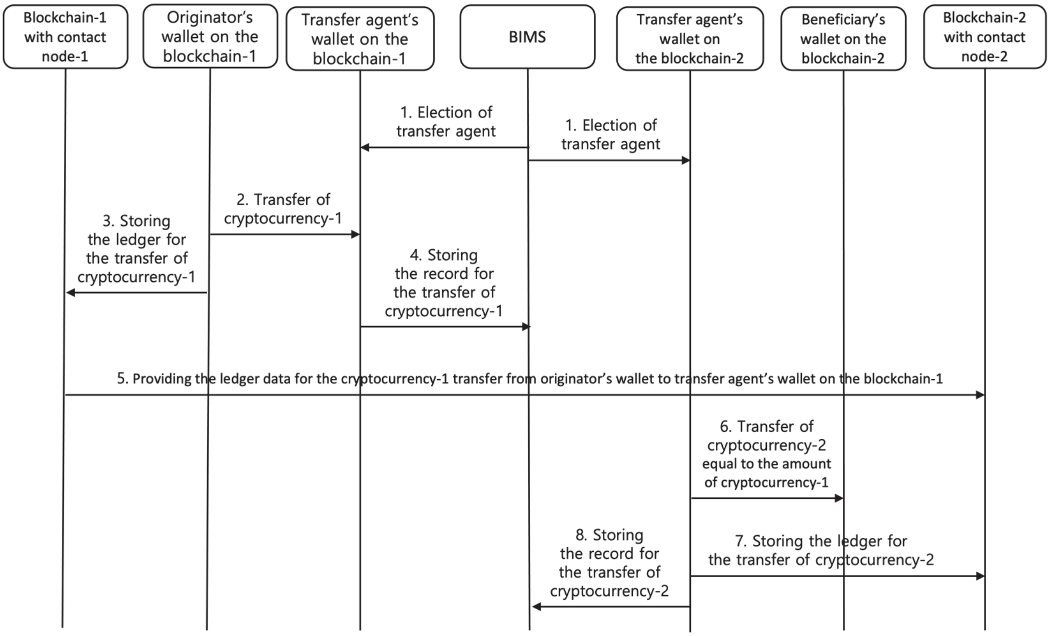

Blockchain-1 with contact node-1
Originator's wallet on the blockchain-1
Transfer agent's wallet on the blockchain-1
BIMS
Transfer agent's wallet on the blockchain-2
Beneficiary's wallet on the blockchain-2
Blockchain-2 with contact node-2
1. Election of transfer agent
1. Election of transfer agent
2. Transfer of cryptocurrency-1
3. Storing the ledger for the transfer of cryptocurrency-1
4. Storing the record for the transfer of cryptocurrency-1
5. Providing the ledger data for the cryptocurrency-1 transfer from originator's wallet to transfer agent's wallet on the blockchain-1
6. Transfer of cryptocurrency-2 equal to the amount of cryptocurrency-1
7. Storing the ledger for the transfer of cryptocurrency-2
8. Storing the record for the transfer of cryptocurrency-2

Figure 6. The service scenario for the transfer between the cryptocurrency-1 and cryptocurrency-2.

5. Security Threats and Requirements

Security threats (STs) to the proposed service model for the transfer between cryptocurrencies and CBDCs are identified and security requirements (SRs) countering the security threats are specified in this section.

5.1. Security Threats

Security threats to the proposed service model for the transfer between blockchain-based heterogeneous cryptocurrencies and CBDCs are identified as follows:

- (ST-1) Breach of contract by originator's transfer agents: If the originator's transfer agents and the beneficiary's transfer agents are not the same entity (for example, see

Figure 3), the originator's transfer agents may not pay the transfer amount excluding the transfer fee to the beneficiary's transfer agents. This threat may lead to the beneficiary's transfer agents not transferring the cryptocurrencies and CBDCs to the beneficiary. As a result, the transfer between cryptocurrencies and CBDCs will fail;

- (ST-2) Ledger data leakage during transmission between contact nodes: The ledger data can be leaked during transmission between the contact nodes running on heterogeneous blockchains. The leaked ledger data can be misused to steal cryptocurrencies and CBDCs;
- (ST-3) Massive ledger data leakage from blockchains: The massive ledger data can be leaked from blockchains registered with BIMS. The contact nodes running on blockchains which is registered with BIMS can request massive ledger data from the contact nodes running on other blockchains registered with BIMS. The leaked massive ledger data can be misused to track cryptocurrencies and CBDCs transfers. This threat can cause privacy issues related to the originators and beneficiaries;
- (ST-4) Monopoly by specific transfer agents: The transfer between cryptocurrencies and CBDCs can be monopolized by specific transfer agents. This threat can allow transfer agents that monopolize transfers to control transfers between cryptocurrencies and CBDCs. Ultimately, this threat can force the originators and beneficiaries to pay higher transfer fees;
- (ST-5) Data request by unauthorized blockchains: The contact nodes running on a blockchain which is not registered with BIMS can request ledger data from the contact nodes running on a blockchain registered with BIMS. The ledger data obtained from the blockchains registered with BIMS can be misused to steal cryptocurrencies and CBDCs.

The security threats are specific to the proposed service model for the transfer between cryptocurrencies and CBDCs, not to the general IT services.

5.2. Security Requirements

Security requirements countering the security threats identified in Section 5.1 are specified as follows:

- (SR-1) Stablecoin deposit: The proposed service model should allow the originator's transfer agents to deposit stablecoins equal to the amount of transfer prior to the transfer. As soon as the transfer from the originator's wallet to the transfer agent's wallet occurs, the stablecoins are automatically held in escrow by the smart contract [22,23] for the beneficiary's transfer agents. The smart contract runs on blockchains for stablecoins, such as Tether coin (USDT) on Ethereum;
- (SR-2) Data encryption in transmission: The proposed service model should provide safe cryptographic protocol (e.g., TLS) [24,25] to prevent ledger data leakage during transmission between the contact nodes running on heterogeneous blockchains. The ledger data should be protected with the cryptographic protocol in the transmission;
- (SR-3) Minimization of the amount of retrieved ledger data: The proposed service model should allow the contact nodes to minimize the amount of ledger data retrieved from the blockchains. More specifically, this can be implemented by narrowing the query conditions to seek ledger data;
- (SR-4) Election of transfer agents by a consensus algorithm: The proposed service model should elect the transfer agents by a consensus algorithm prior to the transfer. The elected originator's transfer agent and beneficiary's transfer agent may or may not be the same. Depending on the type of transfer (e.g., transfer between Bitcoin and Ether, transfer between Bitcoin and Korean CBDC, transfer between Korean CBDC and US CBDC, etc.), the transfer agents should be elected in consideration of the transfer agent's properties (e.g., wallet type, stablecoin deposit amount, transfer fee, etc.);
- (SR-5) Identification and authentication between the contact nodes: The proposed service model should provide an identification and authentication mechanism between

contact nodes. The contact nodes running on heterogeneous blockchains should identify and authenticate each other before sharing ledger data.

In Table 1, SR-1 (stablecoin deposit) can prevent ST-1 (breach of contract by originator's transfer agents). This means that if the originator's transfer agent does not pay the transfer amount, excluding the transfer fee to the beneficiary's transfer agent, the stablecoins deposited by the originator's transfer agent are automatically paid to the beneficiary's transfer agent by the smart contract. SR-2 (data encryption in transmission) can prevent ST-2 (ledger data leakage during transmission between contact nodes). This means that although the ledger data are leaked during transmission between the contact nodes running on heterogeneous blockchains, it is difficult to use the leaked ledger data encrypted with a cryptographic algorithm. SR-3 (minimization of the amount of retrieved ledger data) can prevent ST-3 (massive ledger data leakage from blockchains). This means that it is possible to prevent leakage of massive ledger data from blockchains by narrowing down the query conditions to seek ledger data in the contact nodes. SR-4 (election of transfer agents by a consensus algorithm) can prevent the ST-4 (monopoly by specific transfer agents). This means that the monopoly of specific transfer agents can be prevented by electing transfer agents based on the consensus algorithm for each transfer. SR-5 (identification and authentication between the contact nodes) can prevent ST-5 (data request by unauthorized blockchains). This means that the contact nodes running on blockchains which are not registered with BIMS cannot request ledger data from the contact nodes running on blockchains registered with BIMS, in accordance with the results of mutual authentication between contact nodes.

Table 1. Relationship between security threats and security requirements.

	SR-1 (Stablecoin Deposit)	**SR-2 (Data Encryption in Transmission)**	**SR-3 (Minimization of the Amount of Retrieved Ledger Data)**	**SR-4 (Election of Transfer Agents by a Consensus Algorithm)**	**SR-5 (Identification and Authentication between the Contact Nodes)**
ST-1 (breach of contract by originator's transfer agents)	O				
ST-2 (ledger data leakage during transmission between contact nodes)		O			
ST-3 (massive ledger data leakage from blockchains)			O		
ST-4 (monopoly by specific transfer agents)				O	
ST-5 (data request by unauthorized blockchains)					O

(Note: ST = security threat; SR = security requirement).

6. Discussion and Conclusions

The main objective of this paper is to propose an interoperable architecture between heterogeneous blockchains, and a new decentralized P2P service model for the transfer between blockchain-based heterogeneous cryptocurrencies and CBDCs. The experimental evaluation of the proposed service model could be done as future work.

This paper identifies potential security threats to the proposed service model and describes security requirements to prevent the identified security threats. The proposed service model should be implemented to meet the security requirements.

The interoperable architecture enables the exchange of transaction ledger data of cryptocurrency and CBDC without intermediaries between heterogeneous blockchains. This enables cryptocurrency and CBDC to be transferred by decentralized transfer agents, even if the originator's blockchain and the beneficiary's blockchain are different.

The service scenario in Figure 6 demonstrates that the transfer between an originator and a beneficiary with heterogeneous cryptocurrency and CBDC can be processed very conveniently and usefully. This is because the originator does not have to consider what the beneficiary's wallet type is. Thus, the proposed service model based on the proposed interoperable architecture solves the transfer problem between heterogeneous blockchain-based cryptocurrencies and CBDCs.

There are several advantages of the proposed service model: (1) The proposed interoperable architecture allows the sharing of ledger data between heterogeneous blockchains without intermediaries. (2) BIMS provides the common operations for sharing ledger data between the blockchains, rather than storing and maintaining the ledger data retrieved from the blockchains. (3) The proposed service model allows the transfer between cryptocurrencies, between cryptocurrency and CBDC, and between CBDCs without cryptocurrency exchanges and banks.

There are several reasons why BIMS service provider and transfer users would be interested in accepting the proposed service model: (1) The originator can directly transfer cryptocurrencies and CBDCs regardless of the beneficiary's wallet type. (2) The originator does not need to exchange the cryptocurrency and CBDC to be transferred for the same cryptocurrency and CBDC as the beneficiary's wallet type. (3) Transfer fees to be paid by the originators and beneficiaries are lower than the centralized organization, such as cryptocurrency exchanges, banks, transfer service providers and so on. (4) BIMS service providers are not burdened with storing and maintaining the ledger data retrieved from other blockchains for interoperability.

The proposed interoperable architecture will be developed as an international standard by ITU-T (International Telecommunication Unit) SG17, and the proposed service model will be developed as Korean ICT standard by TTA (Telecommunications Technology Association) PG1006. Private companies will be able to implement the proposed service model based on the interoperable architecture as a decentralized P2P transfer system by technology transfer in the future.

Author Contributions: Conceptualization, K.P.; methodology, K.P.; validation, K.P. and H.-Y.Y.; formal analysis, K.P.; investigation, K.P.; writing—original draft preparation, K.P.; writing—review and editing, K.P. and H.-Y.Y.; supervision, H.-Y.Y.; project administration, H.-Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This research was implemented as part of the project "Standardization Lab. for Next-generation Cybersecurity" (Project Number: 2021-0-00112) supported by MSIT (the Ministry of Science and ICT) and IITP (Institute of Information & Communications Technology Planning & Evaluation).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. CoinMarketCap. Today's Cryptocurrency Prices by Market Cap. Available online: https://coinmarketcap.com (accessed on 20 October 2022).
2. International Monetary Fund (IMF). The Future of Money: Gearing up for Central Bank Digital Currency. Available online: https://www.imf.org/en/News/Articles/2022/02/09/sp020922-the-future-of-money-gearing-up-for-central-bank-digital-currency (accessed on 20 October 2022).
3. The Federal Reserve System. Money and Payments: The U.S. Dollar in the Age of Digital Transformation. Available online: https://www.federalreserve.gov/publications/files/money-and-payments-20220120.pdf (accessed on 20 October 2022).
4. Zhang, T.; Huang, Z. Blockchain and central bank digital currency. *ICT Express* **2022**, *8*, 264–270. Available online: https://www.sciencedirect.com/science/article/pii/S2405959521001399 (accessed on 20 October 2022). [CrossRef]
5. Nakamoto, S. Bitcoin: A Peer-To-Peer Electronic Cash System. Available online: https://bitcoin.org/bitcoin.pdf (accessed on 20 October 2022).
6. Ethereum. Ethereum White paper. Available online: https://ethereum.org/en/whitepaper/ (accessed on 20 October 2022).
7. Blockchain.com. Bitcoin Explorer. Available online: https://www.blockchain.com/explorer?view=btc (accessed on 22 October 2022).
8. Kwon, J.; Buchman, E. Cosmos Whitepaper. Available online: https://v1.cosmos.network/resources/whitepaper (accessed on 24 October 2022).
9. Tendermint. Tendermint Core Documentation. Available online: https://docs.tendermint.com (accessed on 24 October 2022).
10. Castro, M.; Liskov, B. Practical Byzantine Fault Tolerance. Available online: https://pmg.csail.mit.edu/papers/osdi99.pdf (accessed on 24 October 2022).
11. Wood, G. Polkadot: Vision for a Heterogeneous Multi-Chain Framework Draft 1. Available online: https://polkadot.network/PolkaDotPaper.pdf (accessed on 24 October 2022).
12. Wood, G. An Introduction to Polkadot. Available online: https://polkadot.network/Polkadot-lightpaper.pdf (accessed on 24 October 2022).
13. Christodorescu, M.; English, E.; Gu, W.C.; Kreissman, D.; Kumaresan, R.; Minaei, M.; Raghuraman, S.; Sheffield, C.; Wijeyekoon, A.; Zamani, M. Universal Payment Channels: An Interoperability Platform for Digital Currencies. Available online: https://arxiv.org/pdf/2109.12194v2.pdf (accessed on 24 October 2022).
14. Arner, D.; Auer, R.; Frost, J. Stablecoins: Risks, Potential and Regulation. Available online: https://www.bis.org/publ/work905.pdf (accessed on 24 October 2022).
15. Gu, C. Making Digital Currency Interoperable, Visa Shares New Thinking on Cross-Chain Interoperability. Available online: https://usa.visa.com/visa-everywhere/blog/bdp/2021/09/29/making-digital-currency-1632954547520.html (accessed on 24 October 2022).
16. Jung, H.; Jeong, D. Blockchain Implementation Method for Interoperability between CBDCs. *Future Internet* **2021**, *13*, 133. [CrossRef]
17. International Organization for Standardization (ISO). *ISO/IEC 11179-1:2015*; Information Technology—Metadata Registries (MDR)—Part 1: Framework. ISO: Geneva, Switzerland, 2015.
18. Mohanty, D.; Anand, D.; Aljahdali, H.M.; Villar, S.G. Blockchain Interoperability: Towards a Sustainable Payment System. *Sustainability* **2022**, *14*, 913. [CrossRef]
19. Coinbase. Coinbase—Buy & Sell Bitcoin, Ethereum, and More with Trust. Available online: https://www.coinbase.com/ (accessed on 26 October 2022).
20. Binance. Buy/Sell Bitcoin, Ether and Altcoins | Cryptocurrency Exchange | Binance. Available online: https://www.binance.com/en (accessed on 26 October 2022).
21. Mikhaylov, A.Y. Development of Friedrich von Hayek's theory of private money and economic implications for digital currencies. *Terra Econ.* **2021**, *19*, 1. [CrossRef]
22. Zheng, Z.; Xie, S.; Dai, H.-N.; Chen, W.; Chen, X.; Weng, J.; Imran, M. An overview on smart contracts: Challenges, advances and platforms. *Future Gener. Comput. Syst.* **2020**, *105*, 475–491. [CrossRef]
23. Negara, E.S.; Hidayanto, A.N.; Andryani, R.; Syaputra, R. Survey of Smart Contract Framework and Its Application. *Information* **2021**, *12*, 257. [CrossRef]
24. OpenSSL Software Foundation. Vulnerabilities. Available online: https://www.openssl.org/news/vulnerabilities.html (accessed on 5 November 2022).
25. Internet Engineering Task Force (IETF). The Transport Layer Security (TLS) Protocol Version 1.3. Available online: https://tools.ietf.org/html/rfc8446 (accessed on 5 November 2022).

big data and cognitive computing

MDPI

Article

Combining Sociocultural Intelligence with Artificial Intelligence to Increase Organizational Cyber Security Provision through Enhanced Resilience

Peter R. J. Trim [1,*] and Yang-Im Lee [2]

[1] Department of Management, School of Business, Economics and Informatics, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK

[2] Department of Marketing and Business Strategy, Westminster Business School, University of Westminster, 35 Marylebone Road, London NW1 5LS, UK

* Correspondence: p.trim@bbk.ac.uk

Abstract: Although artificial intelligence (AI) and machine learning (ML) can be deployed to improve cyber security management, not all managers understand the different types of AI/ML and how they are to be deployed alongside the benefits associated with sociocultural intelligence. The aim of this paper was to provide a context within which managers can better appreciate the role that sociocultural intelligence plays so that they can better utilize AI/ML to facilitate cyber threat intelligence (CTI). We focused our attention on explaining how different approaches to intelligence (i.e., the intelligence cycle (IC) and the critical thinking process (CTP)) can be combined and linked with cyber threat intelligence (CTI) so that AI/ML is used effectively. A small group interview was undertaken with five senior security managers based in a range of companies, all of whom had extensive security knowledge and industry experience. The findings suggest that organizational learning, transformational leadership, organizational restructuring, crisis management, and corporate intelligence are fundamental components of threat intelligence and provide a basis upon which a cyber threat intelligence cycle process (CTICP) can be developed to aid the resilience building process. The benefit of this is to increase organizational resilience by more firmly integrating the intelligence activities of the business so that a proactive approach to cyber security management is achieved.

Keywords: artificial intelligence; cyber security manager; cyber threat intelligence; learning; resilience; sociocultural intelligence

Citation: Trim, P.R.J.; Lee, Y.-I. Combining Sociocultural Intelligence with Artificial Intelligence to Increase Organizational Cyber Security Provision through Enhanced Resilience. *Big Data Cogn. Comput.* **2022**, *6*, 110. https://doi.org/10.3390/bdcc6040110

Academic Editor: Fabrizio Baiardi

Received: 26 August 2022
Accepted: 1 October 2022
Published: 8 October 2022

1. Introduction

For an organization to become more resilient, top management needs to take heed of the fact that cyber attacks are likely to intensify in the years ahead and because of this, cyber security needs to be placed in a strategic cyber security management context [1]. The need for such an approach is clear, bearing in mind that: "Even with U.S. company losses due to cyberattacks nearing a reported $1 trillion by late 2020, a survey of nearly 1000 organizations found that only 44% had cyber preparedness and incident response plans in place" [2] (p. 2). It seems logical, therefore, for managers to counteract cyber attacks by utilizing cyber security technology more fully, but also for them to discover new ways to engage in cyber security management. A key role of senior management is to help managers draw on operand and operant resources so that they can strengthen the organization's defenses against cyber attacks.

Advice relating to the appropriateness of cyber security technology comes in the form of government advice, highly specialized companies that operate cyber security technological solutions, consultants that have in-depth knowledge of cyber security problems and working practices, and university research teams that develop specific types of security software. There are, of course, other sources of intelligence that originate from government agencies and specialist consultancies, for example. Taking this into account, it can be suggested that managers need to adopt a pro-active approach to cyber security as resilience requires that intelligence-gathering involves the deployment of technology that has the power of human cognition and the ability to learn/reason and hear/see [3] (p. 109). An important point that surfaces, however, is that to be effective, organizational resilience needs to be placed within the context of how organizational staff coordinate investment in cyber security across the supply chain [4] (p. 169). Bearing this in mind, it is pertinent to suggest that cyber security management is to be viewed as a strategic-level capability [5], whereby security is linked with business continuity management and a set of procedures whereby security is placed within a crisis/disaster management setting. The case can be made, therefore, for a cyber security manager to be appointed to take charge of cyber security, which is at the heart of an organization's security [1].

Understanding the motivations of those who carry out a cyber attack means having an in-depth appreciation of human behavior and establishing what causes an individual to behave in an anti-social/illegal manner. The cyber security manager is, therefore, required to have an appreciation of human psychology and possess adequate knowledge of how cyber security policy is formulated and implemented, if they are to provide guidance and advice to a range of functional heads. If a data breach does occur and results in reputational damage and an increase in adverse publicity resulting from a fine imposed by regulators, then cascading effects may have a debilitating effect on the organization and its trading partners. It is for this reason that the cyber security manager needs to have both technical and managerial knowledge relating to cyber security or have expertise available to them that can be drawn on when necessary.

The remit of the cyber security manager is to work with other senior managers and devise, manage, and implement cyber security policy decisions across the organization's networks. The focus of the research is, therefore, to explain how different approaches to intelligence (i.e., the intelligence cycle and the critical thinking process) can be combined and linked with cyber threat intelligence (CTI), which utilizes AI. To explain this, we explore how the cyber security manager can draw on social interaction and establish how it drives cognition [6] (p. 306). This can be viewed as logical in terms of establishing organizational resilience because cyber security management requires the cyber security manager to develop and share cyber security knowledge with individuals that are viewed as first responders. Social interaction is enhanced through trust-based relationships and open communication between staff and provides the basis for institutionalized learning. This gives rise to a defined risk mitigation policy and strategy within partner organizations and the utilization of cyber security models [7].

In terms of AI-based cyber attacks, it can be argued that cyber security experts will be required to intensify their effort to develop AI defense systems [8]. This will require that risk mitigation strategies are put in place to counteract cyber attacks; it also will focus attention on cyber defense from an intellectually driven and holistic perspective. It is with this in mind that the focus of the paper was to outline how sociocultural intelligence can be combined with AI to increase the organizational cyber security provision and enhance an organization's level of resilience. In doing so, we focused our efforts on providing answers to two questions: (1) How can a non-security specialist develop their appreciation and understanding of resilience through undertaking threat intelligence? (2) How can knowledge regarding different types of AI help managers better understand the complexities associated with different algorithms and their functionality vis-à-vis different types of defense system?

To assist us in our task, we drew on the knowledge derived from a small group interview that involved an academic researcher discussing various aspects of intelligence in relation to organizational security with five experts. Each participant had spent over twenty years in security and had worked in different industries and was known to be an expert in the field of organizational resilience. We contribute to the field of cyber security management by combining elements of the intelligence cycle (IC) with the critical thinking process (CTP) [9] (p. 139) to produce a cyber threat intelligence cycle process (CTICP). This should enable staff to adopt a strategic cyber security intelligence perspective. We also highlight the importance of organizational learning and how it facilitates a higher level of intelligence that involves sociocultural interaction and thus makes the organization more resilient. The advantage of this approach is that we reflect on the interplay between centralized versus localized learning and how sociocultural intelligence is viewed as a necessary component of the strategic cyber security management process. Finally, through linking AI with sociocultural intelligence, we outline the steps in the cyber threat intelligence cycle process (CTICP) that enable managers across various industries to adopt a resilience centric approach that hardens the organization.

2. Background

Bearing in mind that those carrying out cyber attacks are becoming more sophisticated and linked more firmly to those carrying out all types of scams, the cyber security manager needs to make a value judgement with regard to how cyber threat intelligence (CTI) is perceived by top management and how, because operant resources are scarce, staff can draw on technological aids such as artificial intelligence (AI) to enhance their cyber threat intelligence (CTI) decision-making capability. Hasan et al. [10] (p. 354) indicated that the advanced persistent threat (APT) is challenging organizational defenses because signature-based defense mechanisms are unable to respond in real-time to new types of malicious code/intelligent mutant codes. It is worth noting that “Conventional cybersecurity tools look for historical matches to known malicious code, so hackers only have to modify small portions of that code to circumvent the defense. AI-enabled tools, on the other hand, can be trained to detect anomalies in broader patterns of network activity, thus presenting a more comprehensive and dynamic barrier to attack” [10] (p. 354).

Surya [11] (p. 991) has provided a useful definition of AI: “Artificial intelligence (AI) refers to the technology involved in the development of smart machines and software. This includes the developments of applications and systems that can reason, collect intelligence, prepare intelligently, learn, interact, interpret, and manipulate objects”. Hence, AI allows users of big data to capture data from a variety of sources, store the data, and apply analytics so that decision-makers can use the outcome [11] (p. 992) in a variety of contexts (e.g., tactical and strategic).

AI can help managers to interpret patterns of cyber attack, and the outcome of a cyber threat intelligence (CTI) analysis can be placed in report form so that senior managers can offer advice based on the type of threat identified with a view to utilizing operand and operant resources. In addition, those charged with managing security can interact more fully with other functional managers and establish how cyber threat intelligence (CTI) can be strengthened using AI. However, it is worth noting that although it is recognized that AI can be used to defend an organization against cyber attacks [12] (p. 363), there are a number of challenges that senior management need to overcome vis-à-vis the use of AI. One such problem is the gap in knowledge relating to what AI/ML represents and how AI/ML can be used by managers operating at different levels of authority. It is possible to suggest that the complexities associated with AI/ML may well militate against individual managers understanding how AI can be used. To overcome the likely resistance of using AI, we propose that managers first develop an appreciation of AI/ML and think of how AI/ML can benefit them in terms of their decision-making so that the day-to-day operations are reinforced through contingency plans.

Managers need to be mindful of the fact that AI is refined through the application of ML but “humans are able to understand the behavior of others in terms of their mental states-intentions, beliefs and desires-by exploiting what is commonly designated as ‘folk psychology’” [13] (p. 279). By acknowledging this, managers can avoid the various pitfalls associated with the use of AI, especially the contradiction whereby chatbots are used to help individuals (i.e., those using an organization’s website) to gain certain information by responding/acting in known and logical ways. Gallese [13] (p. 285) suggests that although it is possible to make sense of how people respond to an event, with regard to human social cognition, "Language is the most specific hallmark of what it means to be human". It is with this in mind that we reflect on and pose the question: how can sociocultural intelligence be linked with AI to increase an organization's resilience?

Before progressing, we consider it necessary to reflect on the notion of what resilience is and to have a clear understanding of what it incorporates. HSSAI [14] (p. 9) provides a useful definition of resilience by indicating that it is "the ability of a system to attain the objectives of resisting, absorbing, and recovering from the impact of an adverse event, before, during, and after its occurrence. It is also a dynamic process that seeks to learn from incidents to strengthen capabilities of the system in meeting future challenges. The goals are to maintain continuity of function, degrading gracefully, and recover system functionality to a pre-designated level, as rapidly as desired and feasible".

The focus is clearly to learn from an event/incident and to make sure that those with operational responsibility can "learn from incidents", as this is what machine learning sets out to achieve. In the context of organizational learning, whereby the focus of attention is on how an individual's skill level is enhanced, Argyris [15] (p. 8) provides guidance by indicating that: "Learning is defined as occurring under two conditions. First, learning occurs when an organization achieves what it intended; that is, there is a match between its design for action and the actuality or outcome. Second, learning occurs when a mismatch between intentions and outcomes is identified and it is corrected; that is, a mismatch it turned into a match".

Whether data are collected, analyzed, and interpreted by humans or are left to a machine(s) is not what is under consideration. What is important to acknowledge is that adequate resilience requires managers to consider how best to utilize intelligence and to make use of limited intelligence. McCreight [16] (p. 5) has offered a comprehensive view as to what resilience encompasses by indicating that there are five main dimensions of resilience, which are: personal and familial socio-psychological well-being; organizational and institutional restoration; economic and commercial resumption of services and productivity; restoring infrastructural systems integrity; and operational regularity of public safety and government. The five dimensions highlighted prove useful with regard to a manager developing a comprehensive understanding of what resilience involves and how to place resilience within an organization–government–society context. Whether the relationships developed are transactional in nature or transformational in nature depends upon the organization's value system, and the leadership style/model in place.

In order to utilize big data to counteract sophisticated cyber attacks, managers are paying increased attention to the capability of AI and its deployment. Hence, it is useful to acknowledge two main but contradictory issues: the volume of data that needs to be processed versus the time available to carry out an analysis, which yields an outcome that has relevance and can be acted upon. Additionally, attention needs to be given to the cost of hiring experts for labeling the data, which relates to the issue of supervised learning, semi-supervised learning, and unsupervised learning. In terms of cyber threat protection, deep learning (DL) is receiving renewed attention. For example, one area that needs immediate attention is ransomware attacks. Andrade and Yoo [17] (p. 2) noted that between 2014 and 2017, 327 families of ransomware were identified that accounted for 184 million attacks. Because cyber criminals are behind such attacks and do, of course, use technology to carry out their actions, it would be logical to suggest that advances in deep learning (DL) will help those involved in cyber security to protect computer systems and networks better. An interesting and relevant point raised by Andrade and Yoo [17] is how cyber security specialists can consider using psychology to enhance cyber security situation(al) awareness and they make clear that cognitive sciences can be utilized to enhance cyber security.

Bearing in mind that the focus of this paper was to deepen our understanding of cyber threat intelligence (CTI) and provide arguments as to how AI/ML can help senior managers to make an organization more resilient, we first need to take cognizance of what Dawson [18] (pp. 268–269) has said about an organization as it provides the basis for better understanding the relational processes that allow individual managers to utilize technology for the benefit of the organization and its partners, and at the same time, provide the basis for strategic cyber security management [1] that is aimed at safeguarding the organization against cyber attack. Dawson [18] (pp. 268–269) highlights seven points that epitomize an organization: (i) an interactive system (e.g., change in one aspect will have repercussions for another); (ii) high level of complexity (e.g., uncertainty is evident); (iii) there is no single way in which to manage a situation; (iv) resources are scare; (v) different interest groups prevail (e.g., conflict, consensus and indifference are evident); (vi) constraints exist that effect action; and (vii) the level of the individual/group needs to be known in order to identify and solve problems. It is with these seven points in mind that we embrace the view that organizational resilience is dependent upon managers having a clear appreciation of what sociocultural intelligence involves and how AI can be utilized to enable managers to make more informed cyber security-based decisions.

3. Placing Sociocultural Intelligence in Perspective

The concept of sociocultural intelligence has been gaining momentum over a number of years and it is clear that the field of intelligence is expanding, and new perspectives are being offered that allow managers such as the cyber security manager to comprehend how intelligence is managed across organizational networks. To ensure that AI is not misused, we advocate a cautious and incremental approach to its use but also advise a wider understanding of AI's application in terms of intelligence provision. What can be deduced from the study of intelligence is that sociocultural intelligence (SOCINT) is purported to include "the process of directing, collecting data related to any of the social sciences, analyzing, producing, and then disseminating such data for situational awareness in any operational environment" [9] (p. 11). This is a well-known and accepted view. To better understand the antecedents of cyber threat intelligence CTI), we suggest that managers take cognizance of the intelligence cycle (IC) process and the critical thinking process (CTP), as outlined by Patton [9] (p. 139). The intelligence cycle (IC) is known to be composed of five separate but linked stages including (i) planning and direction; (ii) collection; (iii) processing; (iv) analysis and production; and (v) dissemination. The critical thinking process (CTP) is known to include eight separate but linked stages: (i) purpose; (ii) question at issue; (iii) information; (iv) interpretation and inference; (v) concepts; (vi) assumptions; (vii) implications and consequences; and (viii) point of view. By merging the intelligence cycle (IC) with the critical thinking process (CTP), it should be possible to establish how AI can be utilized by managers to better understand the role that cyber threat intelligence (CTI) plays and how it is to be managed across organizations. Before we explain this, we need to understand how the differences in learning capabilities associated with AI/ML can be drawn on to provide an intelligence focused appreciation, leading to an enhanced appreciation of resilience. To achieve this, we focused on AI/ML in relation to business so that managers in charge of various business functions can relate better to the learning capabilities afforded by AI/ML, and not worry too much about the technical aspects. Should managers need to, they can deepen their knowledge of AI/ML by consulting those with expert knowledge and/or attend specialist courses of study.

4. Algorithms and Their Learning Capability

Deep learning (DL) is a subset of AI, and it structures algorithms in layers to create an artificial neural network (e.g., a human brain) for filtering information and learning from it and making intelligent/informed decisions. DL applies ML to large datasets. ML uses algorithms to analyze, learn from the data, and make decisions based on the learning. Both DL and ML are subsets of AI. It is useful to note that different algorithms have their own unique functionality and capability for learning, some of which can be used for specific tasks. Table 1 shows different forms of learning in DL. AI systems can be divided into three types such as narrow AI (which is goal-oriented and programmed to perform a single task); general AI (representative of a machine that can learn, understand, and act in a way similar to that of a human in a given situation); and super AI (a hypothetical AI where a machine exhibits intelligence that surpasses the brightest humans).

Table 1. The different types of learning associated with functionality/capability.

Functionality for Different Types of Learning	ML, DL, and AI Learning Style and Algorithm	Use of AI/ML in Business
Mechanical	ML—(un)supervised—minimum degree of learning	To predict similar event in future, e.g., utilize simple tasks such as greeting or simple order taking on behalf of waiters/waitresses; self-service. Algorithms for supervised learning such as: linear and logistic regression, decision tree, k-nearer neighborhood (KNN), naïve Bayes (NB), random Forest, neural network (NN), support vector machine (SVM). Algorithms for unsupervised learning such as: K-means clustering, factor analysis (FA), principal component analysis (PCA), DBSCAN, SVD.
Analytical	**DL—supervised**	Deep learning (DL) is inspired by the way a human brain works for filtering information, which helps a computer model to filter data through layers and classify information. Selecting an advertisement for a particular platform that will gain more popularity via surfing the Internet (to increase clickability based on algorithm learning to make a match between a particular advertisement that was placed and individuals that visit a particular site). Spam filter. DL network architectures are classified such as convolutional neural networks (CNN) and recurrent neural networks (RNN).
	CNN	Used widely for the use for visual image/object analysis, classification, e.g., search engines and recommender systems—Facebook/Google photos suggest tag by recognizing the face.
	RNN	It uses sequential data, which is distinguished by memory, and prior inputs influence current input and output, but the outcome can vary depending on the type of RNN such as for music generation, sentiment classification, and machine, e.g., IBM Watson Studio and Watson Machine Learning, trailers ("binging show") in Netflix, OTT platforms. Monitoring buying habits—particular types of platforms for shopping, surfing the purchasing history of groups of customers and placing them into similar purchasing segments to market specific items among suitable segments [19].
	DL—Semi-supervised learning (SSL)	Combination of supervised and unsupervised learning for data deduction and labeling from large unlabeled data. It can be used for graph-based label propagation; speech recognition; web content classification; text document classification (e.g., URL: https://www.altexsoft.com/blog/ (accessed on 15 June 2021)).
	DL—Unsupervised learning	Without human intervention, algorithms work on datasets to identify hidden patterns or for groupings based on similarities/differences in the data, e.g., useful for building recommend system (look at "rating" and "preference").
	Self-organizing maps	Typically, 2-dimentional representation. Useful for visualization.
	Boltzmann Machines	Handwritten and visual object recognition tasks [20].
Intuitive	AutoEncoders	Useful for reducing audio data, e.g., through anomaly detection algorithms to detect specific fraud.
Intuitive	**Reinforcement learning (RL)—ML/part of AI**	Watson's Jeopardy (question-and-answer system), AlphaGo, Mario, Deepmind, self-driven car, Keras (in libraries), etc. Alpha Zero (RL with AI).
Intuitive	Advanced AI—focused on cognition	Able to make a deductive decision based on the analysis of data and able to predict without data input like human (gut feeling based on cognition of patterns in the data).
Empathetic	Super Advanced AI—Focus on emotion and empathy	Identify consumer emotional status and interact empathetically through the use of natural language processing [21].

Source: The authors.

Managers in various industries such as banking, the motor industry, and health care have paid careful attention to AI implementation in relation to learning capabilities. Retailers utilize augmented reality for a better image (e.g., ASOS, visual search [22] and some retailers such as M&S and Kohl's have partnered with Snapchat and implemented a virtual fitting room [23]). The use of an avatar, a virtual character, with virtual reality and/or with a chatbot, is also gaining the attention of an increasing number of managers in business. It allows them to create virtual social touch points as well as create entertaining effects that result in a richer customer experience and higher customer engagement [24–26].

The application of methods and algorithms in AI/ML varies and produces a specific effect in the way in which the interaction process with end users is managed. Different algorithms also have implications for the types of data that are needed and how the data are captured and analyzed. AI is concerned with designing intelligent systems that exhibit characteristics associated with human intelligence and behavior and involves cognitive processes such as adapting to the latest information and problem-solving [27]. AI's capability varies, for example, Google Home and Alexa, integrate AI and advanced analytics (ML algorithms); chatbots sense the context of the conversation, but cannot perform a set of activities on their own; virtual assistants (e.g., Alexa, Apple Siri, Google assistant or Corona) provide daily activities such as emails or schedule meetings and can crawl through existing resources for a range of requests but with regard to customer service, however, they cannot resolve queries on their own [28] and friendly conversational chatbots such as Mitsuku and Replika, which are humanoid AI, are able to respond to emotional verbal reactions in a meaningful way [29,30]. What can be noted from this is that the communication process between a potential customer and the organization itself can be enriched by staff providing reassurance about the organization in terms of its resilience, which is mapped to an end user's understanding of security awareness. From this, we can identify the following question: how can an individual's learning capability be enhanced through using AI/ML? Finding an answer is important because managers need to link AI learning with the analysis of data and the interpretation of data so that the intelligence derived can be evidence based and used to underpin various plans/strategies. However, we stress that it is not just about AI enhancing what the organization is in terms of its commitment to dealing with customer requests or undertaking cyber threat intelligence (CTI) analysis. It is more about assuring external individuals that the staff are pro-active in terms of security awareness and can link the need for intelligence with the learning capability of those interested in buying the company's products and services, so they feel confident in buying from the company and avoid buying from rogue websites.

Learning can, according to Campbell et al. [31] and Ma and Sun [19] be divided into four types: supervised learning; semi-supervised learning; unsupervised learning; and reinforcement learning. In supervised learning, an expert trains the system by feeding labeled training data and defines variables to algorithms whereas in the case of unsupervised learning, the machine can learn inductively from unlabeled/unorganized data by analyzing the datasets to draw meaningful correlations or inferences by identifying hidden groups or grouping patterns. It can be noted that reinforcement learning (RL) is behavior-driven auto-learning where the algorithm/model (called agent) learns from interaction with its environment (by choosing from a set of possible actions) and their outcome. The sequential order and time plays an important role in reinforcement learning and is linked with a reward or penalty depending on the performance correctness and attempts to maximize the cumulative number of rewards.

The functions of AI/ML in an online business context can be grouped. For example, the basic mechanical function is an analytical tool, and the intuitive function includes humanoid AI [32]. Understanding different functions of AI/ML is useful as it helps managers to choose appropriate AI/ML tools in relation to the company's positioning strategy. We reiterate that the positioning strategy links learning with security awareness and is derived from the leadership style/model and organizational value system.

With regard to the basic mechanical function, it is based on rule-based learning at the minimum and relies on prior knowledge to perform repeated routines and/or transactional tasks (e.g., search engine used by Google or Bing). The analytical function relates to how information is processed for problem-solving in logical reasoning and how AI/ML tools learn from it. It is advanced, rule-based learning that carries out complex tasks and executes rational decisions (e.g., Deep Blue, IBM's chess player). The intuitive function incorporates digital technology that can mimic a human's learning intuitively. It is this, we feel, that can be used to ensure that sociocultural intelligence can be harnessed to get an individual to look more deeply at the issues relating to cyber threat intelligence (CTI) and map the outcomes to their own level of security awareness. Table 1 outlines the different types of AI/ML associated with learning and their use in business and is for illustrative purposes only. The differences in supervised learning, semi-supervised learning, unsupervised learning, and reinforcement learning are discussed next.

4.1. Supervised Learning

There are various algorithms for supervised learning such as a neural network (that has layers of nodes and trains data by mimicking the connectivity of the human brain, through each node being made up of inputs, weights, a threshold (bias)), and output; K-nearest neighbors (for prediction); naïve Bayes (is a classification method and well-used for text mining, spam filtering, and a recommend system); linear regression (used to identify relationships between a dependent and one or more independent variables); logistic regression (used to produce binary output by leveraging linear regression); support vector machine (SVM) (used for both data classification and regression, however, especially useful in the decision boundary to separate classes of data points); and decision tree (based on one input variable, each step split an existing subset into two, and has the capability of intuitive interpretations [19,33].

With respect to the analytical function of AI and ML, there are various levels of sophisticated applications [33,34]. For example, a convolutional neural network (CNN) is normally used for visual image analysis, classification, medial recreation, and is the recommended system, for example, whereas recurrent neural network (RNN) uses sequential data, which is distinguished by memory, and prior inputs influence the current input and output [19,35], but the outcome can vary depending on the type of RNN such as music generation, sentiment classification, and machine translation (e.g., IBM Watson Studio and Watson Machine Learning) [36]. The use of supervised learning in retail allows managers to use a shopper's basket datasheet to further define sub-segment groups by using the price of each product and the budget of an individual. It helps to uncover demand patterns for different products at different stores. For example, the combination of regression techniques may allow a retailer to predict the probability of a target variable (e.g., predict churn and switching behavior) that measures the satisfaction and engagement in the website characteristics and demographic information. This can be considered as confidence building from the perspective of the customer and provides them with a sense of well-being. However, supervised learning requires knowledge and the time to train the model, which can result in human error, which affects whether the algorithm performs as expected. Reflecting on this point, it can be suggested that should an error occur for whatever reason, it is likely that the end user will become less trusting of the technology and therefore seek to purchase another company's product/brand.

4.2. Semi-Supervised Learning (SSL)

The SSL approach is a combination of supervised and unsupervised ML. SSL uses small amounts of labeled data and a large amount of unlabeled data to train a model to label data. It is useful in a situation where limited labeled training data are available with a large amount of unlabeled samples [37]. According to Ouali et al. [38], SSL and its applications can be used to reduce the amount of labeled data required either by developing new methods or adopting existing SSL frameworks for a DL setting. For example, cluster analysis is a method that groups datasets into homogenous subgroups that contain similar characteristics in the data such as the same gender or common group associations as the goal is to identify the similarities and differences between data points. The application of cluster analysis in SSL is to use some known cluster information to classify other unlabeled data, which uses both labeled and unlabeled data. There are various methods and approaches such as consistency regularization (or consistency training) for perturbed vision, for example, proxy-label uses a heuristic approach and leverages trained model on the labeled set to produce training examples by labeling unlabeled sets; generative models use learned features on one task that can be transferred to other tasks; and graph-based methods that propagate labels from labeled nodes to unlabeled nodes by using the similarities of two nodes [38].

4.3. Unsupervised Learning

With regard to the unsupervised machine learning algorithms, these include K-means clustering for identifying groups and iteration, factor analysis (FA), principal component analysis (PCA, to reduce dimensions), DBSCAN (density-based spatial clustering of applications with noise, which are used for data mining), and singular value decomposition (SVD) [19]. In unsupervised deep learning, the learning models such as self-organizing maps (SOMs); Boltzmann machines and AutoEncoders [39,40] are used to reduce dimensionality as the output is always 2-dimentional and is well-used. These allow the user to identify clusters of a specific type of input pattern [41]. The network of Boltzmann machines (or stochastic model) is a systematically connected neuron-like sampling learning algorithm and allows for interesting features in complex training data to be identified [42]. AutoEncoders are used in processing audio raw data into secondary vector space (e.g., word2vec) and have various variations such as spare AutoEndoders (allows a hidden layer and a reduction in overfitting), or contractive AutoEncoders (prevents overfitting and copying of values from hidden layers, add to the loss function), which are useful in terms of building the recommend systems or reducing dimensionality [35].

Unsupervised learning is useful for monitoring a system or building a binary recommend system. For example, it can be used to detect specific types of fraud. The key aspect of unsupervised learning is to unveil hidden patterns or groups from unlabeled large volumes of data, faster than supervised ML can do. Based on past purchase data, unsupervised ML can assist managers to identify trends in the data that can be used to plan a cross-selling strategy through add-on recommendations to customers during the check-out stage [43]. However, there are some aspects that need attention. Issues such as complexity in computation to train high volumes of data, and a lack of clarity as to which data were clustered and how the data were labeled. This means that users need time to understand the labeling and classifications, and interpretation. Unsupervised learning can be used for segmentation or understanding different customer groups, which helps managers to redefine their communication strategy better to fit the needs of certain groups and to monitor for fraudulent transactions or analyze the customer preference based on their search history [44].

4.4. Reinforcement Learning (RL)

Reinforcement learning (RL) models are either positive or negative based. The methods for RL such as SARSA (state-action-reward-state-action for learning Markov decision process policy), n-step method (the increment for rewards is estimated value of at time t, that incorporates n-step backup), actor–critic methods (or TD methods), and Q-learning [45]. Q-learning is value-based learning, which helps the agent (model) determine the optimal action within an environment. Examples of RL are in AlphaGo, Alpha Zero, Mario, Deepmind in Google data centers (with AI), self-driven car (with AI), and Keras in libraries [19]. RL can be applied widely such as self-driving in the automotive industry, for business strategy planning and data processing, but attention is needed in various aspects such as the parameters as this may affect the speed of learning.

Intuitive AI is an artificial natural network based on deep learning that can level up the result of analytics through the emulation of a wide range of human cognition and learning and the adaption of intuitively based understanding (e.g., Google's Deep mind (AI)). In AI development, there are different types of AI such as narrow AI is descriptive and performs one task at a time (answers are provided to the question of what happened); general AI, which is diagnostic (answers to comprehend the question of why did it happen) and makes a decision based on learning (independent); and predictive (answers to the question of what might happen next) [46,47]. Intuitive AI can identify anomalies in the dataset and make a deduction based on analyzed information, which, for example, helps to detect threats in financial services [46,47].

Some applications such as Replica, Sophia, Ellie, Nao, and Kasper recognize emotion and learn and adapt when interacting with humans. Empathy is an important ingredient in social interaction. Through the retailer deploying humanoid AI, they can manage the relationship with customers better as they can respond better to consumer requests by being able to detect the consumer's emotional state [48,49]. This can be looked upon positively as it represents a commitment to the customer centric approach and making the customer feel safe knowing that their needs are understood and that effort has gone into service their requirements, thus ensuring their expectations are met.

From the above, it can be noted that there are many different algorithms with different capabilities and functionalities, which associate with different levels of expert requirements and commitments. Table 1 is useful as it briefly outlines the different types of learning and their capabilities/functionalities and their application, especially in relation to DL. It provides a basic understanding to people who are enthusiastic in terms of using big data, but who have a limited knowledge of information technology and its application. Table 1 can be considered as useful with regard to answering questions such as:

(1) How should individuals make a decision as to what type of algorithm(s) is to be used or combined for the effective use of AI?

(2) What are the differences between supervised learning and unsupervised learning, and the implications regarding commitments vis-à-vis the expected quality outcome and the implication for implementation?

(3) Which aspects of the functionality of a system (e.g., mechanical, analytical or intuitive) should an individual focus on and why?

5. Improving Cyber Security through Utilizing AI

It can be argued therefore that various managers (e.g., marketing managers, logistics and distribution managers, and finance managers) will have knowledge of the use of AI, and will understand the benefits afforded by AI. Hence, it is possible for managers to relate the use of AI from advertising and product promotion to security awareness and counteracting fraud by making staff aware of the need to improve their security behavior. For example, Bresniker et al. [50] (p. 46) provided a number of insights into how AI can be used to aid the cyber security management process, especially from the stance of detecting threats and state: "AI/ML can drive down response times from hundreds of hours to seconds and scale analyst effectiveness from one or two incidents to thousands daily. With an adequate knowledge base, it can preserve corporate knowledge and use that knowledge to automate tasks and train new analysists".

Bresniker et al. [50] (p. 46) indicate that AI/ML will be increasingly used to:

1. Create pattern-matching tools that highlight security issues in networks;
2. Automate mundane tasks so that cyber security staff can use their time to respond to events in real time; and
3. Identify a range of threats and ensure that appropriate action is taken.

Bresniker et al. [50] provide a useful guide as to how AI/ML can enhance cyber security, however, in order for various managers in the organization to work together and provide an integrated approach toward strategic cyber security management [1], whereby the cyber security manager works closely with various other managers including the risk manager, the business continuity manager, the IT manager, and the training manager, for example, it is necessary to match the human dimension of cyber security (e.g., identify human vulnerabilities) with the technical dimension of cyber security (e.g., identify technical vulnerabilities) through the application of the concept of sociocultural intelligence. The reason why matching is necessary is because fake news/disinformation is causing confusion and disruption and is likely to be weaponized further and used to complement various forms of cyber attack.

Fake news is well-orchestrated and targeted [51]. Petratos [52] (p. 764) draws on the United Nations definition and suggests that disinformation has been used "to confuse or manipulate people through delivering dishonest information to them". Bearing in mind that there has been an upward movement in ransomware attacks, managers need to realize that dealing with cyber criminals is not always as straight forward as expected. Drawing on the work of Greenberg, Tatar et al. [53] make known that a ransomware attack may be confused with data destruction malware whereby there is no possibility that the data would be made available to the target because the master boot records are in fact deleted by those carrying out the attack. It is for this reason that senior security managers within organizations need to develop a holistic approach to security because they may not be aware of the subtly behind disinformation. By accepting that disinformation detection requires a large investment in AI/ML, it should be possible for managers to develop resilience-based security by integrating cyber threat detection with security awareness.

6. Materials and Methods

To gain insights into how the concept of resilience can be embedded in the psyche of the organization so that it is a recognized component of the organization's memory, one of the researchers of this paper undertook a small group interview involving five highly knowledgeable organizational security experts. The experts were selected on the basis that they were knowledgeable in terms of strategic intelligence and were well able to place threat intelligence in the context of an organization's commitment to building resilience. The participants were all based in London and received permission from their employer to be involved in the research. Originally, it was envisaged that two small group interviews would be undertaken but it was not possible to organize two separate groups because those approached were busy and had commitments. Those that did attend and participate possessed operational knowledge that allowed them to offer unique insights into the topics under discussion [54] and uncover the underlying conditions [55]. In addition, they were known to have served in various senior security positions within an organization and were able to establish how intelligence and security could be integrated better so that security provision across business functions could be improved. The small group interview method was chosen because it allows for broad based questions to be asked that result in an open-ended group interview [56] (p. 17), whereby the participants can articulate their view, challenge and critique their peers, and then provide unique insights and solutions. Indeed, the selection of the group members (e.g., senior security professionals with work experience gained in both the private sector and the public sector) proved valuable in the sense that it was necessary to establish a group ethos [57] (p. 354) that allowed for meaning through reflection [9] (pp. 116–117). The small group interview was limited to one and a half hours and prior to the group interview commencing, the participants had agreed that the interview could be audio recorded. The researcher-facilitator agreed that specific comments made by individuals would not be attributed to the individuals concerned or the organization that they worked for. The group interview was framed so that the insights provided allowed for a holistic view of security to be derived that could then be interpreted from an organizational intelligence perspective. An interactive style was adopted during the small group interview, and this allowed each participant to explore the subject matter in the way they considered appropriate.

When undertaking a small group interview, it is important for the researcher to give attention to what the purpose of the group interview is and how the group members relate to each other. For the purpose of this research, the objective was not to look at a basic set of conditions or derive insights in relation to government policy. The objective was to bring a highly experienced group of security experts together so that they could provide an in-depth understanding and appreciation of the topics discussed [58]. This was conducted by placing intelligence in the context of organizational resilience and at the same time, allowing each participant to gain intellectual satisfaction and knowledge in relation to perfecting their own organizational resilience policy. A semi-structured, open-ended approach was adopted as this allowed specific questions to be posed and provided the participants with some latitude to branch out and provide answers that incorporated real world examples.

In order to generate the required data, a number of questions were posed during the small group interview that included: How useful is the organizational learning concept in relation to the development of a security culture? How effective is transformational leadership in terms of the strategic intelligence approach? How can organizational vulnerabilities be eradicated through threat intelligence? The advantage of this approach is that the predetermined open-ended questions used were supplemented with additional questions that emerged as the interview progressed [59] (p. 315). The sub-questions that emerged were related to a range of topics that surfaced including crisis management, intelligence tools, networks, organizational skills, outsourcing, transformational leadership, trend analysis, trustworthy behavior, and risk management.

The data collection process was judged important in terms of the evidence and linking theory and practice. However, it was recognized that differences in regulatory conditions meant that senior security managers in one industry operated under different conditions compared to security managers in other industries. Although the view taken by the researchers was that the regulatory conditions exhibited differences, they were differences in degree only.

Immediately after the small group interview had been completed, the transcript was transcribed and then analyzed by the researchers. Each participant was provided with a copy of their portion of the transcript so that they could verify what they had said. Each participant, and indeed the facilitator, were assigned a number as names had not been used, and were identified accordingly. For the data analysis, the inductive approach was used whereby “the patterns, themes, and categories” were derived from the data as opposed to being imposed by the researchers before the data were collected [56] (p. 390). The main themes were identified and reported in [60]. In terms of the analysis of the data, we adapted the process associated with the grounded theory approach whereby we undertook open coding, axial coding, and selective coding [60], and developed a set of themes. The researchers then constructed a narrative in relation to each of the main themes. This would help the non-security specialist to understand how security practitioners placed threat intelligence in a sociocultural context from the perspective of enhancing an organization’s resilience. This allowed the researchers to relate the main themes identified back to the intelligence cycle (IC) and critical thinking process (CTP) so that a cyber threat intelligence cycle process (CTICP) could be produced that was generic in nature and could be extended or adapted by managers in different industries.

7. Results

From the small group interview, it was clear that organizational learning, transformational leadership, organizational restructuring, crisis management, and corporate intelligence emerged as the main management considerations (themes) to be taken into account by top management because together, they provided insights into how threat intelligence was viewed and managed.

Organizational learning is viewed as important because it is a process whereby the mindsets of managers can be changed to embrace organizational values. In relation to how threats can be confronted and communicated to stakeholders, it is important for threat-based intelligence to be shared in real-time. As security covers a range of sensitive topics, it is for this reason that staff are required to understand what trustworthy behavior is and why acts of benevolence are considered important and underpin relationship building. By establishing trust-based relationships, it is easier for individuals to share information when necessary and to safeguard themselves. This can be achieved by working within the organization's ethical code of conduct. Managers need to understand that the insider threat is continually evident, and the best approach appears to be for senior management to establish clearly defined security related roles that individuals can adopt when performing their duties. This means that security training needs to be formalized and a distinction made between training and education. The latter can be viewed as a higher level of knowledge attainment and inclusive of the understanding of what cyber threat intelligence (CTI) involves and how it is used on a day-to-day basis. Although not all staff need to be aware of the technical aspects of cyber security, those in positions of responsibility are required to have an all-round appreciation of the subject. In-house, formal cyber security awareness programs can be organized and administered on a continual basis to up-date staff and to make sure information technology staff talk with staff throughout the organization about security related issues. This should prove beneficial in terms of establishing and maintaining a security culture and ensuring that staff are aware of why and how they are to relate to law enforcement personnel when problems occur such as fraudulent practice, for example.

Transformational leadership was considered as a precursor of organizational change and is brought about through the implementation of strategic vision. Acknowledging that people can become complacent, it is necessary to ensure that people also do not become demotivated and lose sight of important considerations such as day-to-day security. However, transformational leadership is about establishing an organizational security culture, which should be viewed as a collectivist process. Another point that arose was that staff need to develop an understanding of the needs of people in other organizations. This will help staff to recognize symptoms such as corrupt practices and inefficiency in operations that could prove detrimental to the organization and its partners. Part of the transformational process involves staff using their own social network(s) to gain intelligence about cyber related attacks and centralizing this in the form of threat intelligence within a central command and control system within the organization. With regard to the security skill base of employees, managers need to ensure that security is defined in a certain way so that risk management is given adequate attention. To ensure that transformational leadership is effective, people within the organization that are viewed as supportive of security initiatives can adopt the role of champion of the cause and be given prominence to participate in in-house security seminars.

Organizational restructuring can result in an upheaval that places the organization at risk, especially when the management's attention is focused on other, non-security issues. Internal conflict can result in an organization becoming vulnerable because the type of uncertainty being dealt with relates more to an organization's internal situation than an outside threat penetrating the organization's defenses.

Crisis management is considered necessary because it can be assumed that at some point in time, the organization will be penetrated, and it is likely that other partner organizations will also be affected. Although essentially crisis management may be undertaken in different ways (e.g., depending upon the size and complexity of the organization), it must be noted that there are both direct and indirect influences involved. The organization's value system needs to support teamwork and requires that crisis management is viewed as an essential and combined process, whereby senior management make known to employees what a resilient organization is and how such an organization remains resilient. Areas often overlooked or underplayed include cyber insurance, and therefore the risk management process needs to be more formalized than it sometimes is.

Corporate intelligence is aided by the process of risk management and an area of attention is advances in biometrics, which covers threats brought about by fake IDs. Regarding the protection of the identity of employees, managers need to ensure that a person's identity is always safeguarded and because information about an individual can be used against them, attention needs to be given to issues such as identity theft. This means that risk management is viewed from several perspectives and can also be related to human resource management policy and the recruitment of staff both from within the country and from abroad.

The findings from the small group interview highlight various issues that the cyber security manager needs to be aware of. These include the need to define what the organization's stance is in terms of security and resilience; and what the boundaries are that staff need to pay attention to when sharing information. These are important considerations with regard to how staff obtain data and information from outside the organization and share intelligence/knowledge with internal staff so that a cyber attack does not get through the organization's defenses. The quality of the data shared, and the way in which the data are shared, need to be given consideration in advance of a crisis occurring. During a crisis (e.g., an attack has penetrated the organization's defenses and staff struggle to deal with it in real-time), staff need to follow the policy laid down and ensure that cascading effects do not materialize.

Security awareness is, therefore, reflective of the investment in security training and education, however, it is recognized that more investment is needed in making staff aware of the consequences of an impact and convincing them that a proactive approach to gaining cyber security knowledge from appropriate sources is viewed as good practice.

8. Discussion

As well as placing emphasis on the quality of information/data derived from outside the organization, we also focus attention on the use of AI and whether managers can deploy supervised, semi supervised, or unsupervised algorithms for data analysis. This brings to attention whether managers have the knowledge required to interpret the results of the analysis (e.g., through human interpretation or machine interpretation) or whether a higher-level knowledge interpretation is required. Senior managers do need to invest time and effort into discussing these points and will need to put in place a protocol that provides guidance with regard to the analysis of big data. Acknowledging that sociocultural intelligence needs to be analyzed in a certain way and is dependent on the insights of experts brings to the fore the fact that managers need to consider the issue of resource availability.

The findings from the small group interview also highlight the need for a senior security manager/cyber security manager to adopt a transformational approach to security whereby threat analysis is an integral part of intelligence activity. By including current information pertaining to cyber threats, it is possible to highlight the need for cyber threat analysis to be viewed as necessary and to advocate a strategic cyber security management [1] approach. This will provide a basis for cyber security to be more widely appreciated than it is at present by managers that have a non-technical disposition. By adopting a more corporate intelligence focused approach to cyber security, whereby the lead organization takes greater responsibility for security, especially cyber security, guidance and support can be provided to the suppliers. Security staff, and the cyber security manager in particular, can promote the stakeholder view of security whereby supply chain partners take responsibility for updating their security and at the same time, pass threat intelligence data and information onto other stakeholders/network members. This will allow each stakeholder to coordinate their investment in cyber security [4]. The key point to note is that AI/ML can assist managers to undertake cyber threat intelligence (CTI), however, gaining permission across various supply chains is time consuming and requires negotiated access. This involves the sharing of sensitive data and information, and a commitment to building a sociocultural intelligence knowledge base.

To achieve linkage between security and intelligence, it is necessary to have an appropriate leadership model in place that embraces organizational learning and integrates the key aspects of the intelligence cycle (IC) with the critical thinking process (CTP) [9] (p. 139). The COVID-19 pandemic is continuing to have a lasting effect on the international economy, and evidence of this can be seen in the actions of unscrupulous individuals who are intent on exploiting health care provision [61]. By including issues such as fake news, identity theft, and ransomware, for example, in cyber threat intelligence (CTI), it is possible to establish how organized criminals are exploiting the market for legitimate drugs by engaging in online activities in relation to COVID-19 and the methods by which they gain financially. A question that arises is how can senior management devise a strategic approach to cyber security management that results in a collectivist appreciation whereby organizational partners pool resources to mitigate the risks identified? This is a question that top management appears to be discussing but the problem basically remains that not all business relationships are long-term in orientation. Opportunistic behavior may militate against a more structured and integrated approach to cyber security management across supply chains. Another issue that arises is, if partner organizations do not cooperate and share risk related data/information, how is a potential crisis to be effectively dealt with in real-time? Although the cyber security manager may focus on a specific type of cyber threat, it can be suggested that the scale of the problem means that it is necessary to utilize AI/ML to help counteract a range of cyber attacks.

It can be noted that AI is developing through time and its capability is to be viewed as several inter-locking AI and ML capabilities. By progressing from supervised to unsupervised learning and beyond, AI and ML assume a high level of decision-making that is freeing managers to invest time in strategy formulation as they are no longer required to undertake a lot of the analytical tasks themselves. Hence, it can be suggested that managers view the utilization of AI in terms of fostering the strategic capability of the organization and aiding business planning and resilience policy. To understand how AI is to be implemented requires strategic vision and a commitment to investing in a range of platforms (business platforms, enterprise platforms and enabling platforms) [62] that provide the company with a sustainable competitive advantage through relationship building.

Through establishing data-driven knowledge base construction, cyber security staff can guard against the problem of "inaccurate entity recognition and unreliable property/relation discovery due to insufficient training data" [63] (p. 11). In other words, it can be pointed out that cyber security specialists should work with those involved in cognitive sciences such as psychology to better understand how cyber security awareness and other areas of interest such as situational analysis can be incorporated more fully into the process of cyber threat intelligence (CTI) [17]. This should ensure that spikes of activism are noted and linked with disruptive geopolitical campaigns and specific types of hacking activity. Furthermore, emerging trends in fraudulent behavior may be linked to deteriorating economic conditions and the rise in criminal behavior, whereby organized criminal syndicates seek and exploit new market opportunities (e.g., fake websites linked to fictious products and services). Through converting information into intelligence and developing cyber security knowledge, a formalized approach to cyber threat intelligence (CTI) will materialize. Hence, threat actors need to be identified and categorized and this can be conducted by means of a threat template that outlines the opportunities in relation to the selected threat actors [64] (p. 6). By establishing the motivations of threat actors and linking through with their intended actions, it should be possible to understand the nature of the threat(s) and how matters escalate and an impact occurs [64] (p. 8).

Intra- and inter-company relationship building is important from the stance of sharing and utilizing threat-based data and information and can be considered as an integral part of cyber threat intelligence (CTI). Incident analysis tools exist [65] (p. 169) that can undergo further development that results in new initiatives in cyber security provision. It is also hoped that the sharing of such technology will encourage more dialogue between governments and a concerted effort will arise that results in a greater pooling of resources and cutting-edge joint research projects. The logic underpinning this view is to acknowledge that the pressures on managers to analyze big data will increase and new ways of detecting threats need to be found and implemented across industry sectors. Taking note of the risk associated with advanced persistent threats (APTs), it can be suggested that the incident management process needs to be given increased attention. In addition, staff involved in cyber security need to have the confidence to question management practices and lobby for changes in company policy so that improved cyber security occurs at the same time as cyber threat intelligence (CTI) is upgraded.

Bearing the above in mind, we can reflect on the individual stages of the intelligence cycle (IC) and the critical thinking process (CTP) [9] (p. 139) and suggest that cyber threat intelligence (CTI) should be merged into the cyber threat intelligence cycle process (CTICP) so that the following stages are visible: (1) objective resilience (e.g., top management define resilience so that the organization is able to withstand a range of cyber attacks); (2) question framing (e.g., top management establish how the organization is to be made more resilient through human action and the combined usage of AI and ML); (3) threat intelligence (e.g., managers define what is involved and map the identified impacts against possible outcomes); (4) work tasks established (e.g., individual managers and experts are appointed to undertake specific tasks and roles); (5) collection of threat intelligence data and information (e.g., various research and data collection exercises are undertaken but mostly utilize AI and ML); (6) the analysis of threat intelligence data and information (e.g., cause and effect established/patterns in the data are identified that indicate a certain type of attack is occurring/is likely to occur); (7) interpretation of the results (e.g., risk register(s) up-dated within the organization and partner organizations); (8) dissemination of the results (e.g., the cyber security manager liaises with government bodies/agencies, trade associations and various resilience community groups and shares relevant industry information); (9) cyber threat intelligence (CTI) concepts/frameworks/models devised (e.g., industry specific and improved through additional evidence from university research group(s); (10) strategic cyber security management (e.g., assumptions are incorporated into a new way of thinking about the role that cyber security management plays); (11) reflection (e.g., staff focus on how advances in AI and ML will change the nature of future cyber threat intelligence (CTI) analysis and interpretation); and (12) intelligence culture (e.g., promotional activity undertaken within the partnership arrangement and more widely to help people in society prepare for cyber attacks and develop their own level of cyber security awareness so that they are better able to handle the psychological consequences of such attacks).

The benefits of such an approach are clear to see. The cyber security manager and various managers throughout the organization and its partners can utilize sociocultural intelligence to gain a more strategic view of the nature of cyber threats and how various cybers attacks are to be unleashed. The advantage of formalizing cyber threat intelligence (CTI) as opposed to viewing it as ad hoc is clear to see. Cyber threat intelligence (CTI) can be viewed from several stances including allowing "early detection of malicious behavior, preferably before a malicious actor gains a foothold in the network" and aiding the sense-making process by providing "insight into the relevant threat environment to decisionmakers" [66] (p. 301). Cyber threat intelligence (CTI) can therefore improve situational awareness and focus attention on key concerns such as how to guard against bias. Bias originates from cyber threat intelligence (CTI) feeds and/or analysis and can be linked to both criminal groups and state actors [66] (pp. 309–310). Bias is associated with the process itself whereby poisoning attacks occur as a result of training data, derived from open-source platforms, being manipulated/contaminated by malicious actors [67,68].

The cyber threat intelligence cycle process (CTICP) can help managers to identify how malicious actors are targeting organizations and how they are identifying future targets. This is conducted through the cooperation of designated managers, a commitment to using quality data and appropriate data analysis tools, and the sharing of intelligence on malicious actors and their networks. It is also envisaged that a range of ethical concerns will need to be addressed including data privacy, integrity and the accuracy of predictive intelligence [69]. By incorporating ethical issues and concerns into the process, it should be possible for managers to view predictive intelligence from the perspective of the changing needs of society, maintaining individual privacy and meeting legal challenges as and when they occur. In addition, by embracing the sociocultural intelligence approach, the cyber security manager should be well placed to challenge and verify the patterns identified during the analysis of the big data.

9. Conclusions

For managers within an organization that are not familiar with AI to understand more fully what is involved when applying AI to help deal with cyber threats and to deal with cyber attacks when they occur, it is important to understand what cyber threat intelligence (CTI) is and how it feeds into strategic cyber security management [1]. Well-established intelligence concepts can be drawn on and modified to help the cyber security manager devise a cyber threat intelligence (CTI) blueprint that can be used to produce a more generic model or industry specific model, which is aimed at hardening the organization's defenses. By being committed to the use of situational analysis and embracing sociocultural intelligence inputs from external experts as well as in-house company staff, a security culture can be developed that has cyber security at the heart of it.

The advantage of placing cyber security at the center of security is that sociocultural intelligence can be reinforced by AI and in turn, AI can be monitored in terms of its ability to detect fake data and information and counter acts of data poisoning. The greater the quality of the data and the more sophisticated the process of analysis, the more the cyber security manager is able to work alongside colleagues to strengthen the organization's defenses. Through the process of integrating a number of separate but related tasks into a proactive stakeholder approach to cyber security management, the organization's supply chain will become more resilient and better able to withstand various forms of cyber attack.

10. Future Research

It is clear from the forgoing that a follow-up study can be undertaken that focuses more deeply on how AI/ML can enhance cyber security provision from the stance of a coordinated investment in cyber security from the organizations in a specific supply chain. This will provide insights into how organizations with a common trading mandate anticipate and guard against a possible cyber attack(s) and coordinate their defense [4]. The advantage of such a study is that it will provide evidence of a specific type of cyber threat intelligence (CTI) and outline how managers identify and organize supply chain resilience. Another research project that can be undertaken is to establish how managers overcome their lack of knowledge in relation to AI, and how they can develop relevant insights and/or contribute to the development of AI focused cyber security tools that lead to a better understanding of company–industry–society considerations and the need to ensure that AI is regulated appropriately [3] (p. 114). In addition, the insights into knowledge creation through various forms of learning [70] can be drawn on and placed more firmly in the context of managers understanding why network associations are important and how they can be developed through investment in AI.

It would also be appropriate to undertake research that contributes to cyber threat intelligence (CTI) methodology as this would help broaden the base of cyber threat intelligence (CTI) and solve a well-stated problem: "The volume and velocity with which new attacks are reported leads to a high daily influx of many single IoC datapoints that need further triangulation to assess their relevance to the specific threat context" [66] (p.304). Indeed, it should be possible to deploy soft systems methodology [71] and scenario planning [72] to link planning and modeling with strategy formulation and answer "what if" type questions that arise and once answered, enable initiatives in policy to be aimed at solutions to be found through learning.

Author Contributions: Conceptualization, Y.-I.L. and P.R.J.T.; Methodology, Y.-I.L. and P.R.J.T.; Formal analysis, Y.-I.L. and P.R.J.T.; Writing—original draft preparation, Y.-I.L. and P.R.J.T.; Writing—review and editing, Y.-I.L. and P.R.J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to express their gratitude to the reviewers for their in-depth comments and suggestions for improving the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Trim, P.R.J.; Lee, Y.-I. *Strategic Cyber Security Management*; Routledge: London, UK; New York, NY, USA, 2022.
2. Abraham, C.; Sims, R.R. A comprehensive approach to cyber resilience. *MIT Sloan Manag. Rev.* **2021**, *63*, 1–4.
3. Wirkuttis, N.; Klein, H. Artificial intelligence in cybersecurity. *Cyber Intell. Secur.* **2017**, *1*, 103–119.
4. Simon, J.; Omar, A. Cybersecurity investments in the supply chain: Coordination and a strategic attacker. *Eur. J. Oper. Res.* **2020**, *282*, 161–171. [CrossRef]
5. Rajan, R.; Rana, N.P.; Parameswar, N.; Dhir, S.; Sushil; Dwivedi, Y.K. Developing a modified total interpretive structural model (M-TISM) for organizational strategic cybersecurity management. *Technol. Forecast. Chang.* **2021**, *170*, 120872. [CrossRef]
6. Frith, C.D. The social brain? In *Social Intelligence: From Brain to Culture*; Emery, N., Clayton, N., Frith, C., Eds.; Oxford University Press: Oxford, UK, 2008; pp. 297–310.
7. Trim, P.R.J.; Lee, Y.-I. The Global Cyber Security Model: Counteracting cyber attacks through a resilient partnership arrangement. *Big Data Cogn. Comput.* **2021**, *5*, 32. [CrossRef]
8. Yamin, M.M.; Ullah, M.; Ullah, H.; Katt, B. Weaponized AI for cyber attacks. *J. Inf. Secur. Appl.* **2021**, *57*, 102722. [CrossRef]
9. Patton, K. *Sociocultural Intelligence: A New Discipline in Intelligence Studies*; The Continuum International Publishing Group: London, UK, 2010.
10. Hasan, K.; Shetty, S.; Ullah, S. Artificial intelligence empowered cyber threat detection and protection for power utilities. In Proceedings of the IEEE 5th International Conference on Collaboration and Internet Computing, Los Angeles, CA, USA, 12–14 December 2019; pp. 354–359.
11. Surya, L. An exploratory study of AI and Big Data, and it's future in the United States. *Int. J. Creat. Res. Thoughts* **2015**, *3*, 991–995.
12. Hagendorff, T.; Wezel, K. 15 Challenges for AI: Or what AI (currently) can't do. *AI Soc.* **2020**, *35*, 355–365. [CrossRef]
13. Gallese, V. Chapter 12: "Before and below 'theory of mind': Embedded simulation and the neural correlates of social cognition". In *Social Intelligence: From Brain to Culture*; Emery, N., Clayton, N., Frith, C., Eds.; Oxford University Press: Oxford, UK, 2008; pp. 279–296.
14. HSSAI. *Risk and Resilience: Exploring the Relationship*; Department of Homeland Security, Science and Technology Directorate: Arlington, MA, USA, 2010.
15. Argyris, C. *On Organizational Learning*; Blackwell Publishers Limited: Oxford, UK, 1996.
16. McCreight, R. Resilience as a goal and standard in emergency management. *J. Homel. Secur. Emerg. Manag.* **2009**, *7*, 1–7. [CrossRef]
17. Andrade, R.O.; Yoo, S.G. Cognitive security: A comprehensive study of cognitive science in cybersecurity. *J. Inf. Secur. Appl.* **2019**, *48*, 1–13. [CrossRef]
18. Dawson, S. *Analysing Organisations*; Palgrave: Basingstoke, UK, 1996.
19. Ma, L.; Sun, B. Machine learning and AI in marketing—Connecting computing power to human insights. *Int. J. Res. Mark.* **2020**, *37*, 481–504. [CrossRef]
20. Salakhutdinov, R.; Hinton, G. Deep Boltzmann machines. In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS), Clearwater Beach, FL, USA, 16–18 April 2009; pp. 2735–2742. [CrossRef]
21. Moerland, T.M.; Broekens, J.; Jonker, C.M. Emotion in reinforcement learning agents and robots: A survey. *Mach. Learn.* **2018**, *107*, 443–480. [CrossRef]
22. Jones, L. AI Trends in Retail, Retail & E-Commerce, 23 April. Available online: https://www.transperfect.com/blog/2021-ai-trends-retail (accessed on 15 June 2021).
23. Kohl's. 2020 Reimaging the Digital Shopping Experience with Snapchat. Available online: https://corporate.kohls.com/news/archive-/2020/august/reimagining-the-digital-shopping-experience-with-snapchat (accessed on 15 June 2021).
24. Roggeveen, A.L.; Grewal, D.; Karsberg, J.; Noble, S.M.; Nordfält, J.; Patrick, V.M.; Schweiger, E.; Soysal, G.; Dillard, A.; Cooper, N.; et al. Forging meaningful consumer-brand relationships through creative merchandise offerings and innovative merchandising strategies. *J. Retail.* **2021**, *97*, 81–98. [CrossRef]
25. Holzwarth, M.; Janiszewski, C.; Neumann, M.M. The influence of avatars on online consumer shopping behavior. *J. Mark.* **2006**, *70*, 19–36. [CrossRef]
26. Grewal, D.; Noble, S.M.; Roggeveen, A.L.; Nordfalt, J. The future of in-store technology. *J. Acad. Mark. Sci.* **2020**, *48*, 96–113. [CrossRef]

27. Roggeveen, A.L.; Sethuraman, R. Customer-interfacing retail technologies in 2020 & beyond: An integrative framework and research directions. *J. Retail.* **2020**, *96*, 299–309. [CrossRef]
28. Srikanth, A. Virtual Assistants vs Chatbots: What's the Differences & How to Choose the Right One? 2020, Freshdesk Blog. Available online: https://freshdesk.com/customer-engagement/virtual-assistant-chatbot-blog/ (accessed on 16 June 2021).
29. Croes, E.A.J.; Antheunis, M.L. Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *J. Soc. Pers. Relatsh.* **2021**, *38*, 279–300. [CrossRef]
30. Skjuve, M.; Følstad, A.; Fostervold, K.I.; Brandtzaeg, P.B. My chatbot companion—A study of human-chatbot relationships. *Int. J. Hum. Comput. Stud.* **2021**, *149*, 102601. [CrossRef]
31. Campbell, C.; Sands, S.; Ferraro, C.; Tsao, H.Y.; Mavrommatis, A. From data to action: How marketers can leverage AI. *Bus. Horiz.* **2020**, *63*, 227–243. [CrossRef]
32. Huang, M.-H.; Rust, R.T. Artificial intelligence in service. *J. Serv. Res.* **2018**, *21*, 155–172. [CrossRef]
33. Kitchens, B.; Dobolyi, D.; Li, J.; Abbasi, A. Advanced customer analytics: Strategic value through integration of relationship-oriented big data. *J. Manag. Inf. Syst.* **2018**, *35*, 540–574. [CrossRef]
34. Vollrath, M.D.; Villegas, S.G. Avoiding digital marketing analytics myopia: Revisiting the customer decision journey as a strategic marketing framework. *J. Mark. Anal.* **2022**, *10*, 106–113. [CrossRef]
35. Gupta, R. Deep Learning Models—When Should You Use Them? From ANN to AutoEncoders, Towards Data Science, 2019, October. Available online: https://towardsdatascience.com/6-deep-learning-models-10d20afec175 (accessed on 18 June 2021).
36. IBM Cloud Education. Recurrent Neutral Networks, 2020, 14 September. Available online: https://www.ibm.com/cloud/learn/recurrent-neural-networks (accessed on 14 June 2021).
37. Wu, H.; Prasad, S. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Trans. Image Process.* **2018**, *27*, 1259–1270. [CrossRef]
38. Ouali, Y.; Hudelot, C.; Tami, M. An Overview of Deep Semi-Supervised Learning. *arXiv* **2020**, arXiv:2006.05278. Available online: https://arxiv.org/abs/2006.05278 (accessed on 12 September 2022).
39. Manukian, H.; Pei, Y.R.; Bearden, S.R.B.; Di Ventra, M. Mode-Assisted unsupervised learning of restricted Boltzmann machines. *Commun. Phys.* **2020**, *3*, 105. [CrossRef]
40. Sakkari, M.; Zaied, M. A convolutional deep self-organizing map feature extraction for machine learning. *Multimed. Tools Appl.* **2020**, *79*, 19451–19470. [CrossRef]
41. Çelenk, U.; Ertuğrul, Ç.D.; Zontul, M.; Elçi, A.; Uçan, O. Dynamic Quota Calculation System (DQCS): Pricing and Quota Allocation of Telecom Customers via Data Mining Approaches. In *Handbook of Research on Contemporary Perspectives on Web-Based Systems*; Elçi, A., Ed.; IGI Global Publisher: Hershey, PA, USA, 2018. [CrossRef]
42. Hinton, G.E.; Sejnowski, T.J. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 19 June 1983; pp. 448–453.
43. IBM Cloud Education. Machin Leaning, 2020, 15 July. Available online: https://www.ibm.com/cloud/learn/machine-learning (accessed on 5 June 2021).
44. Gavrilova, Y. Artificial Intelligence vs. Machine Learning vs. Deep Learning: Essentials. 2020. Available online: https://serokell.io/blog/ai-ml-dl-difference (accessed on 15 June 2021).
45. Mnih, V.; Badia, A.P.; Mirza, M.; Harley, T.; Lillicrap, T.P.; Silver, D.; Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. *Int. Conf. Mach. Learn.* **2013**, *48*, 1928–1937. [CrossRef]
46. Gazit, M. The Fourth Generation of AI Is Here, and It's Called 'Artificial Intuitions' News, 3 September 2020. Available online: https://thenextweb.com/news/the-fourth-generation-of-ai-is-here-and-its-called-artificial-intuition (accessed on 17 June 2021).
47. Vector ITC. Fourth Generation of AI Arrives: Artificial Intuition, Vector ITC, 1 February. Available online: https://www.vectoritcgroup.com/en/tech-magazine-en/artificial-intelligence-en/fourth-generation-of-ai-arrives-artificial-intuition/ (accessed on 17 June 2021).
48. Yalçın, Ö.N.; DiPaola, S. Modeling Empathy: Building a Link between Affective and Cognitive Processes. *Artif. Intell. Rev.* **2020**, *53*, 2983–3006. [CrossRef]
49. Pizzi, G.; Scarpi, D.; Pantano, E. Artificial intelligence and the new forms of interaction: Who has the control when interacting with a chatbot? *J. Bus. Res.* **2020**, *129*, 878–890. [CrossRef]
50. Bresniker, K.; Gavrilovska, A.; Holt, J.; Milojicic, D.; Tran, T. Grand challenge: Applying artificial intelligence and machine learning to cybersecurity. *Computer* **2019**, *52*, 45–52. [CrossRef]
51. Albright, J. Welcome to the era of fake news. *Media Commun.* **2017**, *5*, 87–89. [CrossRef]
52. Petratos, P.N. Misinformation, disinformation, and fake news: Cyber risks to business. *Bus. Horiz.* **2021**, *64*, 763–774. [CrossRef]
53. Tatar, U.; Nussbaum, B.; Gokce, Y.; Keskin, O.F. Digital force majeure: The Mondelez case, insurance, and the (un)certainty of attribution in cyberattacks. *Bus. Horiz.* **2021**, *64*, 775–785. [CrossRef]
54. Sinkovics, R.R.; Penz, E. Multilingual elite-interviews and software-based analysis: Problems and solutions based on CAQDAS. *Int. J. Mark. Res.* **2011**, *53*, 705–724. [CrossRef]
55. Easterby-Smith, M.; Thorpe, R. Research traditions in management learning. In *Management Learning: Integrating Perspectives in Theory and Practice*; Burgoyne, J., Reynolds, M., Eds.; Sage Publications: London, UK, 1997; pp. 38–53.
56. Patton, M.Q. *Qualitative Evaluation and Research Methods*; Sage Publications: London, UK; New Delhi, India, 1990.

57. Woods, P. Symbolic interaction: Theory and method. In *The Handbook of Qualitative Research in Education*; LeCompte, M.D., Millroy, W.L., Preissle, J., Eds.; Academic Press, Inc.: San Diego, CA, USA, 1992; pp. 337–404.
58. Frey, J.H.; Fontana, A. The group interview in social research. *Soc. Sci. J.* **1991**, *28*, 175–187. [CrossRef]
59. DiCicco-Bloom, B.; Crabtree, B.F. The qualitative research interview. *Med. Educ.* **2006**, *40*, 314–321. [CrossRef]
60. Strauss, A.; Corbin, J. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*; Sage Publications: London, UK, 1998.
61. Pawlicka, A.; Choraś, M.; Pawlicki, M.; Kozik, R. A$10 million question and other cybersecurity-related ethical dilemmas amid the COVID-19 pandemic. *Bus. Horiz.* **2021**, *64*, 729–734. [CrossRef]
62. Carson, B.; Chakravarty, A.; Koh, K.; Thomas, R. *Platform Operating Model for the AI Bank of the Future*; Mckinsey & Company: London, UK, 2021; pp. 1–11.
63. Zhuang, Y.-T.; Wu, F.; Chen, C.; Pan, Y.-H. Challenges and opportunities: From big data to knowledge in AI2.0. *Front. Inf. Technol. Electron. Eng.* **2017**, *18*, 3–14. [CrossRef]
64. Meland, P.H.; Nesheim, A.A.; Bernsmed, K.; Sindre, G. Assessing cyber threats for storyless systems. *J. Inf. Secur. Appl.* **2022**, *64*, 103050. [CrossRef]
65. Settanni, G.; Skopik, F.; Shovgenya, Y.; Fiedler, R.; Carolan, M.; Conroy, D.; Boettinger, K.; Gall, M.; Brost, G.; Ponchel, C.; et al. A collaborative cyber incident management system for European interconnected critical infrastructures. *J. Inf. Secur. Appl.* **2017**, *34*, 166–182. [CrossRef]
66. Oosthoek, K.; Doerr, C. Cyber threat intelligence: A product without a process? *Int. J. Intell. Count.* **2021**, *34*, 300–315. [CrossRef]
67. Ranade, P.; Piplai, A.; Mittal, S.; Joshi, A.; Finin, T. Generating fake cyber threat intelligence using transformer-based models. In Proceedings of the International Joint Conference on Neural Networks, IEEE, Shenzhen, China, 18–22 July 2021; pp. 1–9. [CrossRef]
68. Khurana, N.; Mittal, S.; Joshi, A. Preventing poisoning attacks on AI based threat intelligence systems. In Proceedings of the IEEE 29 International Workshop on Machine Learning for Signal Processing Conference, IEEE, Pittsburgh, PA, USA, 13–16 October 2019. [CrossRef]
69. Tilimbe, J. Ethical implications of predictive risk intelligence. *ORBIT J.* **2019**, *2*, 1–28. [CrossRef]
70. Stella, M.; Kenett, Y.N. (Eds.) *Knowledge Modelling and Learning through Cognitive Networks*; MDPI: Basel, Switzerland, 2022. [CrossRef]
71. Checkland, P.; Scholes, J. *Soft Systems Methodology in Acton*; John Wiley & Sons: Chichester, UK, 2007.
72. Ringland, G. *Scenario Planning*; John Wiley & Sons: Chichester, UK, 2006.

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
www.mdpi.com

Big Data and Cognitive Computing Editorial Office
E-mail: bdcc@mdpi.com
www.mdpi.com/journal/bdcc

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-0365-9645-7