

Advanced Sciences and Technologies for Security Applications

Hamid Jahankhani *Editor*

Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs

Proceedings of the 15th International
Conference on Global Security, Safety
and Sustainability, London,
October 2023

 Springer

Advanced Sciences and Technologies for Security Applications

Editor-in-Chief

Anthony J. Masys, Associate Professor, Director of Global Disaster Management, Humanitarian Assistance and Homeland Security, University of South Florida, Tampa, USA

Advisory Editors

Gisela Bichler, California State University, San Bernardino, CA, USA

Thirimachos Bourlai, Department of Computer Science and Electrical Engineering, Multispectral Imagery Lab (MILab), West Virginia University, Morgantown, WV, USA

Chris Johnson, University of Glasgow, Glasgow, UK

Panagiotis Karampelas, Hellenic Air Force Academy, Attica, Greece

Christian Leuprecht, Royal Military College of Canada, Kingston, ON, Canada

Edward C. Morse, University of California, Berkeley, CA, USA

David Skillicorn, Queen's University, Kingston, ON, Canada

Yoshiki Yamagata, National Institute for Environmental Studies, Tsukuba, Ibaraki, Japan

Indexed by SCOPUS

The series *Advanced Sciences and Technologies for Security Applications* comprises interdisciplinary research covering the theory, foundations and domain-specific topics pertaining to security. Publications within the series are peer-reviewed monographs and edited works in the areas of:

- biological and chemical threat recognition and detection (e.g., biosensors, aerosols, forensics)
- crisis and disaster management
- terrorism
- cyber security and secure information systems (e.g., encryption, optical and photonic systems)
- traditional and non-traditional security
- energy, food and resource security
- economic security and securitization (including associated infrastructures)
- transnational crime
- human security and health security
- social, political and psychological aspects of security
- recognition and identification (e.g., optical imaging, biometrics, authentication and verification)
- smart surveillance systems
- applications of theoretical frameworks and methodologies (e.g., grounded theory, complexity, network sciences, modelling and simulation)

Together, the high-quality contributions to this series provide a cross-disciplinary overview of forefront research endeavours aiming to make the world a safer place.

The editors encourage prospective authors to correspond with them in advance of submitting a manuscript. Submission of manuscripts should be made to the Editor-in-Chief or one of the Editors.

Hamid Jahankhani
Editor

Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs

Proceedings of the 15th International
Conference on Global Security, Safety
and Sustainability, London, October 2023

 Springer

Editor
Hamid Jahankhani
Northumbria University
London, UK

ISSN 1613-5113 ISSN 2363-9466 (electronic)
Advanced Sciences and Technologies for Security Applications
ISBN 978-3-031-47593-1 ISBN 978-3-031-47594-8 (eBook)
<https://doi.org/10.1007/978-3-031-47594-8>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Committee

General Chair

Prof. Hamid Jahankhani, UK
hamid.jahankhani@northumbria.ac.uk

Steering Committee

Prof. Shaun Lawson, Northumbria University, UK
Dr. Julie Walters, Northumbria University, UK
Dr. Arshad Jamal, QAHE and Northumbria University London, UK
Dr. Reza Montasari, University of Swansea, UK
Prof. Amin Hosseinian-Far, University of Northampton, UK
Prof. Mohammad Dastbaz, Deputy Vice-Chancellor, University of Suffolk, UK
Dr. John McCarthy, Director, Oxford Systems, UK
Prof. Ali Hessami, Director Vega Systems, UK
Dr. Sina Pournouri, Sheffield Hallam University, UK
Dr. Rose Cheuk Wai, QAHE and Northumbria University London, UK
Mr. Usman J. Butt, QAHE and Northumbria University London, UK
Prof. Babak Akhgar, Director of CENTRIC, Sheffield Hallam University, UK
Prof. Gianluigi Me, Luiss University, Italy
Prof. Bobby Tait, University of South Africa, South Africa
Dr. John McCarthy, Director, Oxford Systems, UK

Programme Committee

Dr. Omar S. Arabiat, Balqa Applied University, Jordan
Konstantinos Kardaras, Technical Consultant, Greece
Dr. Chafika Benzaid, Computer Security Division, CERIST, University of Sciences and Technology Houari Boumediene, Algeria
Dr. Ken Dick, Nebraska University Centre for Information Assurance, USA
Dr. Elias Pimenidis, University of West of England, UK
UK Professor Murdoch Watney, University of Johannesburg, South Africa
Dr. Mamoun Alazab, Macquarie University, Australia
Dr. Gregory Epiphaniou, Associate Professor of Security Engineering at University of Warwick, UK
Dr. Haider Alkhateeb, Wolverhampton Cyber Research Institute, University of Wolverhampton, UK
Dr. Abul Salih, Northumbria University London, UK
Dr. Carlisle George, Middlesex University, UK
Prof. Gianluigi Me, University of Rome LUISS “Guido Carli”, Italy
Dr. Ameer Al-Nemrat, University of East London, UK
Dr. Nasr Abosata, Northumbria University London, UK
Mrs. Sukhvinder Hara, Middlesex University, UK
Dr. Ian Mitchell, Middlesex University
Dr. Ayman El Hajjar, Westminster University, UK
Mr. Rejwan Bin Sulaiman, Northumbria University London, UK
Dr. George Weir, University of Strathclyde, UK
Dr. Sufian Yousef, Anglia Ruskin University, UK
Dr. Alireza Daneshkhah, Coventry University, UK
Dr. Maryam Farsi, Cranfield University, UK
Mr. Stefan Kendzierskyj, CYFORTIS, UK

About the Conference

2020 will be remembered as a year of COVID-19 pandemic year with a profound impact on our lives and challenging times for governments around the world with technology as a backbone of tools to assist us to work remotely while in self isolations and in total lockdown. Every individual and organization are faced with the challenges of the adoption of technological change and understand that nothing will be the same for some time after and even when normality resumes. AI, blockchain, machine learning, IoT, IoMT, and IIoT all are tested and applied at a very rapid turnaround as we go through the pandemic day after day. However, all this transformation can bring many associated risks regarding data, such as privacy, transparency, exploitation, and ownership.

The advancement of Artificial Intelligence (AI), coupled with the prolificacy of the Internet of Things (IoT) devices is creating smart societies that are interconnected. The expansion of Big Data and volumetric metadata generated and collated by these mechanisms not only gives greater depth of understanding but as change is coming at an unprecedented pace COVID-19 is driving the cultural changes and the public sector innovations to a new height of responsibilities.

Space exploration and satellite, drone, and UAV technology have travelled a long way in recent years and some may debate that we are in the midst of a revolution; in terms of development and the increasing number of these devices being launched. But with this revolutionary progress, it presents itself with new challenges in terms of governance. For example, as satellites sit in a global sphere it makes space governance a complicated dynamic with many entities and stakeholders contributing to their own requirements and not necessarily taking into account all aspects or other factors. It opens up many questions regarding control, data, privacy, security, and so on. Therefore, innovative solutions for a sustainable future into the pressing issues facing space governance today are important in terms of (i) what the current challenges are (ii) how do these devices impact society, and (iii) what the potential negative consequences.

Further development and analysis of defence tactics and countermeasures are also needed fast to understand security vulnerabilities which will be exploited by cyber-criminals by identifying newer technologies that can aid in protecting and securing data with a range of supporting processes to provide a higher level of cyber resilience.

The ethical implications of connecting the physical and digital worlds, and presenting the reality of a truly interconnected society, present the realization of the concept of smart societies in reality.

This Annual International Conference is an established platform in which security, safety, and sustainability issues can be examined from several global perspectives through dialogue between academics, students, government representatives, chief executives, security professionals, and research scientists from the United Kingdom and from around the globe.

The TWO-days Virtual conference will focus on the challenges of complexity, the rapid pace of change, and risk/opportunity issues associated with the 21st-century lifestyle, systems, and infrastructures.

Prof. Hamid Jahankhani
General Chair of the Conference
September 2023

Contents

Strengthening Security Mechanisms of Satellites and UAVs Against Possible Attacks from Quantum Computers	1
Osama A. A. M. Hussien, Isuru S. W. Arachchige, and Hamid Jahankhani	
Evaluating the Security of Open-Source Linux Operating Systems for Unmanned Aerial Vehicles	21
Darshan Patil and Sina Pournouri	
A Proposed Permissioned Blockchain Consensus Algorithm: Consensus Algorithm Genetically Enhanced (CAGE)	51
Ian Mitchell and Kamil Maka	
Data Hiding in Anti-forensics—Exploit Delivery Through Digital Steganography	65
Hans Gashi, Shahrzad Zargari, and Setareh Jalali Ghazaani	
Impact Versus Frequency on Cybersecurity Breach Trends in the Business and Medical Industry to Identify Human Error	77
Galathara Kahanda, Sasha Rider, and Sayantini Mukhopadhyay	
Identifying Daesh-Related Propaganda Using OSINT and Clustering Analysis	97
Gianluigi Me and Maria Felicita Mucci	
Cyberbullying Detection in Twitter Using Deep Learning Model Techniques	147
Anu Ranjana Seetharaman and Hamid Jahankhani	
A Signcryption Approach to Address Security and Privacy Issues in Online Gaming Platforms	169
Mahmut Ahmet Unal and Tasmia Islam	

ChildSecurity: A Web-Based Game to Raise Awareness of Cybersecurity and Privacy in Children 187
 Yang Zou and Tasmina Islam

Subverting Biometric Security Using 3D Printed Biometrics Characteristics 203
 Bobby L. Tait

Image Band-Distributive PCA Based Face Recognition Technique 219
 Henry Ndu, Akbar Sheikh-Akbari, Iosif Mporas, and Jiamei Deng

RAFA Model. Rethinking Cyber Risk Management in Organizations 231
 Jeimy J. Cano M

AI and Ethics: Embedding Good Aspects of AI 245
 Gordon Bowen, Deidre Bowen, and Lisa Bamford

Cyber Security-Related Awareness in Bangladesh: Relevant Challenges and Strategies 259
 Kudrat-E-Khuda

Credit Card Fraud Detection Using Machine Learning 275
 Berlin Srojila Manickam and Hamid Jahankhani

The Role of Trust and Ethics in IT Administration and Its Impact on an Organisation 307
 Bijay Sunny George and Tasmina Islam

How Explainable Artificial Intelligence (XAI) Models Can Be Used Within Intrusion Detection Systems (IDS) to Enhance an Analyst’s Trust and Understanding 321
 Chelsea Shand, Rose Fong, and Usman Butt

Could Adam Smith Live in a Smart City? 343
 Ian Toft and Tasmina Islam

An Analysis of Blockchain Adoption in Zimbabwean Mining Land Title Management (MLTM) Using NVivo 363
 David V. Kilpin, Eustathios Sainidis, Hamid Jahankhani, and Guy Brown

Cybersecurity at the Core: A Study on IT Experts’ Policy Adherence 387
 Rao Faizan Ali, Hamid Jahankhani, and Bilal Hassan

Metaverse Application Forensics: Unravelling the Virtual Truth 399
 M. A. Hannan Bin Azhar and Oliver Rush-Gadsby

**Cloud-Based Secure Electronic Medical Data Sharing System
Using Blockchain Technology (Simulation of a Ransomware
Attack with OWASP) 415**
Rodrigue Ngomsu and Hamid Jahankhani

Strengthening Security Mechanisms of Satellites and UAVs Against Possible Attacks from Quantum Computers



Osama A. A. M. Hussien, Isuru S. W. Arachchige, and Hamid Jahankhani

Abstract The mounting prevalence of quantum computing poses a threat not only to conventional but also to futuristic network systems. As a result, today the focus shifts towards strengthening existing networks to secure sensitive data that may be threatened as a result of the potential for breaking present encryption protocols with the introduction of quantum computing. Cryptography, commonly referred to as encryption, is the underlying principle that enables safe data storage and communication. Cryptography involves key distribution, for which the current best method is a public key distribution based on principles such as Diffie-Hellman key exchange. However, these principles rely on classical computer limitations, such as factoring big numbers, which are expected to be eliminated with quantum computers and algorithms. This research provides an overview of using Quantum Key Distribution (QKD) protocols for satellite-based and Unmanned-Aerial-Vehicle (UAV) related communication. This research also proposes a none QKD protocol to provide an understanding of analysis of the advantages and applications of these protocols. As several QKD protocols have been proposed in the past, identifying the advantages and unique performance characteristics of these protocols is important, especially when adapting these protocols to specific use cases such as satellite-based communications and UAV-related communications.

Keywords QKD · Satellite · UAV · Secure-communication

1 Introduction

Quantum Computing is a cutting-edge technology with applications ranging from quantum physics to biology to economics. Modern-day telecommunication and information security rely on classical encryption algorithms that utilize the computational limitations of existing classical computers. However, quantum computing is

O. A. A. M. Hussien · I. S. W. Arachchige · H. Jahankhani (✉)
Northumbria University, London, UK
e-mail: hamid.jahankhani@northumbria.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_1

a significant threat to information security because of its potential ability to break existing encryption algorithms. Quantum algorithms, such as Shor's algorithm and Grover's search algorithm, create a significant potential threat to classical cryptography. Quantum computing in general is expected to become a much more significant threat to classical cryptography in the coming years, [1].

Even though quantum computing has created a significant threat to information security, it has also introduced several novel solutions; quantum cryptography and quantum communication. Quantum cryptography is based on the fundamental physics principles that govern quantum particles and their interactions and uses the same principles to provide a secure way of communication and encryption. Quantum physics principles such as superposition, uncertainty principle, quantum entanglement, coherence and no-cloning theorem enable the development of secure quantum communication channels and methods for generating secret data encryption keys, [2].

Secure communication is an important aspect of satellite systems and unmanned aerial vehicles (UAVs) due to their various applications in military operations, financial transactions, and sensitive data transmission. Satellites provide essential services such as telecommunications, weather monitoring, and global positioning, while UAVs are employed in applications ranging from surveillance and reconnaissance to package delivery and disaster response. However, these systems face significant security challenges due to their wireless nature, vulnerability to interception or tampering, and the potential for unauthorized access to sensitive information. This is where QKD protocols come into play. With their ability to ensure information security based on the fundamental principles of quantum mechanics, QKD protocols offer a robust solution for secure communication in satellites and UAVs. Moreover, the use of QKD protocols in satellite communication and UAVs is not limited to secure data transmission. These protocols can also be used for secure key exchange, allowing for the establishment of secret keys between remote parties without the risk of interception or eavesdropping, [3].

2 Literature Review

Quantum computing has the potential to break current encryption schemes, such as Rivest-Shamir-Adleman (RSA) as well as Elliptic Curve Cryptography (ECC) [4], which are widely used to secure communication over the internet. Post-quantum cryptography is an emerging field that aims to develop encryption schemes which can resist to quantum attacks. The development of quantum computing raises severe concerns about its potential power to break most standard cryptographic methods. Therefore, post-quantum cryptography has emerged as a promising solution because it seeks to develop cryptographic techniques resistant to quantum computer attacks.

Among other cryptography systems, Quantum Key Distribution (QKD), proposed in the 1980s, is one of the world's most matured quantum technologies. Modern QKD methods are based on two primary methods that were proposed the BB84 protocol

and the E91 protocol, [5]. These two protocols have created the two schemes that all QKD protocols are based on, the prepare-and-measure-based scheme and the entanglement-based scheme. Following the discovery of these two protocols, several other proposals for different QKD protocols emerged. Several of these protocols have been shown in real-world implementations to establish their viability for the future of quantum cryptography, [6].

Blockchains, which serve as decentralized and immutable ledgers for various applications, rely on cryptographic algorithms to secure transactions and protect sensitive data. However, powerful quantum computers pose a significant threat to these classical cryptographic algorithms, potentially compromising the security of blockchain networks. QKD presents an intriguing solution to enhance the security of blockchain networks in the face of quantum computing advancements. While there are practical challenges to overcome, ongoing research and innovation in the integration of QKD in blockchain hold promise for establishing robust security mechanisms that can withstand the potential threats posed by quantum computers, [7].

2.1 Quantum Key Distribution (QKD)

Cryptography, commonly referred to as encryption, is the underlying principle that enables safe data storage and communication. Cryptography involves key distribution, for which the current best method is a public key distribution based on principles such as Diffie-Hellman key exchange. However, one of the key vulnerabilities of classical cryptographic algorithms is their reliance on mathematical problems that can be efficiently solved by quantum computers. Quantum computers have the potential to break commonly used cryptographic techniques, such as asymmetric key encryption and digital signatures, which are the building blocks of blockchain security. QKD can address this vulnerability by providing a method to securely exchange cryptographic keys that are resistant to quantum attacks. Therefore, both governments [8] and corporations [9] have prioritized the development of quantum-safe cryptography methods.

QKD offers a secure and reliable way to generate and distribute encryption keys between the participants of a blockchain network. QKD protocols ensure that any attempt to eavesdrop on the key exchange process will be detectable by leveraging the principles of quantum mechanics, thereby guaranteeing the integrity and confidentiality of the keys. The exchanged keys can then be used to encrypt the data, protecting them from unauthorized access.

At the core of QKD lies the concept of quantum superposition, which allows quantum states to exist in multiple states simultaneously. This property enables the transmission of information encoded in quantum bits, or qubits, that can be in a superposition of 0 and 1. Moreover, the phenomenon of quantum entanglement plays a crucial role in QKD. Entanglement allows for a strong correlation between two or more particles, such that the state of one particle is directly linked to the state of the others, regardless of the physical distance between them. This property ensures

that any attempt to intercept or measure the transmitted qubits will be detected, as it would collapse the entangled states.

QKD protocols serve as the operational frameworks for implementing secure key distribution using quantum states. The most well-known and widely studied QKD protocol is the BB84 protocol, introduced by Bennett and Brassard in 1984 [10]. BB84 employs two mutually orthogonal bases to encode the key information onto the qubits, ensuring that any unauthorized measurement will introduce errors that can be detected by the communicating parties. This protocol forms the foundation for many QKD implementations and serves as a benchmark for evaluating the performance of other protocols.

Another notable QKD protocol is the E91 protocol, proposed by Ekert in 1991, [11]. E91 utilizes quantum entanglement to establish a secure key between two distant parties. It relies on Bell's inequalities to verify the integrity of the entangled states and ensure secure communication. This protocol has the advantage of allowing secure key distribution over longer distances, making it suitable for applications that involve communication between remote locations.

Continuous variable QKD protocols, such as the SARG04 protocol proposed by Scarani et al. in 2004, exploit the properties of continuous variables, such as the amplitude and phase of electromagnetic waves, to encode and transmit quantum information [12]. These protocols are based on the principles of quantum optics and offer advantages in terms of high key rates and compatibility with existing optical communication systems.

Several other QKD protocols have been proposed, each with its unique features, advantages, and limitations. These include protocols based on weak coherent pulses, phase encoding, time-bin encoding, and measurement-device-independent (MDI) QKD. The choice of QKD protocol depends on various factors, including the desired level of security, the resources available, the physical constraints of the communication channel, and the specific application requirements.

Ongoing research and development in the field of QKD focus on advancing the efficiency, security, and practical implementation of these protocols. Areas of exploration include the development of error correction and privacy amplification techniques to improve the reliability of QKD systems, the investigation of quantum repeaters to extend the range of secure communication, and the exploration of device-independent QKD, which aims to ensure security even in the presence of potential loopholes in the implementation.

QKD represents a ground-breaking approach to secure communication, utilizing the principles of quantum mechanics to establish secure cryptographic keys. QKD protocols, such as BB84, E91, and continuous variable protocols, provide the frameworks for implementing secure key distribution using quantum states. These protocols offer unparalleled security and have the potential to revolutionize secure communication in various domains, ranging from telecommunications and financial systems to government and military applications. Ongoing research continues to push the boundaries of QKD, making it a promising technology for the future of secure communication.

2.2 *Post Quantum Network Security Layer*

The Transport Layer Security (TLS) protocol is widely used to secure communications over the Internet. However, the security of TLS relies on public-key cryptography algorithms that are vulnerable to quantum attacks. Therefore, there is a need to design and implement post-quantum TLS (PQ-TLS) protocols that can resist quantum adversaries.

One of the main challenges in designing PQ-TLS protocols is to achieve both security and efficiency. Post-quantum cryptography algorithms typically have larger key sizes and higher computational costs than their classical counterparts, which can affect the performance and bandwidth of TLS connections. Moreover, PQ-TLS protocols need to be compatible with existing TLS infrastructure and applications, which may not support post-quantum algorithms natively, this research proposes a novel PQ-TLS protocol that aims to address these challenges by re-creating a TLS handshake and transport layer, such that the new transport layer will be very similar to the traditional one with few exceptions for example, the destination port (dport) and source port (sport) fields in the TCP header have a bigger field length to support encrypted port numbers, which can prevent port scanning attacks by quantum adversaries, it will also makes use of the reserved bit field in the TCP header as an indicator of PQ-TLS support. When two devices perform a TCP handshake, if the client sends a reserved bit field with a specific value as well as dport set to 0, the server will recognize that the client supports PQ-TLS and proceed accordingly, it will also uses post-quantum algorithms such as NTRU for key exchange, xPoly20chacha1205 for symmetric encryption, FALCON for signatures, and harakav2 for hashing. These algorithms are chosen based on their security, efficiency, and compatibility with existing standards and implementations.

Our protocol is inspired by three recent papers that propose different approaches to PQ-TLS, one of them is Post-Quantum TLS Without Handshake Signatures by Schwabe et al. [13] presents a PQ-TLS protocol that eliminates the use of digital signatures in the TLS handshake, which are typically expensive and require large public keys. Instead, the protocol uses a combination of key encapsulation mechanisms (KEMs) and hash-based signatures to achieve authentication and key agreement. The protocol also supports hybrid modes that combine post-quantum and classical algorithms for backward compatibility and transitional security.

Another approach to this issue was the Post-Quantum Key Exchange for the TLS Protocol from the Ring Learning with Errors Problem by Bos et al. [14] which introduces a PQ-TLS protocol that uses ring learning with errors (RLWE) as the underlying hard problem for key exchange. RLWE is a variant of the learning with errors (LWE) problem that offers smaller key sizes and faster computations than other lattice-based schemes. The protocol also supports hybrid modes and is compatible with existing TLS 1.2 implementations.

And the researchers at Prototyping post-quantum and hybrid key exchange and authentication in TLS and SSH by Crockett et al. [15] describes an experimental

implementation of PQ-TLS and post-quantum SSH protocols based on various post-quantum algorithms, such as NTRU, NewHope, SIDH, Kyber, Dilithium, Picnic, and SPHINCS+. The paper evaluates the performance and security of these protocols in comparison with classical ones and provides recommendations for future standardization efforts.

The three papers that inspire our protocol have made significant contributions to the field of PQ-TLS, but they also have some limitations and challenges that need to be addressed, the paper by Schwabe et al. [13] relies on the assumption that hash-based signatures are secure against quantum attacks, which may not hold in the long term as quantum computers can potentially break some of the properties of hash functions that are essential for the security of hash-based signatures, such as preimage resistance. Preimage resistance means that it is hard to find an input that produces a given output of a hash function. Quantum computers can use quantum algorithms, such as Grover's algorithm, to search a large space of inputs faster than classical algorithms, and thus find preimages or collisions for hash functions more easily than classical adversaries. This can allow quantum adversaries to forge hash-based signatures by finding inputs that match the public keys or the signatures of the legitimate parties. Therefore, hash-based signatures need to use hash functions that are carefully designed and analyzed for quantum resistance, and that have large output sizes and strong internal structures. Some papers that discuss this issue, which present new attacks and schemes for hash-based signatures under quantum scenarios. Moreover, the paper [15] does not provide a formal security analysis or proof of their protocol, which may leave some security gaps or vulnerabilities.

The paper by Bos et al. [14] uses RLWE as the basis for their key exchange scheme, which is a relatively new and untested problem in cryptography. There may be potential attacks or weaknesses that are not yet discovered or exploited by quantum adversaries, and lastly the paper by Crockett et al. [15] provides a valuable experimental evaluation of various post-quantum algorithms in TLS and SSH, but it does not propose a specific PQ-TLS protocol or design but it also does not compare the performance and security of their implementation with other existing or proposed PQ-TLS protocols.

Our protocol differs from these papers in several aspects, as we propose a new way of indicating PQ-TLS support using the reserved bit field in the TCP header, which can avoid modifying the TLS record layer or introducing new extensions or cipher suites as well as that we use encrypted port numbers to enhance the privacy and security of PQ-TLS connections against quantum adversaries who can intercept or manipulate network traffic and our selection of different set of post-quantum algorithms that are based on different hard problems and have different trade-offs in terms of security and efficiency ensures the security of the communication.

Our protocol can offer a viable solution for securing TLS communications in a post-quantum world, while maintaining compatibility and performance with existing systems as it aims to overcome some of these limitations and challenges by proposing a novel PQ-TLS protocol that uses a different set of post-quantum algorithms and techniques, and by providing a preliminary security analysis and proof of concept implementation.

One of the applications of blockchain that can benefit the new TLS protocol is self-sovereign identity (SSI), which is a decentralized approach to digital identity management. SSI empowers users to control their own identity data and credentials, without relying on centralized authorities or intermediaries. SSI uses decentralized identifiers (DID), which are unique identifiers that can be resolved to verifiable credentials stored on a blockchain [16].

By combining SSI, DID and post quantum encryption, blockchain networks can create a secure and privacy-preserving infrastructure for digital identity management. This can enable users to prove their identity and attributes in various domains, such as e-commerce, health care, education and government services [16].

Satellite communication is a vital technology for global connectivity, navigation, surveillance and remote sensing. However, satellite communication also faces various security challenges, such as eavesdropping, jamming, spoofing and denial-of-service attacks. Moreover, satellite communication is vulnerable to quantum computing, which could break some of the current encryption methods used by satellite networks [17].

To address these challenges, some researchers and developers are exploring the use of blockchain (SSI and DID) and post quantum encryption in creating secure satellite communication. Blockchain is a distributed ledger technology that enables peer to peer transactions without intermediaries. It relies on cryptographic algorithms to ensure the integrity and authenticity of data. SSI is a decentralized approach to digital identity management that empowers users to control their own identity data and credentials. DID are unique identifiers that can be resolved to verifiable credentials stored on a blockchain [16].

Some researchers and developers are exploring the use of QKD and blockchain in securing space satellites. QKD is a secure communication technique that uses quantum properties of photons to encrypt secret keys that can be shared by two parties [18].

By combining blockchain (SSI and DID) and quantum resistant encryption, in network communication can create a secure and privacy-preserving infrastructure for data transmission and authentication. This can enable satellite and other networks to resist quantum attacks, prevent unauthorized access, verify data sources and destinations, and enhance network performance [16, 19, 20].

3 QKD for Satellites, Drones and UAVs

Satellites and UAVs have become increasingly important in many fields. Satellites play an essential role in our everyday lives and contribute to our well-being by making it possible for us to meet many important needs and challenges on Earth. They are useful in many fields such as Earth observation, communications, navigation, and science [21].

UAVs are being used for an expanding variety of applications ranging from consumer recreation flights to various military needs, crop monitoring, rail road

inspection, etc., with aircraft sizes from several centimetres to several tens of meters. UAVs have several advantages over satellites. They are less dependent on weather conditions and can hover below the lower cloud boundary and take pictures of the terrain even in the presence of cloudiness [22].

Secure communication for satellites and UAVs is critical due to their increased usage in our day-to-day lives. One approach to achieve this in the post-quantum era is through the implementation of QKD. QKD offers information-theoretic security, based on fundamental quantum properties, which cannot be achieved using widely-used classical methods.

In the context of satellite communication, secure links between ground stations and satellites are essential to protect sensitive data transmission. QKD can provide a reliable method for generating and exchanging encryption keys over long distances, even in the presence of potential eavesdroppers [23]. QKD protocols, such as BB84 or entanglement-based protocols, enable the generation and distribution of quantum states that are used to establish secure cryptographic keys. By leveraging quantum properties, such as the no-cloning theorem and quantum entanglement, QKD protocols can detect any unauthorized interception or tampering attempts, ensuring the confidentiality and integrity of the transmitted data.

Similarly, in the realm of UAVs, secure communication is vital to safeguard sensitive information and maintain the integrity of data transmission. UAVs are often deployed in environments with wireless vulnerabilities, making them susceptible to interception or manipulation by malicious actors. By integrating QKD into UAV systems, secure communication channels can be established between the UAV and ground stations or between multiple UAVs. QKD provides a means to encrypt the transmitted data using quantum keys, ensuring that only authorized parties can access and decrypt the information. This prevents unauthorized interception, tampering, or spoofing attacks, guaranteeing the confidentiality, integrity, and authenticity of the communicated data.

Implementing secure communication for satellites and UAVs, however, presents its own set of challenges. For satellite communication, factors such as atmospheric effects, long-distance transmission, synchronization, and pointing and tracking accuracy need to be addressed to ensure reliable QKD performance. Researchers and engineers have proposed solutions such as adaptive optics systems to mitigate atmospheric turbulence and advanced synchronization techniques to achieve stable QKD operations in satellite links.

In the case of UAVs, wireless vulnerabilities, physical exposure, and limited onboard resources pose challenges for secure communication. Miniaturization of QKD devices, optimized protocols, and efficient key management schemes are areas of active research to enable the integration of QKD into UAV platforms. Furthermore, power consumption, size, weight, and cost considerations are crucial for the practical deployment of QKD systems in UAVs.

3.1 Difficulties in Implementing QKD Protocols

Listed below are the three most popular QKD protocol types and the practical challenges of implementing these protocols.

3.1.1 Measurement-Based QKD

Hardware Requirements: Measurement-based protocols require specialized hardware components, such as single-photon sources and detectors, which can be challenging and expensive to develop and deploy.

Synchronization: Synchronizing the timing of quantum states and measurements over long distances or in the presence of network delays can be technically demanding.

Vulnerability to Side-Channel Attacks: Measurement-based protocols such as BB84 is susceptible to side-channel attacks, where eves exploit auxiliary information obtained through non-quantum means, such as photon number splitting attacks or detector efficiency estimation attacks.

3.1.2 Entanglement-Based QKD

Entanglement Generation: Generating and maintaining entangled states can be technically challenging. Factors such as decoherence, environmental noise, and the need for high-fidelity entanglement can limit the efficiency and performance of the protocol.

Detection Efficiency: The efficiency of measuring entangled states is crucial in Entanglement-based protocols such as E91 protocol. Low detector efficiency can lead to higher error rates and reduce the overall key generation rate.

Distance Limitations: Entanglement-based protocols face distance limitations due to the decoherence effects and losses experienced during long-distance quantum transmission. Overcoming these limitations typically requires the development of quantum repeaters or other mechanisms to extend the range of secure communication.

3.1.3 Continuous Variable QKD

Implementation Complexity: Continuous Variable QKD protocols often involve complex optical setups, including balanced homodyne detection and precise phase stabilization techniques. Implementing and maintaining such setups can be technically demanding.

Noise and Imperfections: Continuous Variable QKD is sensitive to various noise sources, including detection noise, phase noise, and channel losses. These noise

sources can significantly impact the key generation rate and overall performance of the protocol.

Practical Limitations: Achieving high transmission rates in Continuous Variable QKD can be challenging due to factors like limited detector bandwidth and the need for precise calibration of devices. Balancing the trade-off between high transmission rates and maintaining security remains a practical challenge.

3.2 Existing QKD Protocols

Listed below are the known QKD protocols at the time.

QKD protocol	Year	Full form	Type
BB84	1984	Bennett-Brassard-84	Heisenberg uncertainty principle
E91 protocol	1991	Ekert-91	Entanglement
BBM92 protocol	1992	Bennett-Brassard-Mermin-92	Heisenberg uncertainty principle
B92 protocol	1992	Bennett-92	Heisenberg uncertainty principle
MSZ96	1996	Mu-Seberry-Zheng-96	Heisenberg uncertainty principle (CVQKD)
Six-state protocol	1998	Six-state protocol	Heisenberg uncertainty principle
Ping-Pong protocol	2002	Ping-Pong protocol	–
DPS protocol	2002	Differential phase shift protocol	Arbitrary
Decoy state protocol	2003	Decoy state protocol	–
SARG04	2004	Scaraci-acin-ribordy-gisin-04	Heisenberg uncertainty principle (CVQKD)
LM05	2005	Lucamarini-Mancini-05	–
COW protocol	2005	Coherent one-way protocol	Arbitrary
Three-stage quantum cryptography protocol	2006	Three-stage quantum cryptography protocol	–
KMB09 protocol	2009	Khan-murphy-beige-09	Heisenberg uncertainty principle
S09	2012	Serna-09	Public private key cryptography
S13	2013	Serna-13	Heisenberg uncertainty principle

(continued)

(continued)

QKD protocol	Year	Full form	Type
T12	2013	Toshiba twin field-12	–
EQKD	2016	QKD protocol based on entanglement and dense coding	Entanglement and dense coding
MEQKD	2016	Grouping QKD protocol based on entanglement and dense coding	Entanglement and dense coding
P-17	2017	Pastorello-17	Entanglement and GHZ
HDQKD	2017	High dimensional QKD	–
GEQKD	2019	GEQKD	Entanglement and dense coding
SEBV protocol	2021	Symmetrically entangled Bernstein Vazirani algorithm-based protocol	Entanglement and oracle

** Not listed types for protocols were not found in any reviews or papers

While there are numerous QKD protocols that can securely distribute a key pair, most of these have challenges when they are to be implemented practically. As the current developments are at the experimental level none of these QKD protocols can be used directly in an existing satellite or UAV system to provide security from Post-Quantum threats.

4 The Osslayer Protocol Implementation

The Osslayer protocol is a modified version of the transport layer to support encrypted port numbers and reserved bit field. We used Scapy to create a custom TCP header class that inherits from the original TCP header class and extends the dport and sport field. The dport and sport field is a 5-bytes field that replaces the original dport and sport fields, and stores the encrypted port numbers of the client and the server, the dport is set to 0 until the key exchanges finish, sport is set to a random value that will be changed later one to another value when the key exchange is finished. The reserved field will have the same bit size of the original reserved field, and will be used to stores a flag that indicates OssLayer support, other transport layer fields are retained as they were originally, xPoly20chacha1205 is used to encrypt and decrypt the port numbers. We also modified the checksum calculation function to include the new fields in the TCP header, the new layer is shown below, the protocol version in the IP layer is set to 99 (Fig. 1).

This new protocol is works as follows, the client initiate the PQ-TLS handshake by sending a normal TCP SYN packet but with the dport set to 0 and reserved set to 7, indicating PQ-TLS support. It also generates a NTRU public/private key pair and sends the public key as part of the TCP options field.

The server receive this PQ-TLS handshake request from the client and checks if the protocol version in the IP field is set to 99 and the dport is 0 and reserved is 7, and

```

class OssLayer(Packet):
    name = "OssLayer"
    fields_desc = [

        StrFixedLenField("sport", b"\x00\x00\x00\x00", 5),
        # encrypted source port with padding
        StrFixedLenField("dport", b"\x00\x00\x00\x00", 5),
        # encrypted destination port with
        IntField("seq", 0), # sequence number
        IntField("ack", 0), # acknowledgement number
        BitField("dataofs", None, 4), # data offset
        BitField("reserved", 0, 3), # reserved bits
        FlagsField("flags", 0x2, 9, "FSRPAUECN"), # flags
        ShortField("window", 8192), # window size
        XShortField("chksum", None), # checksum
        ShortField("urgptr", 0), # urgent pointer
        PacketListField("options", []), # options
        # StrFixedLenField("EncHeader", b"head", 4),
        # StrFixedLenField("EncTag", b"Tag1", 4),
        # StrFixedLenField("EncNonce", b"Nonce", 12),
    ]

```

Fig. 1 OssLayer fields

if so, it generate the NTRU public/private key and a secret shared key that is derived from the received client public key and uses xPoly20chacha1205 as it's symmetric encryption algorithm and sends it's public key as part of the TCP SYN-ACK packet with the reserved bit set to 6 decremented by 1.

The client will acknowledge the PQ-TLS handshake response from the server. It extracts the NTRU public key from the TCP SYN-ACK packet and uses it to compute the shared secret with its own NTRU private key, the client then decrement the reserved bit field by 1 and reply to the tcp syn ack with an ack.

The server receive the acknowledgement from the client sends it's FALCON based certificate that is hashed using harakav2 to ensure the client it's attempting to connect to the legitimate server, and encrypts it with the negotiated secret key and decrement the reserved field by 1.

The client verify the legitimacy of the PKI Certificate sent by the server, and sends it's hashed authentication id with harakav2 and set the dport of the desired server service and encrypt the transport layer accordingly.

The server then decrypts the transport layer, extract the necessary data and compares it to its database and check if the client is authentic and have the necessary permissions to connect to the requested service, if so, the server start the data exchange process with the client.

All of the algorithms (xPoly20chacha1205, NTRU, FALCON, Harakav2) are improved with an experimental QRNG qiskit 6 qubits circuit and it was used to generate nonces, one each connection the server will generate 2 nonces, a server

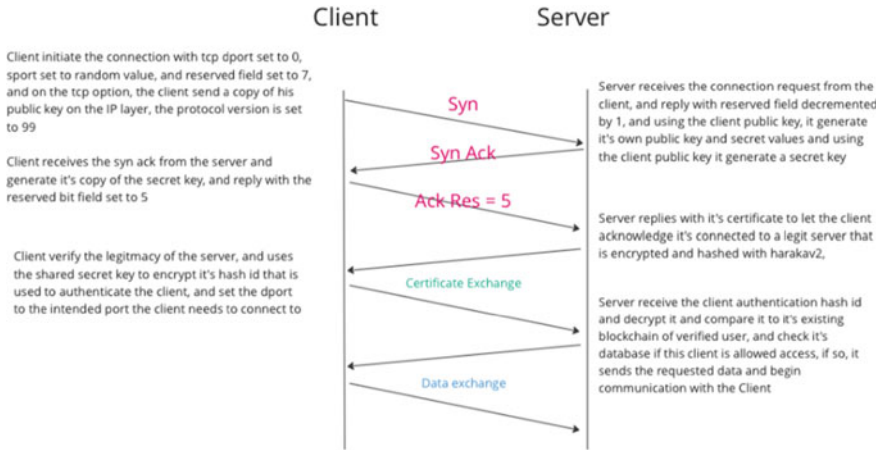


Fig. 2 Proposed architecture

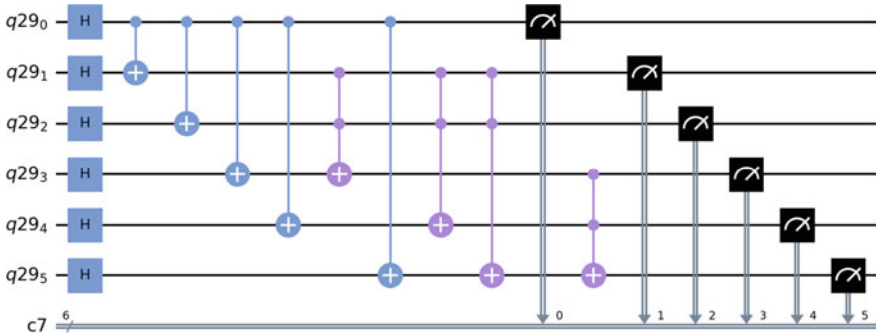


Fig. 3 First experimental QRNG circuit

nonce, and a client nonce that will be sent once the symmetric encryption key is negotiated, this client nonce will be expected to be used by the client on it's next communication link (Figs. 2 and 3).

Source port (5 Bytes)								Destination port (5 bytes)	
Sequence number*									
Acknowledgement number*									
Header length*	Reserved Bits*	U R G*	A C K*	P S H*	R S T*	S S N*	Y I N*	Window size* (Advertisement window)	
Check sum*								Urgent Pointer	

(continued)

(continued)

Source port (5 Bytes)								Destination port (5 bytes)	
Sequence number*									
Acknowledgement number*									
Header length*	Reserved Bits*	U R G*	A C K*	P S H*	R S T*	S Y N*	F I N*	Window size* (Advertisement window)	
Options Used for sharing public keys/Hash ID (Variable Bytes size)									
Data* (Optional)									

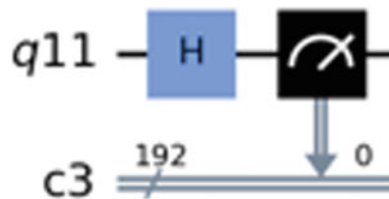
* Denotes that this field was not changed from the default Transport layer

This circuit is used to implement a quantum random number generator by defining a function returns a 192-bit nonce, which is a number that is only used once for cryptographic purposes. The function creates a quantum circuit with six qubits and applies various quantum gates to them, such as Hadamard, CNOT, and Toffoli gates, to create entanglement and superposition among the qubits. Then, the function measures the qubits and stores the outcomes in six classical bits. The function uses the Aer backend of Qiskit to simulate the quantum circuit on a local machine and estimates the number of shots (repetitions) needed to obtain a reliable result based on Hoeffding’s inequality. The function executes the circuit on the backend and obtains the result object, which contains the counts of measurement outcomes. The function converts the counts to a binary string by concatenating the outcomes according to their frequencies and truncates the string to get a 192-bit nonce, this circuit is only used as an example of QRNG, and is not intended to be used in a real application.

Another circuit that this new layer was tested with is as follows (Fig. 4).

This circuit serves the same purpose of the other circuit, but with a simpler quantum circuit. The code creates a quantum register with one qubit and a classical register with 192 bits. The code creates a quantum circuit that applies a Hadamard gate to the qubit 192 times and measures the qubit each time, storing the outcome in one of the classical bits. The Hadamard gate puts the qubit in an equal superposition of $|0\rangle$ and $|1\rangle$ states, making the measurement outcome random. The code executes IBM quantum computing backend with one shot and obtains the result object, which contains the counts of measurement outcomes. The code gets the counts and prints

Fig. 4 Second experimental QRNG circuit



the random number, which is a 192-bit binary string composed of the measurement outcomes.

QRNG can be used to generate the random polynomials for NTRU, the random nonce for xPoly20chacha1205, or the random seed for harakav2. These random numbers need to be uniformly distributed and indistinguishable from truly random by any adversary, including quantum adversaries. QRNG can provide these properties better than classical pseudo-random number generators (PRNGs), which may have biases, patterns, or weaknesses that can be exploited by quantum algorithms.

In real applications, it's possible to use the same approach as the QRNG implemented in a smartphone is the Samsung Quantum phone with SK Telecom, which was launched in 2020 as the world's first QRNG-powered 5G smartphone. The phone features a QRNG chipset developed by ID Quantique, which is the world's smallest QRNG at 2.5 mm square. The chipset uses an LED and a CMOS image sensor to capture the quantum randomness of light noise and produce true random numbers. The phone uses these random numbers to secure various services that require authentication or encryption, such as banking, payment, or biometric applications.

We tested our implementation on a local network using two virtual machines running Ubuntu 20.04 LTS. We ran one instance of our program as a client on one machine, and another instance as a server on another machine. We used Wireshark to capture and analyze the network traffic between them (Figs. 5, 6 and 7).

No.	Time	Source	Destination	Protocol	Length	Info
4	4.227203	192.168.68.143	192.168.68.139	IPv4	60	any private encryption scheme (99)
7	4.360989	192.168.68.139	192.168.68.143	IPv4	60	any private encryption scheme (99)
8	4.461910	192.168.68.143	192.168.68.139	IPv4	60	any private encryption scheme (99)
9	4.530673	192.168.68.139	192.168.68.143	IPv4	125	any private encryption scheme (99)
10	4.679146	192.168.68.143	192.168.68.139	IPv4	116	any private encryption scheme (99)
11	8.183160	192.168.68.139	192.168.68.143	IPv4	448	any private encryption scheme (99)
12	8.673416	192.168.68.143	192.168.68.139	IPv4	428	any private encryption scheme (99)
13	8.839805	192.168.68.139	192.168.68.143	IPv4	60	any private encryption scheme (99)
14	8.930645	192.168.68.139	192.168.68.143	IPv4	60	any private encryption scheme (99)
15	14.942536	192.168.68.143	192.168.68.139	IPv4	392	any private encryption scheme (99)
16	15.492234	192.168.68.139	192.168.68.143	IPv4	440	any private encryption scheme (99)
17	21.438766	192.168.68.143	192.168.68.139	IPv4	60	any private encryption scheme (99)
18	21.495060	192.168.68.139	192.168.68.143	IPv4	60	any private encryption scheme (99)
19	21.566445	192.168.68.143	192.168.68.139	IPv4	60	any private encryption scheme (99)

Fig. 5 Communication with the proposed OssLayer

```

0000  00 0c 29 94 41 88 00 0c 29 9e 1d cc 08 00 45 00  ..).A... ).....E.
0010  00 6f 00 01 00 00 40 63 6f c0 c0 a8 44 8b c0 a8  -o...@c o...D...
0020  44 8f 00 00 00 00 00 00 00 00 00 00 00 00 00  D.....
0030  00 00 00 00 00 18 20 00 00 00 00 00 37 2c 32 39  .....7,29
0040  2c 34 39 31 35 33 31 7c 5b 33 39 34 36 30 39 2c  ,491531| [394609,
0050  20 32 37 36 39 32 2c 20 36 32 33 30 37 2c 20 32  27692, 62307, 2
0060  36 33 30 37 33 2c 20 33 34 36 31 34 39 2c 20 34  63073, 346149, 4
0070  31 35 33 38 2c 20 33 33 39 32 32 35 5d          1538, 339225]
    
```

Fig. 6 Sever sending public keys

0000	00 0c 29 9e 1d cc 00 0c	29 94 41 88 08 00 45 00	..).A...E.
0010	00 66 00 01 00 00 40 63	6f c9 c0 a8 44 8f c0 a8	.f...@c o...D...
0020	44 8b 00 00 00 00 00 00	00 00 00 00 00 00 00 00	D.....
0030	00 00 00 00 00 18 20 00	00 00 00 00 5b 32 38 33[283
0040	38 38 39 2c 20 32 36 39	39 39 32 2c 20 34 38 34	889, 269 992, 484
0050	35 36 38 2c 20 33 35 33	30 35 34 2c 20 31 37 39	568, 353 054, 179
0060	39 39 35 2c 20 31 35 39	32 32 31 2c 20 32 33 35	995, 159 221, 235
0070	34 30 39 5d		409]

Fig. 7 Client public key

The client and server exchanged TCP packets with encrypted port numbers and reserved bit field set to 7.

The client and server exchanged NTRU public keys as part of their TCP packets.

The client and server computed a shared secret using their NTRU private keys and public keys.

The client and server exchanged FALCON public keys and signatures as part of their TCP packets or TLS messages (Figs. 8 and 9).

0000	00 0c 29 94 41 88 00 0c	29 9e 1d cc 08 00 45 00	..).A...E.
0010	01 b2 00 01 00 00 40 63	6e 7d c0 a8 44 8b c0 a8@c n}...D...
0020	44 8f 00 00 00 00 00 00	00 00 00 00 00 00 00 00	D.....
0030	00 00 00 00 00 18 20 00	00 00 00 00 7b 22 6e 6f{"no
0040	6e 63 65 22 3a 20 22 50	50 50 50 50 50 50 50 50	nce": "P PPPPPPPP
0050	50 50 50 50 50 50 50 22	2c 20 22 68 65 61 64 65	PPPPPPP", "head
0060	72 22 3a 20 22 63 32 56	79 64 6d 56 79 58 32 68	r": "c2V ydmVyX2h
0070	68 63 32 67 3d 22 2c 20	22 63 69 70 68 65 72 74	hc2g=", "ciphert
0080	65 78 74 22 3a 20 22 65	64 45 41 6f 30 6d 66 4c	ext": "e dEAo0mfL
0090	33 67 76 39 6a 65 4a 6b	74 43 31 42 6f 77 4b 71	3gv9jeJk tC1BowKq
00a0	4a 4d 70 51 63 51 33 76	4a 48 48 35 6e 57 47 46	JMpQcQ3v JHH5nWGF
00b0	68 6b 45 7a 57 4e 71 6b	7a 37 58 36 6c 53 62 33	hkEzWNqk z7X61Sb3
00c0	49 5a 51 39 44 75 35 74	72 74 73 4b 6d 48 52 59	IzQ9Du5t rtsKmHRY
00d0	69 31 44 45 2b 6d 67 58	69 58 41 54 7a 71 71 52	i1DE+mgX iXATzqqR
00e0	62 6d 67 61 6c 77 65 4a	75 30 2f 67 69 78 70 65	bmgalwEj u0/gixpe
00f0	69 5a 50 30 6d 47 6c 39	67 62 55 53 39 64 68 58	iZP0mG19 gbUS9dhX
0100	4b 70 68 49 4b 52 61 32	6a 70 6d 72 43 63 71 48	KphIKRa2 jpmrCcQH
0110	35 64 2b 42 32 59 36 4a	70 6e 37 58 38 78 41 54	5d+B2Y6J pn7X8xAT
0120	2f 6b 68 74 37 45 42 74	56 36 45 52 48 65 79 6f	/kht7EBt V6ERHeyo
0130	4f 42 78 79 2b 6d 63 30	2b 70 47 5a 72 30 4a 37	OBxy+mc0 +pGZr0J7
0140	2f 52 61 62 64 6f 71 6d	43 73 76 54 71 49 65 6a	/Rabdoqm CsvTqIej
0150	45 4b 6e 2f 47 68 45 69	66 34 54 51 54 6d 76 72	EKn/GhEi f4TQTmvr
0160	65 55 57 46 56 72 2b 41	73 55 69 77 36 66 52 53	eUWKVr+A sUiw6fRS
0170	64 56 6f 63 4e 59 72 36	63 2b 2b 55 6f 36 38 2f	dVocNYr6 c++Uo68/
0180	57 6a 43 46 35 52 6e 53	55 61 36 36 71 51 2b 35	WjCF5RnS Ua66qQ+5
0190	56 6b 34 37 6b 45 6e 62	55 41 3d 22 2c 20 22 74	Vk47kEnb UA=", "t
01a0	61 67 22 3a 20 22 35 33	39 62 6a 4d 4d 45 37 75	ag": "53 9bjMME7u
01b0	6c 74 77 46 73 7a 36 33	4e 45 53 77 3d 3d 22 7d	1twFsZ63 NESw=="}

Fig. 8 Server hashed and encrypted signature

0000	00 0c 29 94 41 88 00 0c	29 9e 1d cc 08 00 45 00	..).A...).....E..
0010	01 aa 00 01 00 00 40 63	6e 85 c0 a8 44 8b c0 a8@c n...D...
0020	44 8f 00 00 00 00 00 00	00 00 00 00 00 00 00 00	D.....
0030	00 00 00 00 00 08 20 00	00 00 00 00 7b 22 6e 6f{"no
0040	6e 63 65 22 3a 20 22 58	58 58 58 58 58 58 58 58	nce": "X XXXXXXXX
0050	58 58 58 58 58 58 58 22	2c 20 22 68 65 61 64 65	XXXXXXX", "heade
0060	72 22 3a 20 22 59 32 78	70 5a 57 35 30 58 33 4a	r": "Y2x pZW50X3J
0070	6c 63 47 78 35 4e 6a 63	3d 22 2c 20 22 63 69 70	lGx5Njc =", "cip
0080	68 65 72 74 65 78 74 22	3a 20 22 68 47 49 79 37	hertext": "hGIy7
0090	58 6f 41 41 51 75 79 2b	4f 42 52 79 54 73 79 6e	XoAAQuy+ OBRYTsyn
00a0	61 6b 49 61 45 46 54 58	51 6d 41 64 6c 43 7a 43	akIaEFTX QmAdlCzC
00b0	51 49 64 6d 4c 7a 38 46	35 61 48 34 51 55 41 6f	QIdmLz8F 5aH4QUAo
00c0	6f 30 51 35 5a 4b 72 65	67 58 50 32 42 47 56 44	o0Q5ZKre gXP2BGVD
00d0	68 30 6e 48 2f 37 75 47	30 62 56 5a 53 6b 52 35	h0nH/7uG 0bVzSkR5
00e0	59 55 2f 64 5a 66 55 6c	6b 5a 78 32 51 44 37 74	YU/dZfUl kZx2QD7t
00f0	73 59 59 4d 78 5a 63 4c	73 61 67 6d 6a 74 78 59	sYYMxZcl sagmjtxY
0100	72 50 43 51 79 56 5a 50	54 55 37 4b 6a 32 45 54	rPCQyVZP TU7Kj2ET
0110	62 69 79 49 62 72 37 32	50 30 54 78 56 4f 44 4f	biyIbr72 P0TxV0DD
0120	71 56 4f 76 65 73 55 4f	4a 43 58 54 2b 4f 55 4a	qV0vesUO JXCt+OUJ
0130	47 66 6c 32 36 33 31 66	53 57 36 69 30 31 64 6f	GfL2631f SW6i01do
0140	66 62 55 6c 71 59 5a 44	44 69 35 35 38 43 7a 2f	fbUlqYZD Di558Cz/
0150	39 46 65 7a 48 44 41 79	36 67 2f 51 48 39 59 45	9FczHDAY 6g/QH9YE
0160	6f 44 73 47 77 4d 76 6e	43 76 41 7a 6e 59 73 2f	oDsGwMvn CvAznYs/
0170	78 44 35 6b 68 4e 32 35	37 58 39 58 35 53 36 6c	x05khN25 7X9X5S61
0180	42 72 47 6b 32 7a 49 41	70 6b 38 45 42 54 2f 37	BrGk2zIA pk8EBT/7
0190	67 77 3d 22 2c 20 22 74	61 67 22 3a 20 22 73 36	gw=", "t ag": "s6
01a0	76 71 32 68 79 6f 53 35	78 67 76 2f 4b 64 71 57	vq2hyoS5 xgv/Kdqw
01b0	62 6b 65 41 3d 3d 22 7d		bkeA="}

Fig. 10 Encrypted communication

```

> Frame 16: 440 bytes on wire (3520 bits), 440 bytes captured (3520 bits) on interface \De
> Ethernet II, Src: VMware_9e:1d:cc (00:0c:29:9e:1d:cc), Dst: VMware_94:41:88 (00:0c:29:94
< Internet Protocol Version 4, Src: 192.168.68.139, Dst: 192.168.68.143
  0100 .... = Version: 4
  .... 0101 = Header Length: 20 bytes (5)
  < Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
    0000 00.. = Differentiated Services Codepoint: Default (0)
    .... ..00 = Explicit Congestion Notification: Not ECN-Capable Transport (0)
  Total Length: 426
  Identification: 0x0001 (1)
  < 000. .... = Flags: 0x0
    0... .... = Reserved bit: Not set
    .0.. .... = Don't fragment: Not set
    ..0. .... = More fragments: Not set
    ...0 0000 0000 0000 = Fragment Offset: 0
  Time to Live: 64
  Protocol: any private encryption scheme (99)
  Header Checksum: 0x6e85 [validation disabled]
  [Header checksum status: Unverified]
  Source Address: 192.168.68.139
  Destination Address: 192.168.68.143
  > Data (406 bytes)

```

Fig. 11 Visible layers in Wireshark

Therefore, a classical alternative like PQ-TLS is identified as a much more suitable solution for Satellite and UAV communication compared to state-of-the-art quantum technologies. However, it should be noted that ongoing research and advancements in QKD technology, along with system optimization and integration efforts, are paving the way for enhanced security in satellite communication and UAV operations using QKD protocols.

The proposed PQ-TLS protocol can be useful for securing communications between satellites and drones, which are emerging applications that require high security and performance. Satellites and drones can communicate with each other or with ground stations using wireless channels, such as radio frequency (RF) or optical links. These channels are vulnerable to eavesdropping, jamming, spoofing, or replay attacks by malicious parties, who may use quantum computers to break the encryption or authentication schemes used in conventional TLS protocols. Therefore, there is a need to use post-quantum cryptography to protect the confidentiality, integrity, and authenticity of the data exchanged between satellites and drones.

The PQ-TLS protocol can offer several advantages for satellite-drone communications, such as:

- Reducing the communication overhead by using encrypted port numbers and reserved bit field to indicate PQ-TLS support, instead of modifying the TLS record layer or introducing new extensions or cipher suites.
- Reducing the computation cost by using key-encapsulation mechanisms (KEMs) instead of signatures for server authentication, which are generally faster and require smaller keys than post-quantum signature schemes.
- Reducing the latency by using a one-round-trip-time (1-RTT) handshake that allows the client to send encrypted application data immediately after receiving the server's public key, without waiting for a signature verification or a certificate validation.
- Enhancing the security by using post-quantum algorithms that are resistant to quantum attacks, such as NTRU, xPoly20chacha1205, FALCON, and harakav2

References

1. Chamola V, Jolfaei A, Chanana V, Parashari P, Hassija V (2021) Information security in the post quantum era for 5G and beyond networks: threats to existing cryptography, and post-quantum cryptography. *Comput Commun* 176:99–118
2. Sasirekha MHN (2014) Quantum cryptography using quantum key distribution and its applications. *Int J Eng Adv Technol (IJEAT)* 3(4)
3. Lo H, Curty M, Tamaki K (2014) Secure quantum key distribution. *Nat Photon* 8:595–604
4. Kirsch Z, Chow M (2015) Quantum computing: the risk to existing encryption methods
5. Ampatzis M, Andronikos T (2021) QKD based on symmetric entangled Bernstein-Vazirani. *Entropy* 23(7)
6. Singh H, Gupta D, Singh A (2014) Quantum key distribution protocols: a review. *IOSR J Comput Eng (IOSR-JCE)* 16(2):1–9

7. Toshiba (2022) Quantum key distribution and Blockchain white paper. [Online]. https://www.global.toshiba/content/dam/toshiba/ww/products-solutions/security-ict/qkd/resources/pdf/Toshiba_QKD_Blockchain_White_Paper.pdf. Accessed 4 July 2023
8. Quantum Computing Cybersecurity Preparedness Act of 2022, Pub. L. No. 117-260, 136 STAT. 2389
9. IBM (2023) IBM Quantum Safe IBM. [Online]. <https://www.ibm.com/quantum/quantum-safe>. Accessed 20 April 2023
10. Bennett CH, Brassard G (1984) Quantum cryptography: public key distribution and coin tossing. *Theor Comput Sci*
11. Ekert AK (1991) Quantum cryptography based on bell's theorem. *Phys Rev Lett* 67(6)
12. Scarani V, Acin A, Ribordy G, Gisin N (2004) Quantum cryptography protocols robust against Photon number Splitting attack. *Phys Rev Lett* 92
13. Post-Quantum TLS Without Handshake Signatures by Peter Schwabe, Douglas Stebila and Thom Wiggers <https://dl.acm.org/doi/abs/https://doi.org/10.1145/3372297.3423350>
14. Post-Quantum Key Exchange for the TLS Protocol from the Ring Learning with Errors Problem by Joppe W. Bos; Craig Costello; Michael Naehrig; Douglas Stebila <https://ieeexplore.ieee.org/abstract/document/7163047>
15. Prototyping post-quantum and hybrid key exchange and authentication in TLS and SSH by Eric Crockett, Christian Paquin, and Douglas Stebila <https://eprint.iacr.org/2019/858>
16. Saha R, Kumar G, Devgun T, Buchanan WJ, Thomas R, Alazab M, Hoon-Kim T, Rodrigues JJPC (2023) A Blockchain framework in post-quantum decentralization. *IEEE Trans Serv Comput* 16:1–12. <https://doi.org/10.1109/TSC.2021.3116896>
17. Bacsardi L, Imre S (2011) Analyzing the quantum based satellite communications. *Procedia Comput Sci* 7:256–257. <https://doi.org/10.1016/j.procs.2011.09.036>
18. Yin J, Cao Y, Li Y-H, Liao S-K, Zhang L, Ren J-G, Cai W-Q, Liu W-Y, Li B, Dai H, Li G-B, Lu Q-M, Gong Y-H, Xu Y, Li S-L, Li F-Z, Yin Y-Y, Jiang Z-Q, Li M, Jia JJ, He D, Ren G (2017) Satellite-based entanglement distribution over 1200 kilometers. *Science* 356:1140–1144. <https://doi.org/10.1126/science.aan3211>
19. Fernández-Caramés TM, Fraga-Lamas P (2020) Towards post-quantum blockchain: a review on blockchain cryptography resistant to quantum computing attacks. *IEEE Access* 8: 21091–21116. <https://doi.org/10.1109/ACCESS.2020.2968985>
20. Fernández-Caramés TM (2020) From pre-quantum to post-quantum IoT security: a survey on quantum-resistant cryptosystems for the internet of things. *IEEE Internet Things J* 7: 6457–6480. <https://doi.org/10.1109/JIOT.2019.2958788>
21. Canadian Space Agency (2020) Satellites serving Earth. [Online]. <https://www.asc-csa.gc.ca/eng/satellites/>. Accessed 20 July 2023
22. Ivashov S, Tataraidze A, Razevig V, Smirnova E (2019) Railway transport infrastructure monitoring by UAVs and satellites. *J Transp Technol* 9(3):342–353
23. Cao Y, Li Y-H, Yang K-X, Jiang Y-F, Li S-L, Hu X-L, Abulizi M, Li C-L, Zhang W, Sun Q-C, Liu W-Y, Jiang X, Liao S-K, Ren J-G, Li H, You L, Wang Z, Yin J, Lu C-Y, Xian Y (2020) Long-Distance Free-Space Measurement-Device-Independent Quantum Key Distribution. *Phys Rev Lett* 125(26)

Evaluating the Security of Open-Source Linux Operating Systems for Unmanned Aerial Vehicles



Darshan Patil and Sina Pournouri

Abstract The proliferation of unmanned aerial vehicles (UAVs), commonly known as drones, has brought substantial benefits across various industries. However, the increasing use of drones has also raised concerns about their security, particularly with regard to open-source operating systems. Open-source operating systems serve as a foundational framework for drone enthusiasts and companies alike. Unfortunately, security considerations have been disregarded by some who directly appropriate open-source code, introducing inadvertent vulnerabilities. Given the evolving role of drones in mission-critical systems, the demand for resilient and secure operating systems becomes increasingly evident. This study addresses the critical need to assess the security of open-source operating systems used in drones.

Keywords Unmanned aerial vehicles · Open-source operating systems · Drones · Security threats · Mission critical systems · Security vulnerabilities

1 Introduction

Unmanned aerial vehicles (UAVs), also known as drones, are becoming increasingly popular. They have enormous potential in a variety of industries, including agriculture, military, transportation, surveillance, supply chain, and film. UAVs are also popular among hobbyists for photography, racing, and other purposes [1].

Despite the many benefits of UAVs, there are also some risks associated with their use. UAVs can be targeted by malicious parties, and if compromised, they could be used to carry out life-threatening activities. As a result, governments all over the world have begun to regulate the use of UAVs [1].

Intelligent and automated defense mechanisms are required for UAVs to ensure the safety of humans, property, and the UAVs themselves. The operating system is a critical component into which we can incorporate the defense mechanism [1].

D. Patil · S. Pournouri (✉)
Sheffield Hallam University, Sheffield, UK
e-mail: S.Pournouri@shu.ac.uk

UAVs use operating systems to manage their functions such as flight control, navigation, and data processing. However, like any other computer system, UAV operating systems are also vulnerable to various security threats. Many hobbyists and enterprises use open-source operating systems as a base or a starting point to build their own UAVs. In fact, some hobbyists simply borrow the open-source operating system and use it directly without even thinking about the security concerns.

The fact that these operating systems are open source means that anyone can submit their code, which might or might not be intentionally vulnerable. As more and more enterprises, such as Red Hat, begin to develop UAVs, and as the functionalities of drones increase and they are used in mission-critical systems, there is a growing need for better operating systems. Therefore, it is essential to evaluate the security of these open-source operating systems.

The focus of this research is to evaluate the security of open-source operating systems of drones. Multiple open-source operating systems will be evaluated to uncover the vulnerabilities. The findings of this research will be helpful for security researchers, hobbyists, and enterprises to come up with better operating systems for UAVs.

2 Literature Review and Background Research

Unmanned Aerial Vehicles (UAVs) or drones are aircraft that fly either with the help of a controller or independently. Where big Organizations like Amazon deploy drones for the purpose of delivery of goods and military drones increasing in number for military operations, we see civilians and hobbyists using them for recreational purposes. They can be used for medicinal services, natural disaster management, construction sites, mining, cinematography and Military Combat [2].

Drones pose a serious threat to privacy and safety hazards. Drones equipped with IR and high-resolution cameras and microphones have many security issues. Vulnerability assessment and investigation of drones have become of utmost importance, post some security incidents where a UAV transfers video or a chance of capturing or sticking a UAV in flight [2].

2.1 *Components of UAS*

The components of a drone can be divided into three parts namely Aerial Platform, Ground Control Station and Communication.

- **Aerial Platform:** This includes the airframe, the navigation system, the power system and the payload. This aerial platform is known as Unnamed Aerial Vehicle (UAV). Aerial Platform includes the aerial frame, onto which the other components such as motors, cameras, video transmitters, sender and receiver, Electronic

Flight Controller (ESC) and the controller also known as the brain of the system. The important point to be noticed here is that the operating system is installed on the Controller [3].

- **Ground Control Station (GCS):** The purpose of the GCS is to provide its operator with the ability to control the Aerial Platform or the UAV [3].
- **Communication System:** This facilitates communication between the Aerial Platform and the Ground Control Station [3].

All the above-mentioned components together are known as Unmanned Aerial System (UAS). Prior research has been done on the security evaluations of drones. However, most of the research does not focus on the security aspects of the operating systems.

Security Vulnerabilities of Unmanned Aerial Vehicles and Countermeasures: An Experimental Study by [4] discusses the security issues of two popular drones, the DJI Phantom 4 Pro and the Parrot Bebop 2 and proposes solutions to improve their security. The vulnerabilities include GPS spoofing, radio signal interception, open Wi-Fi, Telnet, and FTP port. Remedies include WPA security, Telnet password, MAC-filtering, and disguised SSID [4]. The paper also suggests using encryption/packer, an obfuscator, improving SDK authentication, and encrypting firmware code to detect spoofing. It advises future evaluation of P4P drone transmission using SDR equipment.

The paper has strengths, including a detailed analysis of the drones' architecture, internal operations, and governing system. It also conducts experiments to identify security vulnerabilities, compares them in terms of user-friendliness and security issues, proposes countermeasures, and exposes new vulnerabilities. However, the paper's limitations include limited scope, lack of real-world testing, limited countermeasures, and lack of discussion on legal and ethical implications [4].

A review of cybersecurity Vulnerabilities for unmanned aerial vehicles [5] has examined the trade and scientific literature of drones, focusing on cyberattacks in real and simulated environments. It presents a modified taxonomy to organize cyberattacks on UAVs and exploits threats by Attack Vector and Target. The paper concludes that research on countering cybersecurity threats to UAVs focuses on GPS jamming and spoofing, while attacks on controls and data communications stream are ignored [5]. The paper discusses the vulnerability of small commercially available unmanned aerial systems (sUAS) to cyber-attacks and the need for improved security measures. However, the paper's limitations include a limited focus on small commercially available UAVs and lack of discussion on potential solutions or mitigation techniques [5]. The paper's relevance is not directly relevant to the study's objective of evaluating the security of open-source Linux operating systems used in drones, but it provides examples of security vulnerabilities and attacks on different types of drones [5].

A Cyber Security Analysis Used for Unmanned Aerial Vehicles in the Smart City by [6] discusses the use of unmanned aerial vehicles (UAVs) in smart cities and the potential cyber security threats they face. It explores the advantages of UAVs in fields like traffic management, disaster management, and agriculture, but also highlights

the security vulnerabilities they come with [6]. The paper provides recommendations for preventing cyber-attacks on UAVs, such as using strong passwords, fixed MAC addresses, and encryption algorithms. It also identifies various types of cyber-attacks that can target UAVs, such as man-in-the-middle attacks, SYN floods, ACK floods, RST/FIN floods, ICMP floods, UDP floods, PING floods, DNS floods, random recursive GET attacks, HTTP fragmentation attacks, NTP floods, GPS spoofing attacks, brute force attacks, Trojan horse viruses, SQL injection attacks, and phishing/social engineering attacks. It also recommends protecting against these attacks using firewalls, intrusion detection systems, anti-malware/antivirus software, and raising awareness among users about the risks of social engineering attacks [6]. The paper also provides specific examples and case studies to illustrate the potential risks and vulnerabilities of UAVs in smart cities. While not directly relevant to the study of evaluating the security of open-source Linux operating systems used in drones, it provides useful insights into the broader context of UAV security and the potential threats and countermeasures that need to be considered.

A Study on UAV Operating System Security and Future Research Challenges [1] discusses the security challenges in operating systems used in Unmanned Aerial Vehicles (UAVs) and their potential applications. It surveys security issues, investigates existing systems, and addresses research challenges for developing a secure UAV operating system. The paper emphasizes the need for an intelligent and automated defense mechanism to ensure the safety of humans, properties, and UAVs [1]. A secure UAV operating system should ensure information confidentiality, authenticated access, software/data integrity, system availability, and accountability of actions. The paper also discusses possible solutions and the need for a legal framework and appropriate technologies for forensic analysis after illegal drone activities [1].

However, the paper lacks a detailed analysis of the technical implementation of proposed solutions, a comparative analysis of security measures' effectiveness, economic feasibility, and potential impact on UAV performance and functionality. Despite these limitations, the paper is relevant to evaluating the security of open-source Linux operating systems used in drones.

2.2 Background Research Focused on UAV Operating Systems

"A Study on UAV Operating System Security and Future Research Challenges" is the paper available on UAV operating systems and their security [1].

UAV technology is still in its infancy, and the majority of amateur flight controllers do not yet offer a full-fledged operating system. Indeed, many of them (such as Betaflight, Cleanflight, Raceflight, and INAV) just have a scheduler and no real-time kernel. It is unknown whether the drone flight controller requires a full-fledged operating system [1]. A full-fledged operating system is required if the drone is

coupled with a companion computer for enhanced autonomy features. Furthermore, future drones may employ various hardware components, necessitating the use of separate operating systems tailored to their specific duties [1].

- **Evaluated Flight Controllers and their Operating Systems**

1. ArduPilot

One of the first open-source drone software systems was ArduPilot. The Arduino hardware (an ARM 8-bit MCU) was used to start the project, and the software has since been optimized for 32-bit ARM-based MCUs. ArduPilot is compatible with a wide range of unmanned vehicles, including multi-rotors, fixed-wing planes, helicopters, and even submarines. The goal of Ardupilot is to make these vehicles totally self-sufficient. ArduPilot is licensed under the GPL, therefore any changes to the software must be integrated into the parent project. ArduPilot is a real-time operating system that can run on both STM32 and Linux-based boards (e.g., Raspberry PI) [1].

Security Issues

ArduPilot does not have any particular security measures in place to defend it from external threats. It uses the MAVLink protocol to communicate with external devices. The MAVLink protocol is vulnerable to various common attacks and lacks CIA (Confidentiality, Integrity, and Availability) security services.

Eavesdropping and other attacks on confidentiality and privacy are possible with the protocol [1]. Furthermore, the MAVLink protocol does not guard against Man-in-the-middle (MITM) attacks (for example, replay and injection attacks) [1]. In addition, attackers can prevent communications from reaching the UAV, rendering the control station ineffective.

2. PX4

PX4 Autopilot is a complete flying stack consisting of flight controller software, ground control station software, and simulation software [7]. PX4 is largely compatible with Pixhawk flight controllers and is capable of supporting fixed-wing, multi-rotor, single-rotor, and VTOL technology [1]. PX4 is popular among commercial businesses because of its adoption of the BSD license. PX4 interfaces with the QGroundControl ground control station via the MAVlink protocol. PX4 may run its own real-time operating system or on top of Linux, just like ArduPilot [1].

Security Issues

Because PX4 employs MAVLink [1] and we have seen in the security of ArduPilot.

3. Paparazzi

The open-source project Paparazzi UAV is developing a complete drone system, which includes hardware, flight controller software, an autopilot, a Ground Control Station (GCS), and a simulator [1]. The Paparazzi flight controller can be customized

to handle multicopter, fixed-wing, flapping-wing, helicopter, and hybrid aircraft, as well as multiple unmanned aerial vehicles (UAVs). The Paparazzi Ground Control Station (GCS) can be used to configure parameters and read sensor data. We can also design a flight plan that includes personalized guidance, navigation, and control algorithms. Paparazzi, which can be loaded on dedicated flight controllers or Linux-based drones, uses the GPL license [1].

Security Issues

PPRZLINK is the Paparazzi UAV project's communication toolbox (message definition, code generators, libraries, and so on). It distinguishes three types of messages: uplink (datalink), downlink (telemetry), and ground-to-ground (messages transmitted between ground agents) [1]. Secure Pprzlink, an encrypted communication system for unmanned aerial vehicles, was released in 2018 [1]. Chacha20, a Poly1305 authenticator, and HACSL, a formally proven cryptography library, are used. It is the only open-source protocol that allows UAVs to communicate in an encrypted fashion [1].

4. DJI

DJI is a popular consumer drone company. DJI's products are all built with proprietary hardware and software. It also supplies flight controls for amateurs. DJI's devices generally utilize wireless connectivity like as Wi-Fi, OcuSync, and Lightbridge [1]. DJI claims that Lightbridge and OcuSync are proprietary protocols that outperform Wi-Fi. They are more reliable and have a wider bandwidth and range [1].

Security Issues

It is difficult to remark on the security of DJI firmware because the source code for DJI drones is not available [1]. However, researchers were able to compromise DJI drones and identify various security weaknesses. DJI has increased UAV security since then, and communication channels (Lightbridge and OcuSync) can now be encrypted. OcuSync 2.0 is utilized in government and law enforcement drones and offers AES 256 encryption [1].

Even though the above research claims to have evaluated the security of the operating system and identified the security issues with them. However, the identified security issues are associated with the communication part with a missing focus on the operating systems.

The Real Operating systems (RTOS) or Robots Operating systems (ROS) act as a middleware, sitting on top of a core operating system. In this study, core operating systems such as ubuntu core and Raspberry pi OS are used on top of which RTOS such as NuttX OS, Chibi OS and NOOBS will be used.

This study will be evaluating security first of the core operating systems by itself and then with the middleware sitting over it, by performing vulnerability scanning, networking scanning, checking for cryptographic failures, checking for kernel exploits and hashing algorithms used in the operating systems to store passwords etc.

2.3 Background on RTOS and ROS

RTOS

Real Time operating systems generally consists of two parts one the kernel space and the user space. The kernel is the main part of any Operating System, which acts as a bridge between applications and the actual data processing at the hardware. RTOS are defined as the operating systems that have the ability to provide required service in bounded response time and they are categorized into hard and soft real time operating systems.

Hard RTOS

These RTOS carefully follow the deadlines related to the assignments. These systems cannot accept any kind of delay; else, the system will fail. For example, brake systems in automobiles and pacemakers.

Soft RTOS

Missing a deadline is permitted in this RTOS. These systems can endure delays, but they are very undesirable. Online databases, for example. Furthermore, they are categorized according to the types of hardware devices supported (e.g., 8-bit, 16-bit, 32-bit MPU).

2.3.1 Types of Kernels

Monolithic Kernels

Highly prominent in its early days, also known as spaghetti code or the “The Big Mess”. The ins system is a just a collection of procedures where each module calls on to the other module. These types of OS are considered to have a problem of difficult to debug, when it comes to high reliability a monolithic kernel is considered as unstructured.

Mirco Kernel

Code is moved into “user” space as much as possible from the kernel in this case. Message forwarding is used to communicate between user modules. Extending a microkernel and porting the operating system to new architectures are simpler. It is more stable (there is less code to run in kernel mode), reliable and secure.

Exo-Kernel

In this case, the kernel separates hardware from application. The kernel assigns physical resources to applications. It is conceptually based (file systems, virtual address space, schedulers, and sockets). What happens with these sources is determined by the application. It can connect to a lib OS to imitate a traditional operating system. When strong hardware contact is necessary, this method is usually utilized.

2.4 Literature Review for RTOS and ROS

Modeling and Security Analysis of a Commercial Real-Time Operating System Kernel by [8] has the INTEGRITY-178B real-time operating system, developed by GreenHills Software, is a proprietary system for safety-critical systems like avionics. It has been certified to DO-178B Level A and is a good candidate for Common Criteria certification [8]. The evaluation requirements for EAL 5 and above specify formal or semiformal elements, including a security policy model, functional specification, high-level design, low-level design, and representation correspondence. The level of rigor applied to INTEGRITY-178B includes a formal security policy model and semiformal representations of functional and system designs. The paper's strengths include a clear overview of its features, certification requirements, and level of rigor applied to its development and evaluation [8]. However, it is relatively short and may not provide enough detail for readers seeking more in-depth information. The relevance of this paper is not directly relevant to evaluating the security of open-source Linux operating systems used in drones.

Security and Reliability of Safety-Critical RTOS [9] paper reviews the security and reliability of safety-critical real-time operating systems (RTOS), emphasizing the importance of using certified RTOS and the interplay between predictability and security [9]. It highlights current security vulnerabilities and countermeasures and discusses the challenges of developing RTOS for use in domains like aviation, automotive, and medical devices. The paper concludes with the importance of maintaining high standards for safety and reliability in RTOS and future research directions [9]. However, it may be technical and difficult to understand for readers without computer science or engineering background. The paper may be relevant to evaluating the security of open-source Linux operating systems used in drones, as it provides a comprehensive review of RTOS security and reliability.

Real-Time Operating System Security by Yu Weider [10] discusses security issues in embedded systems, particularly in supervisory control and data acquisition (SCADA) systems. It highlights vulnerabilities and suggests common approaches to improve security. The sources cover various aspects of security, including real-time operating systems, MILS architecture, and potential cyber threats [10]. The paper provides examples and cases to illustrate security vulnerabilities and threats, and suggests common approaches to improve security, such as good programming techniques, testing, and firewalls. It emphasizes the importance of security in the development phase and the need for constant monitoring and updates [10]. However, the paper may lack depth in some areas and may not be up to date with the latest developments in security issues. The paper may not be directly relevant to evaluating the security of open-source Linux operating systems used in drones, but it provides general information on security issues in embedded systems.

Security for the Robot Operating System [11], discusses security issues in the Robot Operating System (ROS) and proposes various solutions to address them. It presents common vulnerabilities, attack vectors, and solutions to secure ROS and similar systems [11]. The proposed solutions include an Authentication Server (AS)

for authentication and encryption keys, application-level security, modified communication channels with TLS and DTLS, and modifying the source code of ROS to establish (D)TLS channels for authentication, authorization, and confidentiality. The paper emphasizes the importance of security in robotic systems and the challenges it presents [11].

The paper has strengths, such as a practical testbed to compare plain ROS and hardened versions against fixed attack patterns. It also discusses the importance of security in robotic systems and proposes a security architecture on the application level and the ROS core to harden ROS. It also discusses the importance of accountability in securing ICSs and proposes a blockchain-based logging mechanism [11].

However, the paper lacks a detailed evaluation of the proposed solutions in real-world scenarios, potential limitations, impact on performance, challenges in implementing legacy systems, and potential cost. This paper could provide insights and ideas for evaluating the security of open-source Linux operating systems used in drones and developing solutions to address security issues.

Security Evaluation of a Linux System a Common Criteria EAL4+ certification experience by [12] examines the certification process of FIN.X RTOS, a Linux distribution certified as Common Criteria Evaluation Assurance Level 4+. It covers security targets, requirements, testing, and vulnerability assessment. The FIN.X RTOS successfully passed the evaluation process in May 2014, and the paper discusses security functional requirements, software solutions, and the security test suite used to measure conformance [12]. The article concludes that EAL4 + CC certification is feasible for manufacturers, even those not selling operating systems, and highlights the need for a cooperative and open framework for developing certifiable software. The paper's strengths include a detailed explanation of the Common Criteria scheme and certification process, a discussion of vulnerability assessment, and specific examples of vulnerabilities identified and fixed. However, it also lacks a clear explanation of the certification process's limitations and potential threats. Overall, the paper is relevant for evaluating the security of Linux-based drone operating systems and identifying potential vulnerabilities.

Summary

After performing extensive literature review and background research it can be said that when it comes to the security of drone extensive research has been done on the communication aspect of the drone, and not on the operating system. Since the operating systems used in the drone are either RTOS or ROS, which are also used in satellites, IoT devices etc. Therefore, this work will not only help in improving the security posture UAVs but also the security posture of satellites, IoT devices and all those fields or devices where an RTOS or and ROS has been used.

3 Experiment Methodology

In this section, multiple other research papers are discussed, in order to come up with a methodology for carrying out the experiments better and combine those conclusions along with my observations and knowledge of the domain.

Note: All of the experiments will be carried out on Ubuntu Core, which has real-time kernel and ROS variants.

3.1 *Comparing Multiple Approaches on Evaluating Security of the Linux Operating Systems*

PeX: a permission check analysis framework for Linux Kernel [13]

Methodology—The paper proposes PeX, a Permission Check Analysis Framework tailored to the Linux kernel. PeX seeks to alleviate kernel developers' difficulties in applying and verifying permission checks in an accurate and efficient manner [13]. It offers KIRIN, an accurate and scalable indirect call analysis technique that improves the accuracy of indirect call targets.

The framework aids in the detection of missing, inconsistent, or duplicated permission checks in the Linux Kernel source code. Developers can improve the security of the operating system and confirm the validity of existing permission checks by utilizing PeX [13].

The document does not provide a full overview of all of the content, but it is a useful resource for understanding the PeX architecture and its contributions to increasing the security of the Linux Kernel.

An analytical approach to assess and compare the vulnerability risk of operating systems

The purpose of this work is to give an analytical approach for evaluating and comparing the vulnerability risk of various operating systems. Its goal is to provide insights into the security elements of operating systems and to help users make educated decisions about their use.

To analyze and assess the vulnerability risk, the authors employ a variety of statistical methodologies and modelling approaches such as Stochastic Process, Markov Model, Principal Component Analysis (PCA), Factor Analysis (FA), and Structural Equation Modelling (SEM). They hope to provide a thorough understanding of the aspects contributing to vulnerability risk in operating systems by utilizing these techniques.

The research provided in this paper advances the field of computer network and information security by providing a systematic and analytical approach for assessing and comparing operating system vulnerability risk. It gives useful insights for field researchers, practitioners, and decision-makers, allowing them to make informed

decisions about the selection and use of operating systems based on their vulnerability risk profiles.

Overall, this paper is an excellent resource for anyone interested in understanding and assessing the vulnerability risk associated with various operating systems.

Security evaluation of a linux system: common criteria EAL4+ certification experience [12]

The technique used to execute Linux system security evaluations may involve Common Criteria (CC) certification, vulnerability assessment, and security testing.

CC certification entails assessing security functional and assurance requirements, as well as vulnerability assessment and penetration testing. The certification process provides assurance that the product has been rigorously tested and can be trusted to enforce the security policy set by the product's security features [12].

Vulnerability assessment is finding potential system vulnerabilities and evaluating their impact and possibility of exploitation. Vulnerability assessments can be performed with a variety of tools and methodologies, such as manual code review, automated vulnerability scanners, and penetration testing.

Security testing entails examining the system for security flaws and vulnerabilities using a variety of approaches such as functional testing, penetration testing, and coverage analysis. Security testing can assist in identifying potential vulnerabilities and weaknesses in the system as well as providing insights into how to improve system security [12].

Auditing Linux operating system with center for internet security (CIS) standard [14]

In this research study, the Linux operating system is audited using the Chef Inspec application to perform security audits on the Ubuntu 20.04 operating system. The Centre for Internet Security (CIS) Benchmark is used to assess the operating system's security protections using Chef Inspec [14]. The audit results, including the success, failure, and skipped controls. The audit success rate was 42.44% [14].

This research paper's technique entails installing Ubuntu on a virtual machine and implementing the CIS Benchmark Ubuntu 20.04 on Chef Inspec using the Ruby programming language [14]. The CIS Benchmark controls are translated into Ruby programming language, and the audit is carried out with Chef Inspec. Figure 1 of the study depicts the research flow. The process also includes maintaining and auditing system logs with tools like rsyslog and auditd, as well as appropriately controlling system access, authentication, and authorization. Section 2.3 of the study goes into great length about the technique [14].

Ubuntu Linux [14] "Linux is a free, open-source operating system based on UNIX concepts. The Linux design is sufficiently modular, and Linux servers are thought to be the best for non-local administration. It works in a multi-user scenario. Because of Linux's open-source philosophy, more peer examination of the code is possible in order to detect and repair flaws. However, this does not imply that Linux is secure. Previously, Windows was the most commonly attacked operating system. As a result,

OS	Scanner	Vulnerability
UC 20	Nessus	CWE-200
UC 20	Nessus	IAVT 0001-T-0933
UC 20	Nessus	IAVB 0001-B-0504
UC 20	Nessus	IAVB 0001-B-0515
UC 20	OpenVas	CVE-2018-3639
UC 20	OpenVas	CVE-2017-5715
UC 20	OpenVas	CVE-1999-0524
UC 18	Nessus	CWE-200
UC 18	Nessus	IAVT 0001-T-0933
UC 18	Nessus	IAVB 0001-B-0501
UC 18	Nessus	IAVB 0001-B-0503
UC 18	OpenVas	CVE-2017-5715
UC 18	OpenVas	CVE-2018-3639

Fig. 1 Similar and unsimilar vulnerabilities

most Windows-based servers were converted to Linux. Eventually, the attackers began targeting Linux in this manner.”

Vulnerability profile for Linux [15]

Operating system vulnerabilities are classified using a severity-based vulnerability profiling scheme adapted on upgraded Landwehr categorization [15]. The severity descriptions of common vulnerabilities and exposures (CVE) were used to generate the scoring. The two security additions, hardening and LSM, are combined with the vanilla (simple) version of Linux to create a vulnerability profile that demonstrates the scheme’s utility [15].

Study of security flaws in the Linux Kernel by fuzzing [16]

Kernel fuzzing is performed in this research article using automated software testing techniques to detect potential vulnerabilities by creating wrong input data. Input data might be files handled by the application, information transmitted across network protocols, application interface functions, and so on [16]. During the fuzzing procedure, the researchers employed the KASAN sanitizer to detect kernel security weaknesses such as use-after-free vulnerabilities [16].

The fuzzing process’s outcomes were evaluated by examining the types of crashes detected and the number of instances of each type. Table 1 shows the most prevalent forms of crashes, and the researchers also mentioned the probable implications of UAF exploitation, such as data destruction, emergency system termination, and arbitrary code execution [16].

Table 1 Area of focus in drone security

	Security vulnerabilities of unmanned aerial vehicles and countermeasures: an experimental study [4]	A review of cybersecurity vulnerabilities for unmanned aerial vehicles [5]	A cyber security analysis used for unmanned aerial vehicles in the smart city [6]
Attacks performed	Cracking DJI SDK, GPS Spoofing [4]	GPS spoofing, GPS jamming, Intercept data feed [5]	Man in the middle attacks (MITM), SYN Flood, PING Flood, GPS Spoofing attacks, and UDP flood [6]
Attacks focused on	Some of the attacks focused on the communication link [4]	Focused on the communication link [5]	Focused on the communication link [6]
Focus on operating systems	No	No	No

Security analysis of the Raspbian Linux operating system and its settings to increase resilience against attacks via network interface [17]

The security of the Raspbian Linux operating system and its settings to boost resilience against network interface attacks is performed in this study article by generating five distinct accounts with differing levels of protection and connecting the device to a public internet network [17]. The experiment monitors logs for any suspicious activity from remote attackers and analyses these logs to validate threats and recommend security enhancements. The effectiveness of security measures is assessed based on factors such as the number of successful/failed assaults, the location and type of attacks, and the goal of the attack. The research paper’s technique entails generating five distinct accounts with varied levels of protection and connecting the device to a public internet network. The experiment monitors logs for any suspicious activity from remote attackers and analyses these logs to validate threats and recommend security enhancements [17]. The effectiveness of security measures is assessed based on factors such as the number of successful/failed assaults, the location and type of attacks, and the goal of the attack.

Analysis and study of security mechanisms inside Linux Kernel [7]

The primary security mechanism within the Linux kernel is the Unix-like POSIX permission system, followed by capabilities and access control lists (ACL) authorization as optional components [7]. A mandatory access control (MAC) mechanism imposes a security policy based on specified rules and regulations. It is built into the Linux kernel via the SELinux subsystem, which allows for flexible security policy setup. “The most recent Linux kernel has been integrated with Linux’s mandatory access control mechanism (SELinux).” [7]. The SELinux subsystem is capable of enforcing policy-based Mandatory Access Control (MAC) and providing flexible security policy setup.”

4 Results

4.1 Vulnerability Scanning

Performing vulnerability scanning, using tools like Nessus and OpenVAS (Tables 2 and Fig. 2).

Observations

The number of vulnerabilities detected in the vulnerability scanning phase demonstrates that even without the developers writing their own and deploying on the Ubuntu Core, there are vulnerabilities present which need to be considered.

Though there are vulnerabilities that are the same in both Ubuntu Cores, some distinct vulnerabilities also appear that are unique to their corresponding operating system.

Table 2 Vulnerability scanning results

Sr No	Scanner	Ubuntu core 20	Ubuntu core 18
		Results	Results
1	Nessus scanner	CWE-200	CWE-200
		IAVT 0001-T-0933	IAVT 0001-T-0933
		IAVB 0001-B-0504	IAVB 0001-B-0501
		IAVB 0001-B-0515	IAVB 0001-B-0503
2	OpenVas scanner	CVE-2018-3639	CVE-2017-5715
		CVE-2017-5715	CVE-2018-3639
		CVE-1999-0524	

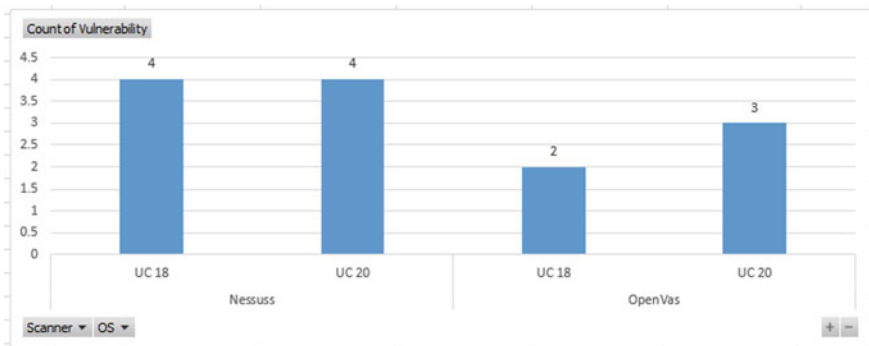


Fig. 2 Results of vulnerability scanning

4.2 Auditing

See Table 3 and Fig. 3.

Observations

Issues with Boot and cryptography did not exist in Ubuntu core 18 but have appeared in the Ubuntu Core 20.

A significant improvement with packages can be seen in the Ubuntu Core 20.

Issues with Accounting, Banner, File system, Hardening logging, Network, ssh and tool seem to persist even in higher version.

4.3 Potential Privilege Escalation Vectors

See Table 4 and Fig. 4.

As we progress to higher versions of Ubuntu Core, we can see major gains in various areas. However, many issues have been brought into newer versions as well.

5 Validation of the Results

For the validation of results, comparative analysis technique will be used, where the same experiments are carried out on the Ubuntu Server, which is the closest to the Ubuntu Core and compare the results.

5.1 Vulnerability Scanning

See Table 5.

5.2 Auditing the System

See Table 6.

Table 3 Auditing results

Sr No	Audit area	Ubuntu core 20	Ubuntu core 18
Warnings			
1	Firewall	Fire-4512	Inptables module(s) loaded, but no rules active
Suggestions			
1	Kernel	KRNL-5820	<p>If not required, consider explicit disabling of core dump in/etc./security/limits.conf file</p> <p>KRNL-5788</p> <p>Determine why/vmlinuz or/boot/vmlinuz is missing on this Debian/Ubuntu system</p>
			<p>If not required, consider explicit disabling of core dump in/etc./security/limits.conf file</p> <p>KRNL-5820</p>
		KRNL-6000	<p>One or more sysctl values differ from the scan profile and could be tweaked</p> <p>KRNL-6000</p> <p>One or more sysctl values differ from the scan profile and could be tweaked</p>
2	Boot	BOOT-5264	<p>Consider hardening system services, run '/usr/bin/systemd-analyze security SERVICE' for each service</p>
3	Authentication	AUTH-9230	<p>Configure password hashing rounds in/etc./login.defs</p> <p>AUTH-9230</p> <p>Configure password hashing rounds in/etc./login.defs</p>
		AUTH-9262	<p>Install a PAM module for password strength testing like pam_cracklib or pam_passwdqc</p> <p>AUTH-9262</p> <p>Install a PAM module for password strength testing like pam_cracklib or pam_passwdqc</p>

(continued)

Table 3 (continued)

Sr No	Audit area	Ubuntu core 20		Ubuntu core 18	
		AUTH-9286	Configure minimum and maximum password age in/etc./login.defs	AUTH-9286	Configure minimum and maximum password age in/etc./login.defs
4	File system	AUTH-9328	Default umask in/etc./login.defs could be stricter like 027	AUTH-9328	Default umask in/etc./login.defs could be stricter like 027
		FILE-6310	Default umask in/etc./login.defs could be stricter like 027	FILE-6310	Default umask in/etc./login.defs could be stricter like 027
		FILE-7524	Consider restricting file permissions	FILE-7524	Consider restricting file permissions
5	Nameserver	NAME-4028	Check DNS configuration for the DNS domain	NAME-4028	Check DNS configuration for the DNS domain
		NAME-4404	Add the IP name and FQDN to/etc./hosts for proper name resolving	NAME-4406	Resolving between localhost and the hostname of the system
6	Packages	PKGS-7398	Install a package audit tool to determine vulnerable packages	PKGS-7346	Purge old/removed packages (4 found) with aptitude purge or dpkg-purge command. This will cleanup old configuration files, cron jobs and startup scripts
				PKGS-7370	Install debsums utility for the verification of packages with known good database

(continued)

Table 3 (continued)

Sr No	Audit area	Ubuntu core 20		Ubuntu core 18	
				PKGS-7394	Install package apt-show-versions for patch management purposes
				PKGS-7398	Install a package audit tool to determine vulnerable packages
				PKGS-7420	Consider using a tool to automatically apply upgrades
7	Network	NETW-3200	Determine if protocol 'dcp', 'setp', 'rds', 'tipc' is really needed on this system	NETW-3200	Determine if protocol 'dcp', 'setp', 'rds', 'tipc' is really needed on this system
8	SSH	SSH-7408	<p>Consider hardening SSH configurations</p> <ul style="list-style-type: none"> - AllowTcpForwarding - ClientAliveCountMax - LogLevel - MaxAuthTries - MaxSessions - Default Port - TCPKeepAlive ara> - X11 Forwarding - AllowAgentForwarding 	<p>Consider hardening SSH configurations</p> <ul style="list-style-type: none"> - AllowTcpForwarding - ClientAliveCountMax - LogLevel - MaxAuthTries - MaxSessions - Default Port - TCPKeepAlive - X11 Forwarding - AllowAgentForwarding - Compression - UsePrivilegeSeparation 	
9	Logging	LOGG-2154	Enable logging to an external logging host for archiving purposes and additional protection	LOGG-2154	Enable logging to an external logging host for archiving purposes and additional protection

(continued)

Table 3 (continued)

Sr No	Audit area	Ubuntu core 20		Ubuntu core 18	
		BANN-7126	Add a legal banner to/etc./issue, to warn unauthorized users	BANN-7126	Add a legal banner to/etc./issue, to warn unauthorized users
10	Banner	BANN-7130	Add legal banner to/etc./issue.net, to warn unauthorized users	BANN-7130	Add legal banner to/etc./issue.net, to warn unauthorized users
11	Accounting	ACCT-9622	Enable process accounting	ACCT-9622	Enable process accounting
		ACCT-9626	Enable sysstat to collect accounting	ACCT-9626	Enable sysstat to collect accounting
		ACCT-9628	Enable audit to collect audit information	ACCT-9628	Enable audit to collect audit information
12	Cryptography	CRYP-7902	Check available certificates for expiration		
13	File integrity	FINT-4350	Install a file integrity tool to monitor changes to critical and sensitive files	FINT-4350	Install a file integrity tool to monitor changes to critical and sensitive files
14	Tools	TOOL-5002	Determine if automation tools are present for system management	TOOL-5002	Determine if automation tools are present for system management
15	Hardening	HRDN-7230	Harden the system by installing at least one malware scanner, to perform periodic file system scans	HRDN-7230	Harden the system by installing at least one malware scanner, to perform periodic file system scans

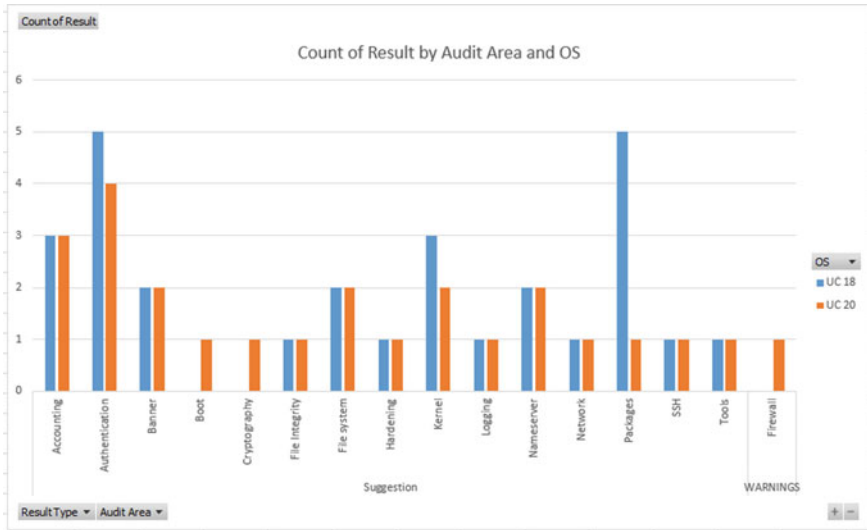


Fig. 3 Comparison of the results of audit

Table 4 Results of privilege escalation scan

Sr No	Potential PE vector	Ubuntu core 20		Ubuntu core 18	
1	Exploits suggested by exploit suggester	CVE-2021-3156	Sudo Baron Samedit 2/ Samedit	CVE-2021-3156	Sudo Baron Samedit 2/ Samedit
		CVE-2022-32250	nft_object UAF (NFT_MSG_NEWSET)	CVE-2022-32250	nft_object UAF (NFT_MSG_NEWSET)
		CVE-2022-2586	nft_object UAF	CVE-2022-2586	nft_object UAF
		CVE-2021-27365	Linux-iscsi	CVE-2021-27365	Linux-iscsi
		CVE-2021-22555	Netfilter heap out-of-bounds write	CVE-2021-22555	Netfilter heap out-of-bounds write
		CVE-2019-18634	Sudo pwfeedback	CVE-2019-18634	Sudo pwfeedback
2	SUID—bit set	/passwd		/passwd	
		/sudo		/sudo	
		/chfn		/chfn	
		/newgrp		/newgrp	

(continued)

Table 4 (continued)

Sr No	Potential PE vector	Ubuntu core 20	Ubuntu core 18
		/pppd	/pppd
		/mount	/mount
		/umount	/umount
3	Capabilities	cap_chown	cap_chown
		cap_net_bind_services	cap_net_bind_services
		cap_net_admin	cap_net_admin
		cap_net_raw	cap_net_raw
4	Other Priv Esc vectors	Writable socket files	Writable socket files
		Writeable timer files	Writeable timer files
			Sudo version 1.8.16

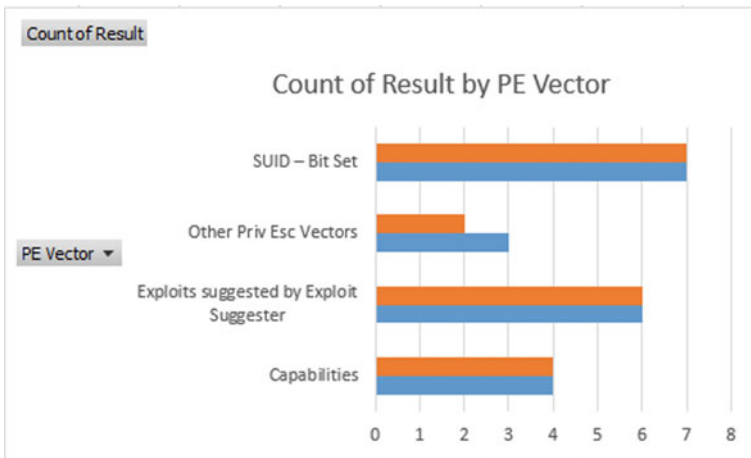


Fig. 4 Comparison of the privilege escalation results

5.3 Privilege Escalation

Upon comparing the results of security evaluations of the Ubuntu core 18 and 20, it can be noticed that the Ubuntu Core operating systems have inherited a lot of security of issues from the Ubuntu server, which also leads to a conclusion that, any exploits developed which work successfully on either of the variant, might also work on the other variant (Table 7).

Table 5 Validation of vulnerability scanning

Sr No	Scanner	Ubuntu core 20	Ubuntu server	Observations
		Results	Results	
1	Nessus scanner	CWE-200	CWE-200	On comparison it is found that some vulnerabilities have are similar and in both the operating systems and each of them have their distinct vulnerabilities as well
		IAVT 0001-T-0933	IAVT 0001-T-0933	
		IAVB 0001-B-0504	IAVT 0001-T-0502	
		IAVB 0001-B-0515	IAVB 0001-T-05016	
2	OpenVAS scanner	CVE-2018-3639	CVE-2018-12207	
		CVE-2017-5715	CVE-2018-3639	
		CVE-1999-0524	CVE-2022-40982	
			CVE-2022-29900, CVE-2022-29901	
			CVE-1999-0524	

Table 6 Validation of results of auditing

Sr No	Audit area	Ubuntu core results	Ubuntu server results	Observations
1	Kernel	KRNL-5788	KRNL-5820	The issues found with the Kernel with of the Ubuntu core and the Ubuntu server do not tend to deviate much
		KRNL-5820	KRNL-6000	
		KRNL-6000		
2	Boot	BOOT-5264	BOOT-5122	Issue with the boot also tend to be the same
			BOOT-5264	
3	Authentication	AUTH-9230	AUTH-9230	The Ubuntu core seems to have inherited the authentication and file system issues as it is from the server version the Ubuntu operating system
		AUTH-9262	AUTH-9262	
		AUTH-9286	AUTH-9286	
		AUTH-9328	AUTH-9328	
		AUTH-9228	AUTH-9228	
4	File system	FILE-6310	FILE-6310	
		FILE-7524	FILE-7524	
5	Nameserver	NAME-4028	NAME-4028	Suggestion have been made to check the DNS info across both the core and the server
		NAME-4404		
		NAME-4406		

(continued)

Table 6 (continued)

Sr No	Audit area	Ubuntu core results	Ubuntu server results	Observations
6	Packages	PKGS-7346	PKGS-7370	Majority issues identified regarding Package management seem to have been delt with. However, issues such as missing debsusms utility for package verification, missing packages for patch management and missing package audit tool still persist
		PKGS-7370	PKGS-7394	
		PKGS-7394	PKGS-7398	
		PKGS-7398		
		PKGS-7420		
7	Network	NETW-3200	NETW-3200	
8	SSH	SSH-7408		While auditing SSH a list of suggestions for improving SSH on Ubuntu core for both 18 and 20 version was provided, contrary to be the result of auditing SSH on the Ubuntu Server, not even a single improvement was suggested
9	Logging	LOGG-2154	LOGG-2154	Issues with logging, banner and accounting, tends to persist
10	Banner	BANN-7126	BANN-7126	
		BANN-7130	BANN-7130	
11	Accounting	ACCT-9622	ACCT-9622	
		ACCT-9626	ACCT-9626	
		ACCT-9628	ACCT-9628	
12	Cryptography	CRYP-7902		No issues with cryptography were detected on the Ubuntu server
13	File integrity	FINT-4350	FINT-4350	The issues with File integrity, tools and hardening are the same on both
14	Tools	TOOL-5002	TOOL-5002	
15	Hardening	HRDN-7230	HRDN-7230	

6 Discussion and Conclusions

For various reasons, this study project on analyzing the security of open-source operating systems used in drones is critical.

Table 7 Validation of privilege escalation scan results

Sr No	Potential PE vector	Ubuntu core 20	Ubuntu server	Observations
1	Exploits suggested by exploit suggester	CVE-2021-3156	CVE-2022-2586	Most of the exploits identified on Ubuntu core 18, Ubuntu core 20 and Ubuntu server 20 are the same
		CVE-2022-32250	CVE-2021-4034	
		CVE-2022-2586	CVE-2021-3156	
		CVE-2021-27365	CVE-2021-22555	
		CVE-2021-22555	CVE-2022-32250	
		CVE-2019-18634	CVE-2017-5618	
2	SUID—bit set	/passwd	/snap-confine	The binaries on which the SUID bit is set are identical on the Ubuntu Cores and some all of them also exist on the Ubuntu Server, existence of pkexec binary, also confirms the PwnKit exploit identified by the exploit suggester
		/sudo	/umount	
		/chfn	/mount	
		/newgrp	/pkexec	
		/pppd	/sudo	
		/mount	/passwd	
		/umount	/newgrp	
		/at		
3	Capabilities	cap_chown	cap_chown	The capabilities identified are exactly the same, across all the variants of the Ubuntu core and Ubuntu Server tested
		cap_net_bind_services	cap_net_bind_services	
		cap_net_admin	cap_net_admin	
		cap_net_raw	cap_net_raw	
4	Other Priv Esc vectors	Writable Socket files	Writable Socket files	Vulnerable sudo version exists on the Ubuntu Core 18 and the Ubuntu Server 20. There are writable socket files with helps attackers gain persistence on the system. Timer file work similar to the cron jobs, having corrects permissions set on them is crucial
		Writeable timer files	Writeable timer files	
		Sudo version 1.8.16 (UC 18)	Sudo version 1.8.31	

To begin, the growing popularity and broad use of unmanned aerial vehicles (UAVs) or drones has resulted in significant benefits in a variety of industries. However, with growing usage comes the potential of malicious actors posing security vulnerabilities. Drones can be targeted and compromised, potentially causing harm to people, property, and the drones themselves. As a result, it is critical to ensure the safety and security of these devices.

Second, both enthusiasts and companies build drones using open-source operating systems as a basis or starting point. However, some enthusiasts have been caught directly stealing open-source operating systems without considering the security consequences. Because these operating systems are open source, anyone can

contribute code to them, potentially introducing accidental flaws. As the capabilities and applications of drones in mission-critical systems develop, the need for robust and secure operating systems becomes increasingly important.

The project attempts to bridge a knowledge gap in the security evaluation of open-source drone operating systems. While these systems are popular among enthusiasts and some businesses, they are not generally used in the consumer market. As a result, their vulnerabilities may not have been identified or addressed adequately. This study aims to discover potential security flaws and vulnerabilities by executing purple teaming operations on drones equipped with several open-source operating systems.

It is necessary to close this knowledge gap in order to have a better understanding of the security threats connected with open-source drone operating systems. It will provide vital insights into the vulnerabilities present in these systems, allowing security researchers, hobbyists, and businesses to take the required safeguards and design better UAV operating systems. The research findings can be used to build more secure and resilient drone operating systems in the future. Furthermore, it will spur further research in the subject of drone security, resulting in innovations and safer drone operations in a variety of applications and businesses.

The studies are conducted in a simulated environment using acceptable and available opensource tools, installing the operating systems on the Raspberry Pi 4 and the tools on a Kali Linux Virtual Linux.

6.1 Limitations of the Study

1. Samples and selection:

Because the abstract does not specifically indicate the selection method for the drones or open-source operating systems to be examined, the sample selection criteria may be opaque. The abstract makes no mention of the number or features of the drones or operating systems used, which could lead to bias in the findings.

The abstract states that “multiple open-source operating systems will be evaluated,” but no precise number or reason for the sample size is given. It may be difficult to generalize the findings to a larger population of drones and operating systems without a sufficient sample size.

2. Lack of previous research studies on the topic:

According to the abstract, open-source operating systems for drones are not widely employed in the consumer sector, and their vulnerabilities have not yet been discovered. Because there have been no previous research studies on the issue, the availability of related material and prior information to build on in the study may be limited.

3. **Limited access to data:**

Because the study focuses on analyzing the security of open-source operating systems used in drones, there may be a scarcity of relevant and thorough data on these specific operating systems. Access to proprietary or sensitive data from organizations or manufacturers may be restricted, making it difficult for researchers to undertake a full examination.

4. **Time constraints for a master's dissertation:**

The study must be finished within three months, which can place severe time limits on numerous research tasks. Conducting extensive vulnerability scanning, networking scanning, and other security assessments on numerous open-source operating systems can take time, which can limit the depth and breadth of the research. The time constraint may also have an impact on the study's breadth, restricting the inclusion of a larger sample size or a more complete analysis of vulnerabilities.

5. **Scope limitations:**

The time constraints may limit the scope of the study, preventing the researchers from exploring certain aspects of the research question in greater detail. With a limited timeframe, the study might have to focus on a specific subset of open-source operating systems or perform a less comprehensive assessment than would be ideal.

6. **Resource constraints:**

Conducting purple teaming activities on drones and assessing alternative open-source operating systems may necessitate the use of specialized resources, equipment, and experience, which may be limited by time constraints and academic constraints. These methodological restrictions can have an impact on the study's results and may prevent researchers from delving further into specific parts of the research subject. To address these limits, researchers should consider optimizing their research approach, focusing on essential areas of risk assessment, and clearly noting any limitations owing to limited data access and time in the study's discussion. Given the limits they faced, they may also emphasize the significance of future studies to supplement and build on their findings. To address these limitations, the researchers could acknowledge these constraints in their study's discussion section and provide transparency regarding the scope and generalizability of their findings. They might also highlight areas where future research with more extensive data and time resources could enhance the understanding of UAV operating system security.

6.2 *Further Work*

1. **Expanded data access and analysis:**

In the future, efforts can be made to gain greater and more thorough access to data relevant to open-source operating systems used in UAVs. Researchers can collaborate with industry partners or manufacturers to gain access to confidential data that can shed light on the security of these operating systems. With a larger and more diverse dataset, researchers may conduct a more thorough inquiry and validate the current study's conclusions.

2. **Long-term and in-depth analysis:**

Given the time constraints of a master's dissertation, lengthier study periods can be designed in the future to allow for more extensive and in-depth analysis of the security of open-source operating systems. With extended study durations, researchers will be able to investigate other areas of security evaluation, such as real-world testing, simulation-based assessments, and analyzing growing threats and vulnerabilities over time.

3. **Comparative analysis with proprietary systems:**

A comparison of open-source operating systems with proprietary operating systems used in UAVs could be part of future study. Researchers can acquire useful insights into the merits and shortcomings of open-source solutions by comparing the security features, vulnerabilities, and performance of both types of systems.

4. **Real-world drone testing and red teaming:**

Future study could integrate real-world drone testing, in which UAVs with open-source operating systems are deployed in controlled areas to assess their security under actual conditions, to supplement the findings from purple teaming operations. Red teaming exercises can also be used to simulate genuine cyber-attacks on UAVs and highlight potential flaws that may not be revealed during controlled purple teaming.

5. **User behavior and security awareness studies:**

These methodological restrictions can have an impact on the study's results and may prevent researchers from delving further into specific parts of the research subject. To address these limits, researchers should consider optimizing their research approach, focusing on essential areas of risk assessment, and clearly noting any limitations owing to limited data access and time in the study's discussion. Given the limits they faced, they may also emphasize the significance of future study to supplement and build on their findings. Given the potential impact of human behavior on UAV security, further research should involve surveys or interviews with hobbyists, businesses, and other stakeholders to better understand their security concerns and practices. By investigating user behavior and security awareness, researchers can identify opportunities for educational initiatives and better security practices.

6. Integration of machine learning and AI-based defense mechanisms:

Future studies could look into incorporating machine learning and artificial intelligence (AI) approaches as part of UAV defense measures. These technologies have the potential to improve autonomous threat detection, response, and adaptation to new security concerns.

7. Scalability and performance evaluation:

Future research can focus on analyzing the scalability and performance of open-source operating systems in managing sophisticated UAV operations and data processing as drone functionality evolves. The study's limitations of limited data access and time limits can be alleviated by addressing the above areas in future research, and a more thorough understanding of UAV operating system security can be gained.

Overall, we have discussed the aims and objectives of this study and whether or not all the objectives were achieved. The key takeaways could be that despite the best effort of the Ubuntu Core team to make the Ubuntu Core operating system secure, it still has a lot of vulnerabilities. The security issues of the Ubuntu Core also exist on the Ubuntu Server or even vice versa. When it comes to the drones most of the issues that are discussed are of the network part, which draws the attention away from the operating system and its security flaws.

References

1. Iqbal S (2021). A study on UAV operating system security and future research challenges. <https://doi.org/10.1109/CCWC51732.2021.9376151>
2. Kumar CRS, Mohanty S (2021) Current trends in cyber security for drones. In: 2021 International Carnahan Conference on Security Technology (ICCST). IEEE, pp 1–5
3. Marshall DM, Barnhart RK, Hottman SB, Shappee E, Thomas M (2016). Introduction to unmanned aircraft systems. CRC Press
4. Dey V, Pudi V, Chattopadhyay A, Elovici Y (2018) Security vulnerabilities of unmanned aerial vehicles and countermeasures: an experimental study. In: 31st International conference on VLSI design and 2018 17th international conference on embedded systems (VLSID). <https://doi.org/10.1109/vlsid.2018.97>
5. Krishna CGL, Murphy RR (2017) A review on cybersecurity vulnerabilities for unmanned aerial vehicles. In: IEEE international symposium on safety, security and rescue robotics (SSRR). <https://doi.org/10.1109/ssr.2017.8088163>
6. Cebeloglu FS, Karakose M (2019). A cyber security analysis used for unmanned aerial vehicles in the smart city. <https://doi.org/10.1109/UBMYK48245.2019.8965591>
7. Zhai G, Li Y (2008) Analysis and study of security mechanisms inside Linux Kernel. International conference on security technology. <https://doi.org/10.1109/sectech.2008.17>
8. Richards RJ (2010) Modeling and security analysis of a commercial real-time operating system kernel. Springer EBooks, pp 301–322. https://doi.org/10.1007/978-1-4419-1539-9_10

9. Luna R, Islam SA (2021) Security and reliability of safety-critical RTOS. *SN Comput Sci* 2(5). <https://doi.org/10.1007/s42979-021-00753-y>
10. Yu Weider D (2010) Real-time operating system security
11. Dieber B, Breiling B, Taurer S, Kacianka S, Rass S, Schartner P (2017) Security for the robot operating system. *Robot Auton Syst* 98:192–203. <https://doi.org/10.1016/j.robot.2017.09.017>
12. Recchia L, Procopio G, Onofrii A, Rogo F (2014) Security evaluation of a Linux system: common criteria EAL4+ certification experience. <https://doi.org/10.1109/ISSREW.2014.45>
13. Zhang T, Tech V, Shen W, University Z, Lee D, Azab A, Wang R (2019) PeX: a permission check analysis framework for Linux Kernel PeX: a permission check analysis framework for Linux Kernel. <https://www.usenix.org/system/files/sec19-zhang-tong.pdf>
14. Sedano WK, Salman M (2021) Auditing Linux operating system with center for internet security (CIS) standard. <https://doi.org/10.1109/ICIT52682.2021.9491663>
15. Wita R, Teng-Amnuay Y (2005). Vulnerability profile for Linux. <https://doi.org/10.1109/AINA.2005.343>
16. Teplyuk PA, Yakunin AG, Sharlaev EV (2020) Study of security flaws in the Linux Kernel by fuzzing. In: International multi-conference on industrial engineering and modern technologies (FarEastCon). <https://doi.org/10.1109/foreastcon50210.2020.9271516>
17. Gallus P, Frantis P (2021) Security analysis of the Raspbian Linux operating system and its settings to increase resilience against attacks via network interface. In: International conference on military technologies (ICMT). <https://doi.org/10.1109/icmt52455.2021.9502746>

A Proposed Permissioned Blockchain Consensus Algorithm: Consensus Algorithm Genetically Enhanced (CAGE)



Ian Mitchell  and Kamil Maka

Abstract Blockchain technology is disruptive and relies on the development of consensus algorithms, CAs. Currently, there are many CAs proposed and this paper explores a new avenue of nature-inspired CAs. In particular, emphasis is on selection techniques used in canonical genetic algorithms, GAs, combined with a stamina attribute. The premise is that selection techniques used in GAs could be adapted and used to select nodes responsible for adding blocks in permissioned blockchain networks. Various selection techniques were researched and the Fitness Proportionate Selection was chosen as the most suitable, which was tested on a blockchain network and executed many times and compared to another CA for permission blockchain networks, Proof-of-Elapsed-Time, PoET. The results of the experiments show an improvement in transaction throughput and block creation when compared with PoET. Our initial results, however, had a biased node selection and could be vulnerable to take-over attacks. Therefore, a stamina element was introduced and the experiments were repeated with the results showing a less biased node selection. The overall contribution is the exploration of nature as a solution to CAs. The initial results are comparable and in places better than existing CAs, such as PoET.

Keywords Blockchain · Consensus algorithms · Evolutionary computation

1 Introduction

The perception of Blockchain Technology, BCT, is inevitably linked to cryptocurrency. This has been referred to as Blockchain 3.0, and more generically as Web 3.0. Whilst the advances in cryptocurrency assets cannot be denied or ignored, this perception that all blockchain applications are cryptocurrency related is shortsighted. In [1], Blockchain 3.0 is described as, "... wider spectrum of non cryptocurrency-related

I. Mitchell (✉) · K. Maka
Middlesex University, London NW44BT, UK
e-mail: i.mitchell@mdx.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_3

distributed ledger uses...”. These applications can be completed on cryptocurrency platforms, whilst still being independent of the cryptocurrency itself, and have wide and varied applications, e.g., [2, 3]. Many of these applications are completed on permissioned blockchains that do not require cryptocurrency and therefore, do not require incentivised consensus algorithms.

In [4] the authors go further and identify ten domains where BCT is used in industry: Finance; Healthcare; Logistics; Manufacturing; Energy; Agriculture and Food; Robotics; Entertainment (Gaming); Construction; and, Telecommunications. There are many other industries that are interested in blockchain, particularly those that have a central body of authority and require auditing of members, e.g., gun ownership [5]; chain of custody used in forensic sciences [6]; industries that require provenance, such as antiquities [7] or arts [8]; and, finally, there are applications to IoT [9].

So, BCT can be applied to wide and varied domains and is the motivation for researching components. This paper is focused on consensus algorithms for fault-tolerant permissioned BCT.

1.1 Ethics and Sustainability

Before embarking on this journey of research there are some considerations to be made about Ethics and Sustainability that are discussed briefly in this subsection.

The eradication of data stored on a blockchain is difficult due to its decentralised nature, where there are 100 s of nodes then each of these nodes would have to delete the data. This is true of permissionless BCT, however, for permissioned BCT there is some control over the distribution of the nodes. Therefore, eradication is made considerably easier. In addition to this, the data generated for each block is random and does not include personal details, which would be a mistake for any BCT. Ethical approval was granted for the project and was deemed a minimal risk.

BCT has been identified in [10] as the next General Purpose Technology, GPT. History has taught us that the introduction of GPTs can have a detrimental impact on the environment and the UNs Sustainability Development Goals, SDG, and energy consumption caused by Proof-of-Work Consensus Algorithms is significant [11] and in some cases even considered unsustainable [12]. BCT challenges item 12 in the UN’s SDG (see [13]). This can lead to resistance to adopting BCT, which can further prejudice the understanding and thwart innovation of such technologies [14].

CAGE is an algorithm that is trying to reduce the energy in finding consensus in BCT, and therefore, achieve a responsible use of energy in the application of BCT.

2 Blockchain Anatomy

There are two types of BCT: permissioned and permissionless. Permissionless BCT, as the name suggests, is an open invitation to all and sundry to join the network. This approach creates a trilemma, detailed in [15], between Scalability, Security and Decentralisation. Want an increase in Scalability, then this cannot happen without compromising security or decentralisation.

Permissioned BCT has distributed nodes that are verified members. Membership restrictions prevent all and sundry to join and thus become less susceptible to some cyber-attacks, e.g., Sybil attacks. Permissioned BCTs are required to be only crash fault tolerant. So, BCT is possible due to the following characteristics.

Security: transactions are completed using encryption, this provides a secure environment that every industry is concerned with. It also promotes anonymity and via hashing [16] provides the immutability or append-only property that is conducive and attractive to many of the domains aforementioned.

Smart Contracts, SC: An important development of BCT, although optional. Smart Contracts, SC, are coded contracts that allow autonomous updates or transactions. When certain pre-conditions are reached the SC can autonomously change state. SCs are not an integral part of the Consensus Algorithm, and therefore not considered in this paper.

Transaction: A transaction is an exchange of information between users. Naturally, this leads to cryptocurrency and value, however, in many BCTs the information exchange is important, e.g., in [6] the exchange of evidence between custodians contributes to the chain of custody and ultimately to evidential integrity, which should never be altered and therefore BCT is conducive to solving this problem.

Block: A block can contain one or many transactions. The transactions are pooled together. A block is created and includes the hash of the previous block and hence forms a blockchain. The exception to this is the first block, often referred to as the genesis block.

Trust: In any centralised system trust is given to the governing body, blockchain is decentralised and trust is distributed by independent nodes. How these nodes interact with the users and add blocks to the system is down to the consensus algorithm. All nodes and users have a mutual distrust of each other, the consensus algorithm is responsible for independently creating the trust between users making the transaction. In other systems, the level of trust between users is high, which usually results in a highly centralised system. Often distributed ledger technology has been referred to as trustless systems. With permissioned BCT the level of trust is higher, and this results in a less resource-intensive CA. This therefore brings us to a general rule, an increase in trust \propto a decrease in CA computation.

Repudiation: BCT has been referred to as distributed ledger technology. The distributed ledger is important and each node in the network has a copy of the ledger, a blockchain. Every addition to the blockchain records one of the transactions, and therefore, this makes this repudiation proof.

Transparency: The SC is normally stored on the blockchain and therefore is transparent, as is any transaction. Normally, a transaction will be made visible to members of the blockchain and can depend on their access controls.

Immutability: The much-hyped immutability is possible due to the above characteristics. What makes immutability possible is that when a block is added, it contains the hash of the previous block. The blocks cannot be changed without the approval of the consensus algorithm, the recursive hash also makes this computationally expensive as the block height increases.

Consensus Algorithms, CA: CAs are essential to any BCT network. The CA has to be fault tolerant in permissioned BCT and solve Byzantine Fault Tolerance in permissionless networks. The CA requires an agreement between the nodes, once this is reached the block can be prepared with the previous block hash and appended to the chain.

3 Genetic Algorithms

The proposed Consensus Algorithm, CA, is designed for permissioned blockchains. This means an approach can be used to reduce the computational redundancy that exists in many other CAs on permissionless blockchains. A single node is selected to calculate the block, and if two-thirds of the nodes agree in the network, then the block is then added to the blockchain. There is some repetition in the latter part of the consensus, but the former part of the CA is done only once. Therefore, this reduces computations and thus energy consumption.

The selection of the node is of importance in permissioned networks, and that is what the rest of the paper discusses.

The canonical Genetic Algorithm [17] GA, has five basic stages: Initialisation; Selection; Crossover; Mutation; and, Replacement. There are many varieties of each stage, but the stage of interest in this paper are the Selection Techniques [18].

The selection algorithm identified as most appropriate for BCT CA is based on fitness-proportionate selection. There were many considerations for this option are detailed as follows:

Latency: Unlike nature, BCT does not have millions of years to evolve. To be a comparable CA it is known that the selection method would work in milliseconds, see [19], and therefore fitness proportionate selection is efficient if the evaluation is instant. The speed at which the nodes' calculation of fitness needs to be considered when designing the CA, and is in milliseconds.

Trust: Selecting weaker individuals increases diversity in evolutionary algorithms, whilst stronger individuals increase the fitness of the population. This battle between exploration (low selection pressure) and exploitation (high selection pressure) is an important balance to get right. However, the selection pressure in a permissioned BCT is reflected in wrong and right answers. Simply put, if the selected node produces an error, then it is heavily penalised via its fitness and not to be trusted. This leads to lower selection pressures when all nodes are trustworthy.

Offspring: There is no crossover or mutation. The re-evaluation of offspring is unnecessary. Once a node has been selected, then its fitness and stamina are re-evaluated.

Decentralisation: Diversity is important to both GAs and CA but for different reasons. The diversity of node selection in BCT is required to maintain the decentralisation characteristics, e.g. if the same node was selected for every transaction then the system has an architecture of decentralisation but implemented in a centralised way.

Speed: Typically, permissioned BCT has 10s of nodes, not 1000s. They also remained fixed sized and rarely change. Unlike permissionless networks that encourage nodes to be added or removed from the network. The constant and small number of nodes in the network means that tournament selection and other dynamic selection techniques can be ruled out. More traditional fitness proportionate selection techniques can be deployed. The other issue with tournament size is the time it takes to select the nodes to compete in the tournament. Normally, tournament selection is preferred for its efficiency [20], however, due to the modifications on fitness proportionate selection, there is less time spent on fitness scaling.

Fair: The calculation of the fitness and evaluation functions can be computationally expensive in GAs. Unlike in nature, fairness was required to explore, rather than exploit, and weaken the selection pressure. This is at the heart of decentralisation and results are analysed for the distribution of the selection of nodes.

Stamina: Age-based [21, 22] GA strategies can allow the selection or survival of an individual into the next generation. The age-based strategy cannot deploy node death (for fault tolerance the node can be temporarily unavailable), however, after selection this can reduce the likelihood of immediate selection thereafter. This concept is similar to the coin age in Proof-of-Stake explained in [23].

Consideration and recommendations for developing a novel CA are as follows:

- Comparison and similarity.
- Measure node distribution.
- Measure transactions per second.
- Measure blocks per second.

From the above, there are some clear points to favour the deployment of fitness proportionate selection to select nodes in a CA.

4 Cage

In [24] there is a taxonomy of Consensus Algorithms. The proposed CAGE will fall under the non-incentivised category. These algorithms are required to be crash fault tolerant, where $n = 3f + 1$. The non-incentivised CAs work well on permissioned BCT networks and in certain situations can process 1000 s of transactions per second (tps) [19].

Briefly, the BCT network has n validator nodes and m partial nodes. The validator nodes are responsible for the confirmation, before each block can be added a validation of the block is required. The partial nodes are users and not considered in the experiments.

A fitness proportionate selection method is chosen and in particular Roulette Wheel Method is implemented, whereby the fitness of the individual is based on the calculation returned. The individuals are synonymous with nodes in the permissioned blockchain system and therefore, nodes returning invalid calculations or are unreliable and penalised. The fitness of that node is decreased and therefore they are less likely to be selected in the near future. If a node has a valid selection then the fitness increases. Unlike GAs, whereby the genetic value alters, here there is no calculation necessary. This could cause an issue with the distribution of nodes selected, typically leading to biasedness caused by exploitation. Therefore, a stamina attribute is added, and when a node has been selected and returns a valid calculation then the stamina is reduced and the fitness increased. This is to ensure low-selection pressure and therefore better node distribution. The results not only monitor latency and throughput but also the distribution of selected nodes.

The schema in Fig. 1 outlines the proposed CA, CAGE. The schema illustrates the important areas of any CA development: node initialisation; transaction setup; BCT initialisation; Consensus; and, after confirmation adding the block to the chain. There are some differences between CAGE and other CAs and these are the use of Roulette Wheel to select the node, and the re-evaluation of the node (node info). These are calculated differently in CAGE and it is this difference that is proposed, analysed and summarised in the rest of this paper.

All nodes start with equal fitness and stamina values of 10 and 100, respectively. Calculate all fitness values with the stamina of 100 and likewise for nodes with less than 100 stamina. The selection process is based on a random number that is between 0 and the sum of the fitness values; the sum can be calculated as a moving total and not recalculated after each generation. Initially, each node has an equal probability of being selected. However, when a node is selected the fitness and stamina value are modified as follows:

- Node selection is based on a combination of fitness and stamina
- Node fitness is increased upon selection and reliability
- Node stamina is decreased upon selection
- Node fitness is decreased when developing a fault
- Node stamina remains when not selected.

4.1 Differences Between PoET and CAGE

PoET predominantly relies on random selection and therefore it is expected that the distribution between nodes is even. With CAGE it is expected that the distribution is biased. The difference between them is illustrated and demonstrated in Figs. 4 and 5. The main difference is due to feedback after the node calculation and the re-evaluation

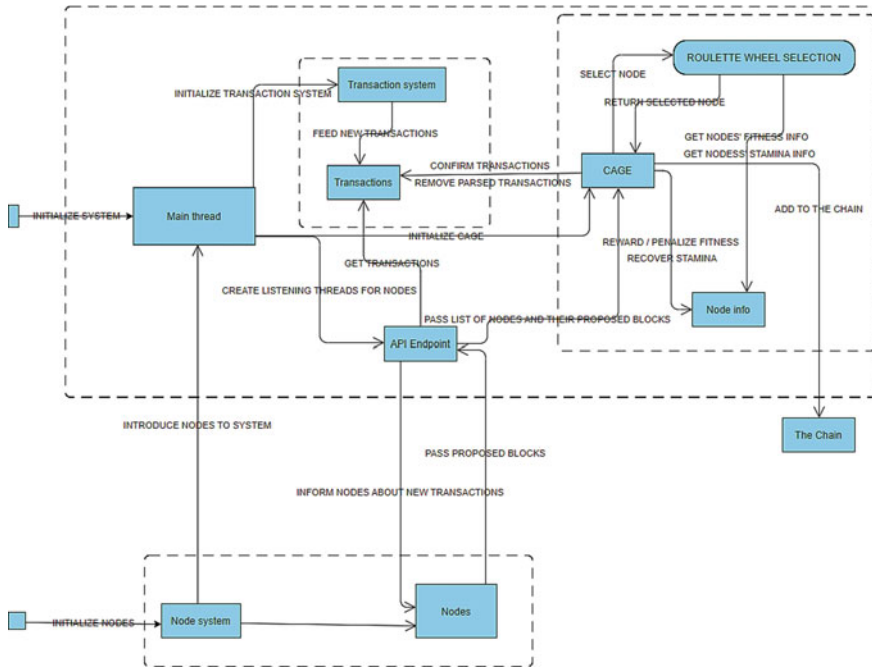


Fig. 1 Schema diagram for software implementation of CAGE

of the node’s fitness. This is also compounded by any fault tolerance. When a fault is detected, be it due to unavailability or inaccuracy, the node is penalised.

5 Method

The experiment will compare PoET and CAGE and the performances on:

- Transactions: the number of transactions per second
- Blocks: the number of blocks produced per second
- Nodes: the distribution of node selection

Some parameters were chosen and each run lasted 300,000 ms. This was considered adequate time to get enough data for comparative analysis. It has to be fault tolerant, FT, not Byzantine Fault Tolerant, BFT. This was simulated using approx. 5% randomised node failures, which were then identified by the CAGE as unreliable and a reduction in the node’s fitness attribute. The changing parameters are identified in Table 1.

Three nodes are difficult to test for fault tolerance since it can be difficult to reach confirmation with 33% of the nodes unavailable. Therefore the 3-node network was

Table 1 Parameters for testing: there were a variety of number of nodes using both PoET and CAGE. Each test was run three times over 300,000 ms. The block size was chosen at random between 1 and 5 transactions

Parameters	Values
Nodes	3, 10, 20
CA	PoET, CAGE
Time	300,000 ms
Block size	1–5 transactions
Runs	3

faultless, and no simulation of fault tolerance was tested. This test was completed as a control experiment and used to see if the system initially works. These are preliminary results and are not subject to the same tests completed on 10 and 20 nodes. It is included to show all results collated for the system.

6 Results and Analysis

The results are comparable and explained in each of the following subsections, covering the main objectives: Transactions; Blocks; and, Nodes.

Initial results of the distribution of the node selection showed there was a bias towards certain nodes, the larger the number of nodes in the system the larger the differences were exposed. Figure 5 illustrates a clear bias towards certain node selections compared with PoET which is more even. However, this bias does not exclude the selection of all nodes and therefore, is acceptable.

6.1 Transactions

Normally, transactions are combined such that a single block may contain multiple transactions. There can be more transactions than blocks, due to a single block having multiple transactions. The number of transactions in each block was chosen at random between 1 and 5. This is an effort to simulate other BCT that has multiple transactions per block. Figure 2 shows how different configurations of CAGE performed. As expected more nodes resulted in fewer transactions, due to increasing overheads. The results are comparable with PoET, which are slightly lower at 20 nodes.

The TPS for this experiment decreases as the number of nodes increase, so does the efficiency of TPS and CAGE operates at approx. 41%, whereas PoET has approx. 32% efficiency.

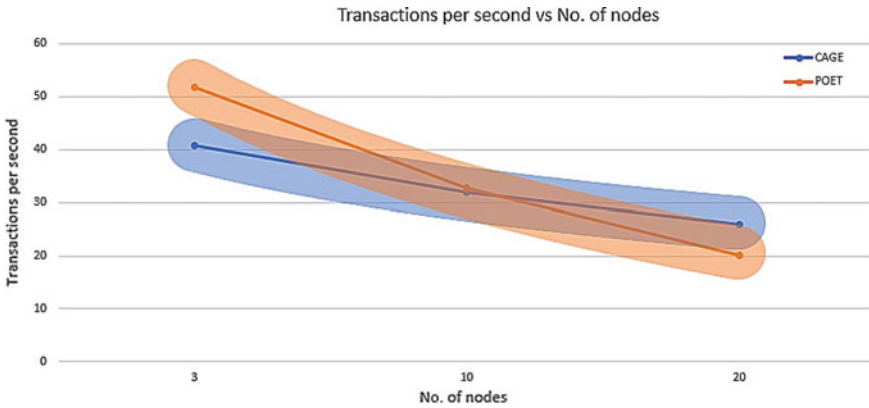


Fig. 2 Transactions per second, TPS, plotted against each experiment for 3, 10 and 20 nodes. Linear interpolation used for number of nodes between 3–10 and 10–20

6.2 Block Latency

Figure 3 shows the number of Blocks Per Second, BPS, compared between the two models. As expected, as the number of nodes increase, there is an overhead of reaching consensus between more nodes and therefore the number of BPS decreases.

It can be seen that CAGE slightly outperforms PoET. With 10 nodes there are approx. 24 blocks per second created for CAGE, compared to approx. 12 blocks per second for PoET. With 20 nodes there are approx. 18 blocks per second for CAGE,

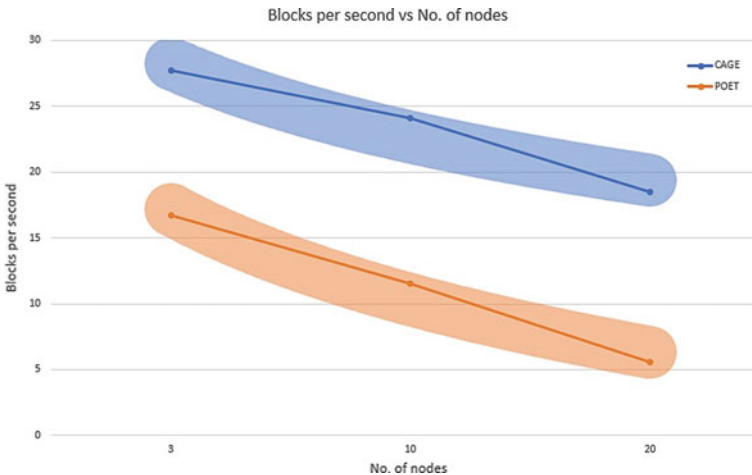


Fig. 3 Blocks per second, BPS, plotted against each experiment for 3, 10 and 20 nodes. Linear interpolation used for number of nodes between 3–10 and 10–20

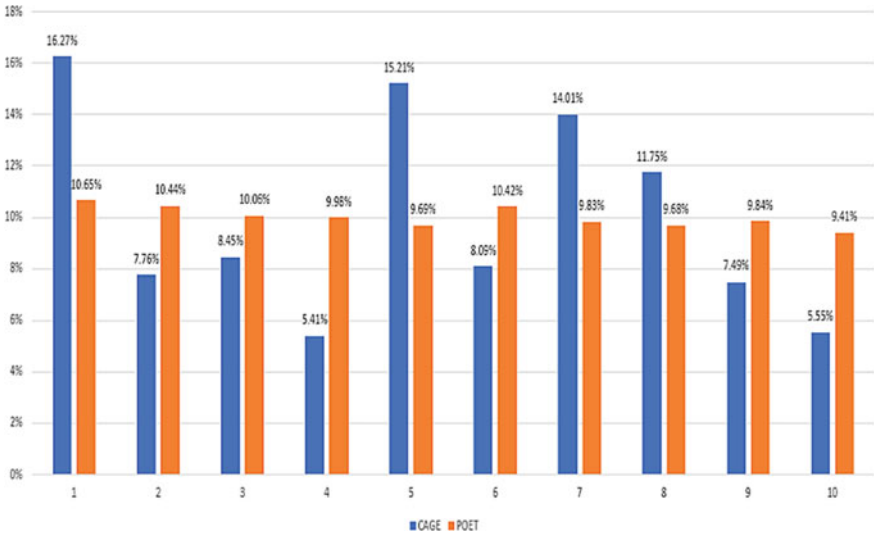


Fig. 4 Node distribution for 10 node BCT system using PoET or CAGE CAs. PoET has a fair and equal distribution. CAGE favours nodes 1, 5 and 7

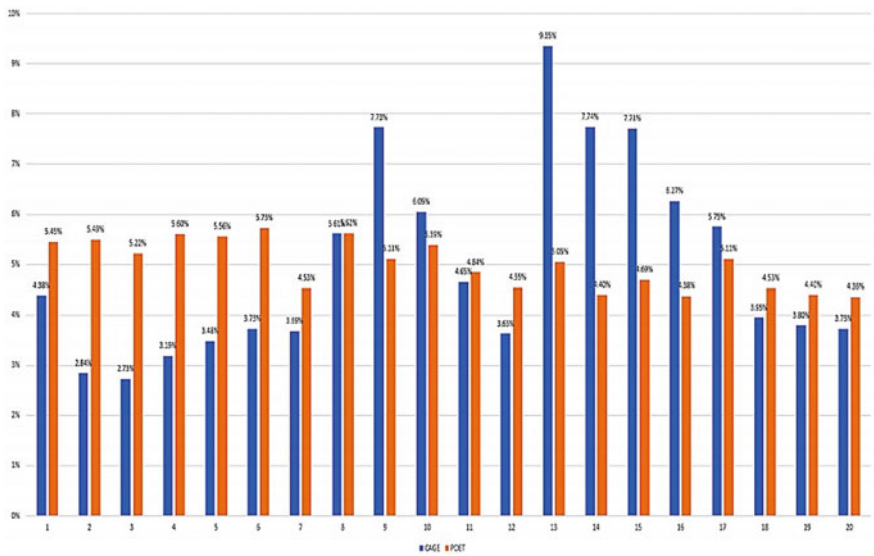


Fig. 5 Node distribution for 20 node BCT system using PoET or CAGE CAs. PoET has a fair and equal distribution. CAGE is slightly biased towards nodes 9, 13, 14 and 15. The variance for node distributions are 3.7 and 0.2 for CAGE and PoET, respectively

compared to approx. 6 blocks per second for POET. On average and overall CAGE performs twice as better in the creation of blocks per second.

6.3 Node Distribution

It would be devoid of decentralised systems existence to produce a node selection that is centralised and favours a single or a few nodes. In GAs this would be compared to over exploitation of a single parent, such that the individual makes copies of itself and dominates the gene pool. This generates a stagnant population that cannot find an optimal solution, without changing the selection algorithm or increasing the rate of mutation. The analogy in BCT would be to produce a centralised system reliant on a single node for all calculations which is something that is to be avoided.

The results in Figs. 4 and 5 indicate that whilst nodes are favoured with stronger fitness, there is still a selection of nodes with weaker fitness and over exploitation is neutralised.

The results in Fig. 5 illustrate that whilst CAGE is not as even as PoET, it does not stagnate or become centralised. Therefore, the idea of applying fitness and stamina attributes to node selection are at least successful and comparable to PoET.

It may concern the reader that there are favourable selections over time and this may contravene the decentralised process. There are some contradictions even in the most successful decentralised systems such as Bitcoin, in [25] it is reported that fewer than 16 nodes control over 90% of mining power. There are other considerations in such networks, such as incentivised CAs, connectivity, and resources, but this indicates that a biased decentralise network can still succeed. There must be caution when drawing such comparisons between FT and BFT BCT networks, however, it does indicate that node selection is an important to preventing a drift from decentralise to centralise networks.

The results in Fig. 5 illustrate that whilst CAGE is not as even as PoET, it does not stagnate or become centralised. Therefore, the idea of applying fitness and stamina attributes to node selection is demonstrated as successful and comparable to PoET.

7 Conclusion

When starting on this journey, it wasn't the idea that was the problem. It was the design, development, implementation and analysis. The results did reveal that straightforward implementation of selection algorithms, results in the uneven distribution of node selection. With hindsight this was to be expected, it has been observed in optimisation of GAs getting stuck in local minima due to the stagnation of the population. The population diversity required is often attributed to good and fair selection techniques, different cross-over techniques and mutation. This experiment was not exempt from stagnation and it was found that the distribution of node selection was

biased. To address this a new set of experiments were tried and a stamina attribute was added to each node. After completing this the node selection was distributed. So, the measurements of the experiments were: throughput, latency and node distribution. Any new consensus algorithm should consider these three metrics, with the latter being important in a decentralised system.

7.1 Future Work

Nature has many solutions to technological problems, here is no exception. The future work would be to investigate other areas of AI and GAs to create an algorithm that would be Byzantine fault tolerant and work on permissionless blockchain technology. This new algorithm is unlikely to challenge the latency and throughput capacity of existing algorithms. This would adapt to tortoise-like blockchains.

Some measurements could be improved upon during the experiment, these would be to record the lag in transactions due to node selection, block generation and confirmation are detailed in [26].

7.2 Discussion

Is low latency always and high throughput always desirable for a consensus algorithm? In blockchain technology, it seems that low latency and high throughput are attractive metrics of any consensus algorithm. Until now this is true, however, energy consumption has raised questions over the use of certain consensus algorithms. Therefore, what if latency and throughput came second to energy consumption? This would mean that the information in the transactions would not be required imminently. There are plenty of systems that do not require imminent transactions, e.g., annual audits [27, 28], exam papers [3], processes and procedures relating to audits and accreditation [29, 30]. Such events could take place on slow blockchains, where it is important to record the event, just not immediately. The advantage of tortoise-like blockchains would be their efficient use of energy. There would have to be an issue of spending and any expenses incurred by processing the transaction. Such actions could be to lock the customer until the transaction has been completed. Yes, this may be preventative in terms of speed, but in slow networks the problem being solve would have to be conducive to this.

For example, if there is a token for the accreditation of annual membership to an organisation. The applicant would apply for membership by spending the token. The transaction would not need to be immediate. The applicant would have their approval removed for spending tokens. However, they would not need to spend this until the following year. Such a problem would be conducive to a tortoise-like blockchain.

8 Summary

CAGE is a proposed consensus algorithm inspired by nature and is a new area of research. The selection of nodes is similar to the selection of individuals in genetic algorithms with an additional stamina attribute. CAGE was tested and compared to PoET on a permissioned network. The results show that with a low number of nodes PoET outperforms CAGE, but with higher numbers CAGE outperforms PoET. CAGE node selection distribution is slightly biased, but not centralised, whereas PoET has a more even distribution.

References

1. Maesa DDF, Mori P (2020) Blockchain 3.0 applications survey. *J Parallel Distrib Comput* 138:99–114
2. Azaria A, Ekblaw A, Vieira T, Lippman A (2016) Medrec: using blockchain for medical data access and permission management. In: *International conference on open and big data*. pp 25–30
3. Mitchell I, Sheriff M, Hara S (2019) DappER: decentralised application for exam reviews. *Global security, safety and sustainability. The Security Challenges of the Connected World*
4. Al-Jaroodi J, Mohamed N (2019) Blockchain in industries: a survey. *IEEE Access* 7:36500–36515
5. Lokre SS, Naman V, Priya S, Panda Sk (2021) Gun tracking system using blockchain technology. In: *Blockchain Technology: applications and challenges*. Springer, pp 285–300
6. Mitchell I, Hara S, Jahankhani H, Neilson D (2020) Blockchain of custody, BoC. In: *Cyber security practitioner's guide*. World Scientific, pp 365–397
7. Whitaker A, Bracegirdle A, de Menil S, Gitlitz MA, Saltos L (2021) Art, antiquities, and blockchain: new approaches to the restitution of cultural heritage. *Int J Cult Policy* 27(3):312–329
8. Wang Z, Yang L, Wang Q, Liu D, Xu Z, Liu S (2019) ArtChain: blockchain-enabled platform for art marketplace. In: *international conference on blockchain*. IEEE, pp 447–454
9. Novo O (2018) Blockchain meets IoT: an architecture for scalable access management in IoT. *IEEE Internet Things J* 5(2):1184–1195
10. Filippova E (2019) Empirical evidence and economic implications of blockchain as a general purpose technology. In: *2019 IEEE technology and engineering management conference (TEMSCON)*. pp 1–8
11. Gallersdörfer U, Klaaßen L, Stoll C (2020) Energy consumption of cryptocurrencies beyond bitcoin. *Joule* 4(9):1843–1846
12. Sutherland BR (2019) Blockchain's first consensus implementation is unsustainable. *Joule* 3(4):917–919
13. ISO26000 (2018) How to Contribute to sustainable development. International Organisation for Standardization (ISO). Geneva, CH
14. Sedlmeir J, Buhl HU, Fridgen G, Keller R (2020) The energy consumption of blockchain technology: beyond myth. *Bus Inf Syst Eng* 62(6):599–608
15. Hafid A, Hafid AS, Samih M (2020) Scaling blockchains: a comprehensive survey. *IEEE Access* 8:125244–125262
16. National Institute of Standards and Technology (2015) Secure Hash Standard. Tech. Rep. Federal Information Processing Standards Publications (FIPS PUBS) 180–4, U.S. Department of Commerce, Washington, DC
17. Holland JH (1992) Genetic algorithms. *Sci Am* 267(1):66–73

18. Shukla A, Pandey HM, Mehrotra D (2015) Comparative review of selection techniques in genetic algorithm. In: International conference on futuristic trends on computational analysis and knowledge management (ABLAZE). IEEE, pp 515–519
19. Thakkar P, Nathan S, Vishwanathan B (2018) Performance benchmarking and optimizing hyperledger fabric blockchain platform. arXiv preprint [arXiv:1805.11390](https://arxiv.org/abs/1805.11390)
20. Razali NM, Geraghty J, et al (2011) Genetic algorithm performance with different selection strategies in solving TSP. In: Proceedings of the world congress on engineering, vol 2. International Association of Engineers, Hong Kong, China, pp 1–6
21. Burkhardt M, Yampolskiy RV (2021) Death in genetic algorithms. CoRR abs/2109.13744, <https://arxiv.org/abs/2109.13744>
22. Dimopoulos C, Papageorgis P, Boustras G, Efstathiades C (2017) The concept of ageing in evolutionary algorithms: discussion and inspirations for human ageing. *Mech Ageing Dev* 163:8–14
23. Mingxiao D, Xiaofeng M, Zhe Z, Xiangwei W, Qijun C (2019) A review on consensus algorithm of blockchain. In: IEEE international conference on systems, man, and cybernetics (SMC). pp 2567–2572
24. Ferdous MS, Chowdhury MJM, Hoque MA, Colman A (2020) Blockchain consensus algorithms: a survey. arXiv preprint [arXiv:2001.07091](https://arxiv.org/abs/2001.07091)
25. Gencer AE, Basu S, Eyal I, Van Renesse R, Sirer EG (2019) Decentralization in bitcoin and ethereum networks. In: Financial cryptography and data security: 22nd international conference, FC 2018, Nieuwpoort, Curaçao, February 26–March 2, 2018, Revised Selected Papers 22. Springer, pp 439–457
26. Aslam T, Maqbool A, Akhtar M, Mirza A, Khan MA, Khan WZ (2022) Blockchain based enhanced ERP transaction integrity architecture and PoET consensus. *Comput Mater Contin* 70(1):1089–1109
27. Bonyuet D (2020) Overview and impact of blockchain on auditing. *Int J Digit Account Res* 20:31–43
28. Snow P, Deery B, Kirby P, Johnston D (2015) Factom ledger by consensus. Retrieved from Factom website: <http://www.factom.org>. Accessed June 2023
29. Cahyadi D, Faturahman A, Haryani H, Dolan E et al (2021) BCS: Blockchain smart curriculum system for verification student accreditation. *Int J Cyber IT Serv Manag* 1(1):65–83
30. Tariq A, Haq HB, Ali ST (2022) Cerberus: a blockchain-based accreditation and degree verification system. *IEEE Trans Comput Soc Syst*

Data Hiding in Anti-forensics—Exploit Delivery Through Digital Steganography



Hans Gashi, Shahrzad Zargari, and Setareh Jalali Ghazaani

Abstract Developments in digital forensics investigations have occurred along with those in anti-forensics. Legal issues involving cybercrime are difficult to investigate and even more difficult to prosecute since a forensic investigator must often develop a case by examining artefacts left on a device or network. When cyber criminals became more aware of the techniques utilized in digital forensics, countermeasures to these approaches were developed. The goal of these procedures is to sabotage forensic investigations, and many of them are readily available and simple to use. The purpose of this research is to improve our understanding of these Anti-Forensic technologies by doing in-depth individual analyses and discussing the functionality and methods, as well as the possibilities of mitigation. The topic of this Anti-Forensics study is within Data Hiding; there are different ways available; however, this project focuses on a steganography tool known as Stegosploit and looks to see if embedding JPG images with malicious code without visual distortion of the image is conceivable.

Keywords Anti-forensics · Digital forensics · Data hiding · Steganography · Browser exploit delivery

1 Introduction

In the digital era, with the rapid evolution of cybercrime, both in methodology and frequency of occurrence, digital forensics has become a crucial element of the judicial system. While digital forensics aims to legally acquire and examine collected evidence, Anti-Forensics aims to hinder that process by applying a variety of methods such as manipulating, hiding, or wiping data through respective tools. Cybercriminals may even aim to diminish the credibility of evidence as this is often enough to disrupt a forensic investigation. This work presents a data hiding method that allows for embedding images with undetected malicious code that can be later extracted.

H. Gashi · S. Zargari · S. Jalali Ghazaani (✉)
Sheffield Hallam University, Sheffield, UK
e-mail: Setareh.ghazaani@gmail.com

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_4

Hiding data in images is known as Steganography. The focus of this study is a steganography method called Stegosplit, this tool can deliver browser exploits while avoiding detection. Understanding the foundations of tools of this nature as well as the options available for mitigation is necessary in order to advance our understanding of Anti-Forensics.

Current research of individual Anti-Forensic tools is quite limited, this may be because efforts for countering AF are generally only made once necessary, for example, should a forensic investigator find evidence that an AF (Anti-Forensic) process has been used within a case, they may then proceed towards mitigation of that method. "The lack of true innovation in the forensic world is because there's no pressure to do so" [1]. This is not the case for Anti-Forensics, as constant innovation is a requisite in order to reliably hinder forensic investigations.

This study aims to further our understanding of AF techniques and how they may impact the investigative process. For every new digital forensic technique developed, cybercriminals look to implementing a related counter-technique [2]. This creates an endless cycle, with this constant evolution of Anti-Forensics, it became necessary to consider as a part of DFIR (Digital Forensics and Incident Response). In order to achieve better results when countering Anti-Forensics, performing individual analysis of these tools is a vital step in understanding the underlying mechanisms, this will aid in the efforts for defining and categorizing of Anti-Forensics which began in the early 2000s [3].

A data set of 308 known tools that can be considered as anti-forensic based on their characteristics was proposed by Rogers [4] as part of an anti-forensic taxonomy that defines and categorizes these tools. Conlan et al. [5] devised an extended AF taxonomy in the hopes of facilitating a form of standardization for AF. However, while research on Data hiding and Steganography is plentiful, very little research could be found that investigates Stegosplit or any other Drive-by (malicious content that is able to exploit vulnerabilities in a browser without the users knowledge) browser exploits. This study aims to fill this gap in research and explicitly detail the mechanisms of Stegosplit and undetected payload delivery in browsers. Therefore, in contribution to these efforts, the following research questions have been formulated.

How can digital steganography techniques be utilized to embed hidden data into the pixels of a JPG file without altering the image, and what strategies can be employed to mitigate and categorize Stegosplit within a standardized taxonomy of Anti-Forensics (AF) tools?

The aim of this research is to address the research questions by pursuing several objectives and deliverables. The objectives include evaluating current literature on browser exploits to enhance our understanding of Anti-Forensics (AF) tools through individual analysis. Additionally, conducting the Stegosplit experiment in a secure virtual environment with simulated parameters to assess the feasibility of delivering hidden data through JPG files as carriers. Furthermore, measuring the effectiveness of the Anti-Forensic Tool and documenting the process of hiding, preserving, and delivering information using Stegosplit. Lastly, exploring countermeasures against this AF method and evaluating their success in detecting suspicious content. By

accomplishing these objectives, this research aims to contribute to the field of tackling Anti-Forensics.

2 Literature Review

The following section highlights research material that was gathered to aid with the development of the study.

2.1 Introduction to Digital Forensics, Anti-forensics and Current Challenges

Digital Forensics in simple terms, is the extraction of any information that may be useful to a forensic investigation contained on an electronic device [6]. In 2001 the first Digital Forensic Research Workshop took place, this workshop intended to set the foundations of a wider community of specialists and academics for Digital Forensics. It aimed to set in motion works towards a clear definition of DF as well as furthering our understanding of the greatest challenges in the field. They would proceed to define Digital Forensic Science as.

“The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal or helping to anticipate unauthorized actions shown to be disruptive to planned operations.” [7].

Although AF is not a new or even recent concept, it lacks a universal standardized definition [8]. A myriad of studies would posit that the bedrock of AF is the process a cybercriminal observes when attempting to hide their electronic footprint so that it remains undetectable during a forensic investigation [9]. Whereas other studies such as [1] would suggest that AF is the act of avoiding detection and breaking tools or [10] who defined AF as hiding attempts at system intrusion.

The relationship between Digital Forensics & Anti-Forensics has grown to a stage where should we fail to find a clear research strategy upon which collective works can be developed with a clear goal, advancements in DF (Digital Forensics) will start trailing behind those in AF. This was addressed by Beebe [11] which argues that DF is no longer a niche discipline, and it is now common knowledge that any action taken on an electronic device or network leaves a digital footprint. Comparatively, other sources would state that despite the development of AF tools, many of them freely available, AF largely remains present on only the most complex DF investigations (POST 2016).

The “Golden Age” of DF is said to have occurred between 1990 and 2010. Vincze [12] suggests that throughout those years DF could be defined by several characteristics such as:

- Relatively few file formats.
- Investigations would generally be within a single device.
- Majority of computer systems operating on Microsoft Windows.

However, the capabilities of DF developed over the last decade are at risk of becoming obsolete in the future due to the transformation of computing technology. Garfinkel [13] when discussing challenges in DF in the next 10 years suggests that some of the contributing factors may be:

- Considerably larger storage on devices, increasing the time it takes when creating and processing a forensic image.
- Development of Embedded Storage, this makes it so storage on devices can no longer easily be removed and imaged.
- The propagation of different file formats and OS has greatly increased complexity when dealing with data processing tools.
- The scope of forensic investigations is largely defined by regulations and other legal challenges.

2.2 Anti-forensics—Data Hiding

The practice of storing information in locations where it is unlikely to be found is known as Data Hiding. Blunden [14] defines data hiding as “a method of security through obscurity”, this refers to act of integrating an obscure piece of data into a host source whilst maintaining the integrity of said source, this would operate through a different communication scheme when compared to how normal data transfer and communication occurs, meaning that hidden data relies on the host source for transmission [15]. The difference between normal data communication and communication of hidden data is that hidden data requires a host signal in which a secret message is embedded and then must be extracted upon transmission, whereas traditional data communication simply sends and receives message signals through a transmission channel.

There is a multitude of methods available to hide data, some of the more traditional methods include tampering with the metadata of certain file extensions, hiding data in unallocated or slack space, use of MBR (Master Boot Record), device driver registers and tables which may be allocated but are not in use or even partitions on a device hard drive that are potentially hidden and encrypted, however the vast majority of these methods can be discovered by forensic tools [16].

In contrast to this, there is still methods of data hiding that make use of unreachable locations such as unallocated space that are ignored by available forensic tools, a compiled list of such methods is shown in Table 1.

Table 1 Data hiding—available file system tools

Data hiding tool	Functionality	Source
Metasploit's slacker	Hides data within slack space of NTFS or FAT file systems	Göbel and Baier [17]
FragFS	Hides data in the NTFS MFT (Master file Table)	[3]
RuneFS	Stores data in “bad blocks”	Grugq [18]
Waffen FS	Hides data within ext3 journal file	Eckstein and Jahnke [19]
KY FS	Stores data in directories	Grugq [18]
Data mule FS	Hides data in inode reserved space	Grugq [18]

2.3 Steganography Techniques

An example of one of the earliest effective methods of data hiding is Steganography. It can be summarized as the act of hiding secret information inside a carrier file [20]. These carrier files can consist of a variety of different file types, including, audio or image files, certain video file types and executable files [21].

As the scope of this study is steganography in images and whether it's possible to encode and deliver exploits without affecting the image itself, a review was conducted examining research on the effectiveness of known methods of image steganography. The most popular survey of steganography techniques is that of Johnson [22], while the majority of the techniques discussed in that survey are still relevant today, we must factor in that this work was published over 20 years ago. We can reasonably assume that in that time digital steganography has developed drastically and new methods have become available. This was addressed by Hamid [23] who would devise a more current taxonomy of image steganography methods. In comparison, whilst Johnson [22] discussed the ability of the reviewed techniques to hide information, Hamid [23] took a more in-depth approach by also accounting for how much information could be hidden and the robustness of these image processing tools. Referring to this papers research gap, regardless of in-depth surveys such as these, little research is available that individually analyzes and investigates available steganography tools beyond the most popular options.

When discussing the effectiveness of image steganography methods, [24] states that each technique can be weighed against the following characteristics:

- **Robustness**—This is the ability of an image to remain intact should it go through a process that changes the image itself such as resizing or cropping after embedding.
- **High-Capacity Perceptual Transparency**—Refers to the amount of degradation a carrier image undergoes after the maximum amount of information has been embedded into the image, meaning the perceived quality of the embedded image when compared to the original.
- **Temper Resistance**—This is how difficult it would be to change the information after embedding into an image.

- **Computation Complexity**—Measures how expensive or resource intensive the process of embedding and extracting a hidden message is.

Though there are many methods of image steganography such as Marking & Filtering, Distortion Techniques and Domain Transformation Techniques, the work performed on this paper looks to make use of hiding data by changing the least significant bit (LSB) values within the image pixels. This technique is known as Spatial Domain. During his work surveying image steganography techniques [25] would broadly classify Spatial Domain Techniques into the following categories:

- Random Pixel Embedding (RPE)
- Pixel Value Differencing (PVD)
- Least Significant bit (LSB)
- Histogram Shifting method
- Pixel intensity-based methods.

The developments made in the field of AF in recent years through constant updates and release of new methods have created an ever-growing gap between those in Digital Forensics. The major gap identified in the research stage of this study is the lack of in-depth individual analysis of current AF tools, beyond proposed taxonomies and surveys that aim to categorize AF tools in the hope of aiding our understanding, minimal research is available that sufficiently inquires about the fundamental mechanisms of these tools or that discusses specific mitigation methods. This gap further widens when discussing methods developed in recent years. Therefore, this study will focus on the examination of a Steganography tool known as Stegosplit of which no current research papers could be found, in addition to this, this paper also looks to investigate the possibility of mitigation of this tool in the hopes of aiding future forensic investigations being confronted by this AF method.

Throughout this literature review, much new information relating to the broad spectrum of data hiding as well as steganography itself was gained, it was interesting to delve into the current stenographic techniques and discover the true depth of this field. The sheer number of techniques available only highlights the problem digital forensics faces. Another interesting aspect was the consideration that AF should not be considered as a tool used exclusively by the criminal class and that it does in fact have legitimate uses by those seeking to protect their own privacy. Whether it is beneficial to excuse the use of some tools despite them being AF in nature is yet to be seen. For example, one might ask what possible use an average user could find in Digital Steganography except for acting on malicious intent.

Furthermore, it was interesting to find that despite AF being an issue shared by the entire digital forensic community, there is still many differences of opinions when attempting to define AF itself, its objectives and whether its existence serves only to hinder forensic investigations.

3 Results and Discussion

Any digital image in essence is simply a correlation of pixels that have been arranged so that the desired output is displayed. Every individual pixel is characterized by 3 color channels known as RGB (Red, Green, Blue), where each of these channels is represented by its own 8-bit value, which in turn allows for a total of 256 possible levels of color. For example, if we were to analyze an image with only a black and white or greyscale color scheme, we would find that the values for RGB would be the same throughout each pixel. In order to aid with the visualization of this, consider an image that has been sliced into 8 separate panes where each individual pane is its own bit layer. The first bit layer (Bit layer 0) naturally has the least impact on the overall image and is therefore known as the images LSB (Least-Significant Bit). Thus pixels in bit layer 1 are composed of values of the second least-significant bit. This in turn would make bit layer 7 the most important layer as it has the most effect on the output of the image itself, for this reason, it is called the Most-Significant Bit (MSB) [26].

Accordingly, encoding an exploit onto the higher bit layers of an image (Layers 3,4,5,6 and 7) will result distortion of the image, significant enough that it will be noticeable to the human eye. Therefore, encoding the exploit into the lower bit layers should show minimal or possibly no aberration of the image, even to close inspection.

When replacing values of the LSB with the with the code representing the hidden data, in our case a JavaScript exploit, individual binary bits of the image pixels are changed to fit the encoded hidden data so that the visual aberration of the image is negligible to the human eye. If we were to select a specific pixel color channel bit stream from the original image such as:

```
10010101 00001101 11001001
10010110 00001111 11001010
10011111 00010000 11001011
```

Taking the following random 8 bits from the converted hidden data embedded as an example,

```
101101101
```

Hiding those 9 bits from a segment of our hidden data in a selected pixel is simply a matter of changing the last bit in each sequence of 8 bits from the original image pixel bit stream as needed. Resulting in:

```
10010101 00001100 11001001
10010111 00001110 11001011
10011111 00010000 11001011
```

By using the method this example shows that it is only necessary to change the four out of the nine LSB in those nine bytes [27]. Changing the last bit results in a slightly different colored pixel but remains indistinguishable from the original when viewing the image.

However, regardless of bit layers, pixel loss is always a possibility when altering the bits of an image. To overcome this, an iterative encoding technique is used which allowed for error-free decoding of the exploit code.

The study specification set out at the beginning of this work was to determine “Can digital steganography be used to embed hidden data into pixels of a JPG file without affecting the image?”. This research question remained the same throughout this study as the existing literature review conducted showed that there currently a gap in research when discussing individual analysis of AF tools that sufficiently evaluates their mechanisms in order to better understand how they can affect DF investigations. No research could be found that relates to Stegosploit and the data hiding method that it employs.

To ensure that this research question is answered, and the stated research gap addressed, the experiment conducted in this study provides a detailed overview of the process that Stegosploit uses to embed an image with hidden data, discussing the intricate mechanisms of the tool such as the process it undertakes when encoding and decoding hidden data in a JPG image via conversion of the data into a bit stream that is then sporadically hidden within its pixels in the form of a grid. In order to adequately analyze this AF tool so that future DF investigations concerning this steganography technique are able to efficiently recognize when these actions have been appropriated, in addition to detailing the process that this AF tool undertakes, the experiment also demonstrates how a cybercriminal may then use a stenographically encoded image to initiate an attack. For this purpose, the embedded image is used as a part of a browser exploit delivery method where the embedded JavaScript exploit is executed once the web page is visited, or the image is clicked.

From the results gained in the experiment it was found that the embedded exploit code becomes much more visually apparent on pixel layers 3, 4, 5, 6 and 7, therefore it is safe to assume should an attacker employ this tool, they would naturally hide their data in the lower layers of an image as to make it impossible to notice only through visual inspection of the image.

Addressing the research question of this study and the initial hypothesis: Encoding JPG images with hidden data is possible with no noticeable visual distortion of the image.

The results show that the hypothesis was correct when stating that it is possible to encode JPG images with hidden data with no visual aberration of the image provided that the data being hidden is encoded onto Bit-Layers 0–2, as using layers above those has a more prominent effect on the overall image, so much so that even a visual inspection of the image in passing is enough to detect that the image has been tampered with defeating the primary purpose of this stenographic technique.

Stegosploit is a relatively new AF tool, only a handful of researchers have ever conducted works involving it. While there is plentiful research available when discussing the methods available for steganalysis and detecting stenographically altered content, techniques that look to counter images generated by Stegosploit are practically non-existent. In relation, Park [28] published works showing possible methods available to protect networks from hidden exploits and [29] discussing their analysis of Stegosploit images or Harblson [30] who analyses techniques for hacking

with images through Stegosplit. However, these authors all discuss in detail various characteristics of Stegosplit, there is no focus on the means of mitigating of this tool.

As Stegosplit hides the malicious data within an image acting as a carrier file, it can easily evade even modern anti-virus software as it is treated as a regular JPG file that does not raise any flags, this is due to the malicious code embedded into the image not being directly visible.

The main focus of this study being an in-depth analysis of the functionalities of Stegosplit, showing its mechanisms as well as how it may be used maliciously, so that future forensic investigations are able to efficiently respond. It is beyond the scope of this study to establish how direct mitigation of Stegosplit can be implemented. Regardless of this, some research was conducted in order to allow future researchers an avenue in which they may develop a fixed method of mitigation.

Countering and mitigation of Stegosplit may prove more successful if the focus is shifted from attempting to directly detect encoded images to the detection of polyglots as this is the primary route Stegosplit takes for browser exploit delivery. An example of this is the work conducted by Vaidya and Rughani [31], who presented a method of polyglot detection by analyzing signatures of both a regular JPG and one that has been stenographically encoded through HxD (Hex Editor) and designing a python script that detects polyglots based on this observation.

The initial goals of this study were to contribute to the immense work that is required to create a standardized definition and understanding of Anti-Forensics so that we may allow for purposeful development of Digital-Forensics tools and techniques. This is outlined in the work of many authors who looked to create a grand taxonomy of Anti-Forensic tools so that a systematic analysis of these methods could be performed. It is only by inquiring about individual tools available through collective works that are frequently adapted that we may hope to truly bring to light the scale of the threat that Anti-Forensics poses to forensic investigations.

The experiment performed in this study successfully detailed the processes that the Anti-Forensic tool Stegosplit undertakes when stenographically hiding data in the pixel layers of JPG file. An in-depth analysis was performed of the encoding and decoding mechanisms of the tool as well as how it is able to deliver malicious code without user knowledge via polyglots. This meticulous work along with the options available for mitigation to the best of the researcher's ability, allows for this tool to be added to future granular taxonomies of Anti-Forensic tools hoping to further digital forensic science or already available works such as that of [32] extended granular taxonomy of Anti-Forensic tools or Katamara [33] taxonomy of countermeasures to Anti-Forensic tools.

3.1 Limitation Od Study and Future Work

Throughout this study, a number of unanticipated obstacles were met. The most apparent was the researchers lack of experience with the JavaScript programming

language which affected the ability to use JavaScript exploits when embedding the image. Some difficulty was found when attempting to appropriate the syntax to fit Stegosploit. Due to this, a simple JavaScript alert was used in the place of an exploit to show the execution of the code upon delivery. Future researchers more familiar with JavaScript should have no issues when implementing exploits into the tool following the methodology shown in this paper. Another limitation of the study is the tools lack of support for other file types such as PNG or GIF files, Stegosploit is unable to encode and decode JavaScript exploits on any other file type excluding JPG due to the differences in file infrastructure. Lastly, there are some browsers available that do not support JavaScript and therefore inhibit the process that Stegosploit takes, making it impossible to execute exploits in the manner in which the tool is designed.

Future research in the field of Anti-Forensics and Steganography tools should look towards establishing a direct mitigation and detection method of malicious images as well as conceptualizing an expanded steganography subcategory for tools akin to Stegosploit in order to aid with categorization in future Anti-Forensic tool taxonomies. Regarding Stegosploit, future directions for research could include discovering a method that allows multiple browser compatibility for steganographically embedded JPG images as well as how these images can withstand resizing and compression without data loss.

4 Conclusion

To conclude this study, the initial aim of this research was to aid in the understanding of Anti-Forensic techniques through in-depth individual analysis so we may then be able to better recognize how they affect forensic investigations. From the literature review conducted, it was noted that the availability of research that looks to examine the intricate functionalities of specific tools was lacking. This was even more apparent when delving into Anti-Forensics through Steganography. Minimal Research was available that thoroughly discussed the relevant parameters such as the purpose, functionality and mitigation. Many authors discussing current Anti-Forensic tools would postulate that in order to close the gap between developments made in Anti-Forensics to those in Digital forensics, a comprehensive study of specific tools is necessary. Therefore, this study attempts to fill this gap in research having thoroughly analyzed the steganography tool Stegosploit, meeting all the relevant parameters mentioned above so that our understanding of this tool is adequate enough that it is ready to be added to taxonomies of Anti-Forensic methods easing the transition into real world application should a digital forensic investigator be met by this data hiding tool.

When referring to the research question proposed at the beginning of this study as well as the hypothesis, it can be said that they have been answered successfully and the objectives met. The initial hypothesis stating: "Encoding JPG images with hidden data is possible with no noticeable visual distortion of the image", upon experimenting the results showed that the hypothesis was correct in that it is possible

to hide data within a JPG file without distortion noticeable to the human eye. However, it was found that this only remains true should the data be hidden in the lower bit-layers of the image (LSB) as embedded data into the higher layers or the MSB causes significant aberration of the image and would therefore defeat the purpose of the tool by making it considerably easier to discover that the image has been tampered with.

This research also looked further to inquire the mitigation methods available for Stegosplit as well as how it may be categorized in taxonomies looking to standardize Anti-Forensic tools and methodologies. Upon research, it was found that at the time of this study, options targeting direct mitigation of this tool are currently widely unavailable, this may be due to the fact that it is a relatively new Anti-Forensic method and remains a somewhat niche technique for cyber criminals to implement. However, it is safe to assume that this will change in the future and that options for mitigation need to be considered. The possibility of implementing a method that looks to shift the focus from detecting malicious images to the detection of polyglots for Stegosplit is mentioned as upon analysis of the tool, this mitigation method is the most appropriate. When looking towards categorization, Stegosplit falls under Steganography related Anti-Forensic tools, however, based on the work conducted on this report, creation of further subcategories for Steganography tools should be considered. This is due to the nature in which Stegosplit performs its stenographic process by altering specific pixels and the means by which it exploits vulnerabilities.

References

1. Foster MA, Liu V (2005) Catch Me If You Can. Congressional Research Service
2. Raghavan S (2013) Digital forensic research: current state of the art. CSIT 1:91–114
3. Garfinkel S (2007) Anti-forensics: techniques, detection and countermeasure. In: 2nd international conference on i-warfare and security, vol 20087. pp 77–84
4. Rogers M (2006) Anti-forensics: the coming wave in digital forensics
5. Conlan K, Baggili I, Breitingner F (2016) Anti-forensics: furthering digital forensic science through a new extended, granular taxonomy. Digit Investig. 18. <https://doi.org/10.1016/j.diin.2016.04.006>
6. Etow TR (2020) Impact of anti-forensics techniques on digital forensics investigation. <http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-97116>
7. DFRWS Technical Committee (DFRWS) (2001) A road map for digital forensic research: DFRWS technical report. DTR-T001-01
8. Harris R (2006) Arriving at an anti-forensics consensus: examining how to define and control the antiforensics problem. In: Proceedings of the 2006 digital forensics research workshop. Digit Investig 3(S):S44–S49. <http://dfrws.org/2006/proceedings/6-Harris.pdf>
9. Mothukur A, Balla A, Taylor D, Sirimalla S, Elleithy K (2019). Investigation of countermeasures to anti-forensic methods. 1–6. <https://doi.org/10.1109/LISAT.2019.8816826>.
10. Shirani B (2002) Anti-forensics. High Technology Crime Investigation Association. <http://www.aversion.net/presentations/HTCIA-02/anti-forensics.ppt>
11. Beebe N (2009) Digital forensic research: the good, the bad and the unaddressed. In IFIP International conference on digital forensics. Springer, Berlin, Heidelberg, pp 17–36
12. Vincze EA (2016) Challenges in digital forensics. Police Pract Res 17(2):183–194
13. Garfinkel SL (2010) Digital forensics research: the next 10 years. Digit Investig 7:S64–S73

14. Blunden B (2009) *The rootkit arsenal escape and evasion is the dark corners of the system*. Wordware Publishing
15. Sencar HT, Memon N (2008) Overview of state-of-the-art in digital image forensics. In: *Statistical science and interdisciplinary research: algorithms, architectures and information systems security*, pp 325–347. https://doi.org/10.1142/9789812836243_0015
16. Budimir N, Slay J (2007) Identifying non-volatile data storage areas: unique notebook identification information as digital evidence. *J Digit Forensics, Secur Law* 2(1):75–91
17. Göbel T, Baier H (2018) Anti-forensics in ext4: on secrecy and usability of timestamp-based data hiding. *Digital Invest* 24(Supplement), S111–S120. <https://doi.org/10.1016/j.diin.2018.01.014>
18. Grugq (2003) *To the art of defiling*. Black Hat Asia 2003 presentation. [online] <http://openres.thebunker.net/pub/mirrors/blackhat/presentations/bh-asia-03/bh-asia-03-grugq/bh-asia03-grugq.pdf>
19. Eckstein K, Jahnke M (2005) Data hiding in journaling file systems. In: *The digital forensic research conference DFRWS 2005 USA Proceedings*
20. Abboud G, Marean J, Yampolskiy RV (2010) Steganography and visual cryptography in computer forensics. In: *Fifth IEEE international workshop on systematic approaches to digital forensic engineering*. <https://doi.org/10.1109/sadfe.2010.14>
21. StegoArchive.com (2005) Stego Archive Web site. <http://www.stegoarchive.com>
22. Johnson NF, Katzenbeisser S (2000) A survey of steganographic techniques. In: *Information hiding*. 43–78
23. Hamid N, Yahya A, Ahmad RB, Al-Qershi OM (2012) Image steganography techniques: an overview. *Int J Comput Sci Secur (IJCSS)* 6(3):168–187
24. Lin ET, Delp EJ (1999) A review of data hiding in digital images. In: *IS&T's 1999 PICS conference proceedings*. Video and Image Processing Laboratory (VIPER), School of Electrical and Computer Engineering
25. Hussain M, Chen D, Cheng A, Wei H, Stanley D (2013) Change detection from remotely sensed images: from pixel-based to object-based approaches. *ISPRS J Photogram Remote Sens* 80:91–106. <https://doi.org/10.1016/j.isprsjprs.2013.03.006>
26. Rustad S, Setiadi DRIM, Syukur A, Andono PN (2022) Inverted LSB image steganography using adaptive pattern to improve imperceptibility. *J King Saud Univ—Comput Inf Sci* 34(6, Part B):3559–3568. <https://doi.org/10.1016/j.jksuci.2020.12.017>
27. Warkentin M, Bekkering E, Schmidt MB (2008) *Steganography: forensic, security, and legal issues*. The Association of Digital Forensics, Security and Law (ADFSL)
28. Park B, Kim D, Shin D (2015) A study on a method protecting a secure network against a hidden malicious code in the image. *Indian J Sci Technol* 8(26)
29. Jeyasekar JJ, Saravanan P (2016) Science maps of global and Indian wildlife forensics: a comparative analysis. *Libr Philos Pract* 2016:519. ISSN 1522-0222
30. Harblson C (2015) *Hacking with pictures; new stegosplit tool hides malware inside internet images for instant drive-by pwning*
31. Vaidya N, Rughani P (2019) An efficient technique to detect stegosplit generated images on windows and Linux subsystem on windows. *Int J Comput Sci Eng* 7(12):21–26. <https://doi.org/10.26438/ijcse/v7i12.2126>
32. Conlan K, Baggili I, Breitinger F (2016) Anti-forensics: furthering digital forensic science through a new extended, granular taxonomy. *Digital Invest* 18(Supplement):S66–S75. <https://doi.org/10.1016/j.diin.2016.04.006>
33. Katamara Z (2020) *Taxonomy for anti-forensics techniques and countermeasures*. *Culminating Studys in Information Assurance*. 109. https://repository.stcloudstate.edu/msia_etds/109

Impact Versus Frequency on Cybersecurity Breach Trends in the Business and Medical Industry to Identify Human Error



Galathara Kahanda, Sasha Rider, and Sayantini Mukhopadhyay

Abstract As life becomes more digitally oriented, we need to emphasize cybersecurity and its awareness. Particularly in industries such as medicine and business that contain sensitive information, like social security numbers or payment information, which if stolen can result in identity theft. With the right security measures in place, we can maintain a new standard of data privacy as the response to data breaches becomes more efficient. This paper aims to analyze two datasets to compare and recognize any ongoing data breach trends from 2005 to 2019 and from 2009 to 2022. After identifying these trends, we will determine whether these cybersecurity threats are mostly due to human error or if the organizations targeted most have weaker cybersecurity systems. Through our analysis, we found that the business sector loses thirty-nine times more records than medical, however medical is targeted significantly more. Whereas in business the top type of attack is usually hacking or malware, most medical breaches are either unintended disclosures or physical loss which is why fewer records are compromised. With deliberation, we conclude that human error is an element in the majority of large-scale breaches, as well as pointing out the necessity for stronger cybersecurity systems and teams.

Keywords Data breach · Unintended disclosure · Cybersecurity awareness · HIPAA · Protected health information · Yahoo data breaches · DDoS attack · Anthem hack · River City Media · Zero trust architecture

1 Introduction

The research paper begins with formulating a null and alternative hypothesis. Our null hypothesis is that the medical industry has weaker cybersecurity systems, thus leading to an increase in data breaches. The alternative hypothesis is that human error is one of the top reasons for data loss incidents. Through our analysis of the

G. Kahanda (✉) · S. Rider · S. Mukhopadhyay
Rutgers University, Newark, NJ, USA
e-mail: gk451@newark.rutgers.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_5

Privacy Rights Dataset, we have found that business is hit hardest in terms of records lost. So much so, that when combining all other industries' total number of records lost, they do not account for even half of the almost 10 billion lost in business from 2005 to 2019. Finding data on cybersecurity in the business sector is challenging as most of it is private, and even when found is costly. A meeting with Advisen [1], a data collection company that has a large real-time dataset on cybersecurity threats, revealed that their dataset cost between \$15,000 and \$17,000 for a yearly license for universities. Since obtaining data on business cybersecurity is scarce and expensive, in the second dataset we will focus on the industry that gets attacked most frequently and pays the most for data breach violations: "The Medical Industry". This second dataset was found on the U.S. Department of Health and Human Services Office for Civil Rights website and provides free archived information about data breach investigations in the medical industry. Breaches in the medical industry are so costly because PHI or private health information contains the most sensitive data, so for each record lost the medical facility must report and pay a HIPAA fine to the US Department of Health and Human Services. These fines can range from \$100 to \$50,000 per record lost [2] depending on severity and response time, so great precaution to preserve this data must be exhibited. Whether it is establishing strong cybersecurity systems or minimizing the possibility of human error, it is vital to understand the cause of the data breach trends to prevent them.

1.1 Synopsis on Cybersecurity Breaches

Digitization, globalization and smart technologies have elevated the chances as well as the adversities of cybercrime. As stated by [3], the occurrences of cybercrime have been predicted to have impacted the worldwide economy just beneath USD 1 trillion in the year 2020, depicting an increment of over 50% from 2018. Based on the same, the repercussions of cybersecurity impose substantiated corporate risks with business interruption, privacy breach and loss of finances. On the other hand, [4] emphasized that the exposure of privacy-sensitive information in consideration to security breach incidents could lead into serious losses towards both organizations and businesses. Hence, in consideration to the same, the exposure of sensitive information owing to unauthorized cyber access might act detrimental to the long-term stability of a business.

1.2 Human Factors Relative to the Occurrences of Cybersecurity Breach

Insider threat in assumption to the participation of cybersecurity transgressions is accompanied by malicious intent for either personal or financial gains. In retrospect,

[5] indicated that human driven breaches are aligned with carelessness, apathy and negligence. Thereby, associates of businesses might partake within adverse cybersecurity behavior entailing higher tolerance of risks for being efficient. Additionally, phishing and ransomware also pertain to cyber-security risk originating within individual involvement with either intentional or unintentional coherence. Diversely, based on the presentation of [6], it has been observed that the usage of smart devices acts as a primary source of privacy breaches due to the presence of software vulnerabilities and human error posing unauthorized access to the databases. Hence, the magnitude of the information breaches to an extent is confined with the probabilities of exposed records due to human error.

1.3 Cyber-Security Concern in Health Records

The advancement in the information and communication technology has been supportive to the healthcare sector based on the replacement of paper-driven systems with electronic health records (EHRs). Considering the equivalent, EHRs act facilitative to the enhancement of patient care with streamlining the disease diagnosis catering efficiency in the patient care [6]. However, the incorporation of a digitized health system, pertaining to convenient access might also align with security failures owing to the likelihood of unauthorized access to the databases. In reference, [7] indicated that the healthcare sector is on a continuum of being victimized to higher rate of data breaches. The reason behind the fact encompasses the targeting of innovative medical equipment and healthcare applications by the hackers. Relating to the same, businesses on a regular interval are being compelled to pay ransom in relation to the time-sensitiveness of the healthcare sector.

2 Methodology

2.1 Research Design

Comparative study design has been utilized within the premises of this research in order to analyze and evaluate databases on the frequency and trends of cybersecurity breaches within different time frames. In consideration to the same, the databases from 2005 to 2019 and 2009 to 2022, in order to interpret the intensity and drift of cybersecurity breaches based on business and medical industry prospects. The comparisons have helped in the derivation of distinguished insights on the repercussions of cybersecurity violations that have taken place within the considered time frame in order to determine the future propensity of the same.

2.2 Data Collection

The collection of information has been through existing databases in compliance with the key variables of the research. Key concepts in the research include unintended disclosure, data breach, malware, medical, business, human error and others. Complying the same, the databases that have been sorted are of PRC organizations, state-based data, government reports, organizational sources and medical database configuration information on the cyber breaches that have taken place within the time-frame of 2005–2019 and 2009–2022. This has enabled the analysis of the impact of cybersecurity violations on the medical and business scenario resembling leakage of sensitive data intervened by humans.

2.3 Data Analysis

Frequency and trend analysis have been undertaken through the basis of extracting information from the existing sources. The frequency analysis has facilitated in undermining the intensity of data breaches commenced due to human interpretation. On the other hand, trend analysis has been significant in terms of outlining the influence of cybersecurity breaches on the data management of the businesses and medical records. Proceedings of the frequency analysis, entailed counting of events of data breaches at diverse consideration within the period range of 2005–2019 and 2009–2022. Additionally, trend analysis has been conferred on the basis of understanding the pattern of adversities through databases based on similar time frames. These analytical techniques have acted prolific in analyzing the outcome of data breaches.

3 Datasets

Our first dataset was found in [8] research paper where they combined three datasets of cybersecurity breaches in an attempt to create a larger dataset to analyze. The one we chose to start our research comes from PRC or Privacy Rights Clearinghouse, which is a U.S. nonprofit organization that has an assortment of data breaches from 2005 to 2019. This dataset contains 9000 rows and 16 columns such as the type of breach and type of organization with abbreviations that are explained below (Figs. 1 and 2).

Now it is important to note that hacking and malware are not the same things, hence the combined classification in the first dataset is ambiguous. Whereas a hacker is actively working to disable security systems, malware is passive software usually sent over the internet that works to gain unauthorized access to a network server or computer. Furthermore, the second dataset is from the U.S. Department of Health and Human Services and includes 5000 rows with 9 columns, some of which are

Fig. 1 PRC organizations explained

Types of organizations	
Abbreviations	Meaning
BSF	Business financial/insurance
BSR	Business retail
MED	Healthcare, medicine
BSO	Business other
GOV	Government/military
NGO	Nonprof organization
EDU	Education institutions
UNKN	Unknown

Fig. 2 PRC breach types explained

Type of breach definitions	
Abbreviations	Meaning
HACK	Hacking/malware
PORT	Portable device
DISC	Unintended disclosure
PHYS	Physical loss
UNKN	Unknown
INSD	Insider
STAT	Stationary device
CARD	Payment card fraud

the type of breach, and the number of individuals affected. This spans from 2009 to 2022 and gives us a significant number of years to compare with the PRC dataset. By contrasting similar columns in both of the datasets, we will gain an in-depth understanding of data breach trends from 2005 to 2022. In the analysis section below, we will discuss how data breaches are affecting industries, primarily business and medical, and what some of the consequential effects are.

4 Evaluation/Analysis

4.1 Malicious Versus Accidental Attacks

Cybersecurity incidents are serious and affect millions of people. However, there are levels of severity when it comes to types of breaches. On one hand, hacks are malicious and always intend to cause harm, whereas unintended disclosure is usually an accidental miscommunication of sensitive information. Often, a disclosure is revealed to someone who does not intend to misuse the information. For example, a doctor sending medical records that get faxed to the wrong location is much different from an unintended disclosure where a database is left unencrypted with medical records easily accessible. Lack of protocol and proper training for employees are the primary reasons for a mistake like this, so essentially human error. Moreover, hacks tend to result in the highest number of record losses, so the worst kind of severity

in the dataset is large medical hacks. This is due to the sensitive nature of health data, so in this case, it is more fatal than a business hack. For example, private health information almost always contains social security numbers as well as treatment plan information, so tampering with this could result in not only identity theft but harm to the patient.

To show the impact that different types of breaches have, we grouped the data based on the type of breach column in both datasets, then summed each type based on the number of records lost or individuals affected. We then plotted this into two pie charts to show the drastic effect that hacking and malware have.

In Fig. 3 we can see that unintended disclosure accounts for 25.3% of all records lost in our first dataset. Whereas hacking takes the majority of total records lost with 69.4%, meaning that hacking and malware is the biggest threat when it comes to cybersecurity. Similarly, if we take a look at Fig. 4 of the second dataset, we can see that hacking/IT Incident leads to 82% of total individuals affected. Interestingly, unauthorized access/disclosure only has a 6% impact as compared to theft at 7%. So, although things like improper disposal and unintended disclosure may happen, they are very minimal threats in comparison to hacking or malware. From this, we know that we must provide recommendations to mitigate this issue.

Even in 2022, IBM’s cost of a data breach report states that ransomware grew 41% in the last year, took 49 days longer to identify and contain, and increased in

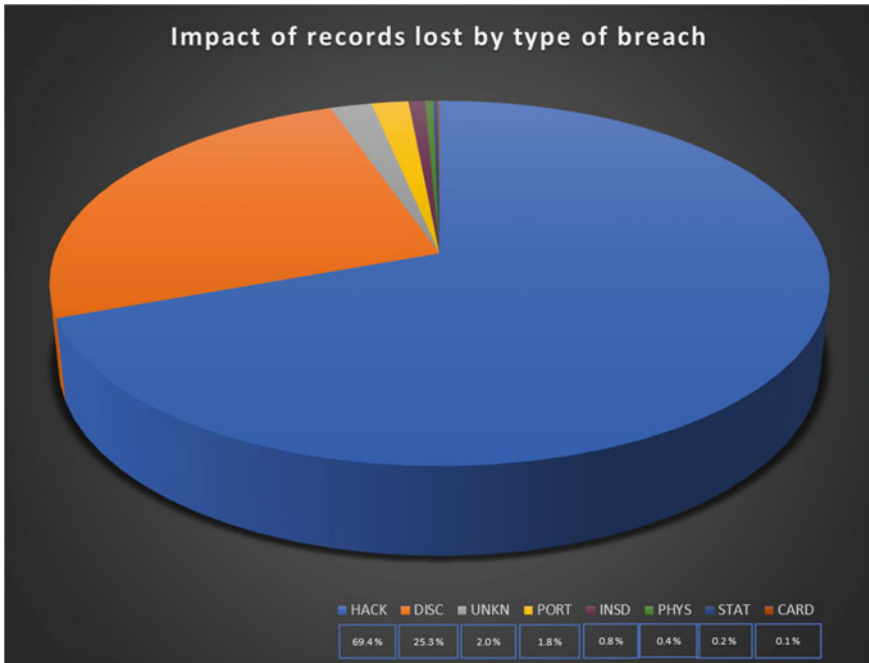


Fig. 3 Sum on total records lost by type of breach

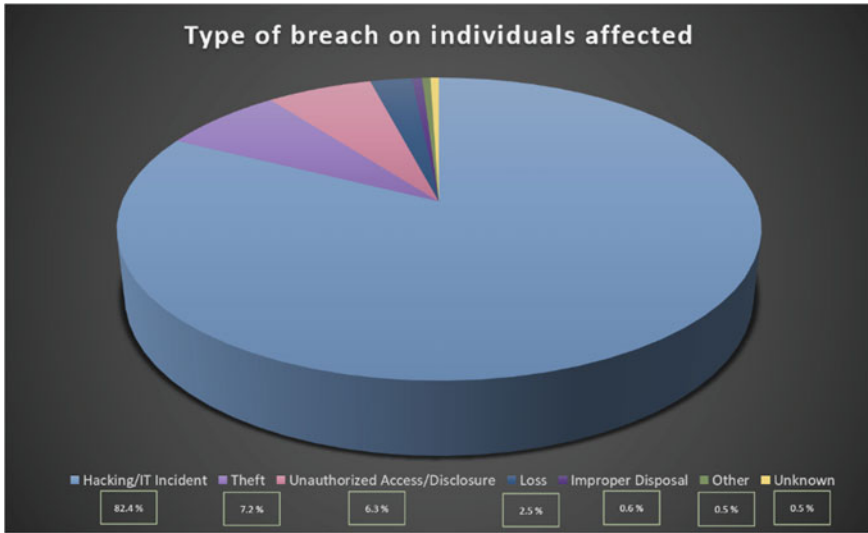


Fig. 4 Sum on total individuals affected by type of breach

cost [9]. This means that hacking and malware are steadily increasing and the issue is only getting worse. Thus, it is crucial to inform people on the best cybersecurity measures to practice as well as research techniques like artificial intelligence which can be implemented to decrease these attacks.

4.2 Impact Versus Frequency

Now we must shift our focus to understanding impact versus frequency. We briefly touched on impact in the last section, so here we will compare impact or total record loss/individuals affected to frequency or number of occurrences throughout different columns.

From Fig. 5 we can observe that the medical industry has the greatest number of breaches throughout every breach type. This means that the medical industry is one of the most sought targets due to the nature of the data. Reference [10, Table 1] had similar findings, as they show medical with 61.55% of data breaches from 2005 to 2019 as compared to BSO (business other) at 6.7%. Nevertheless, it is crucial to understand the effect that these attacks have, which is why we must also look at the impact. In Fig. 6 we can see that in terms of records lost, BSO is almost at 8 billion. This is just one category of business and when adding all three, the total loss of records in business amounts to 9.7 billion. A staggering number when compared to the 246 million records lost in the medical industry. So, although the medical industry has a higher occurrence of breaches, the business sector loses far more records.

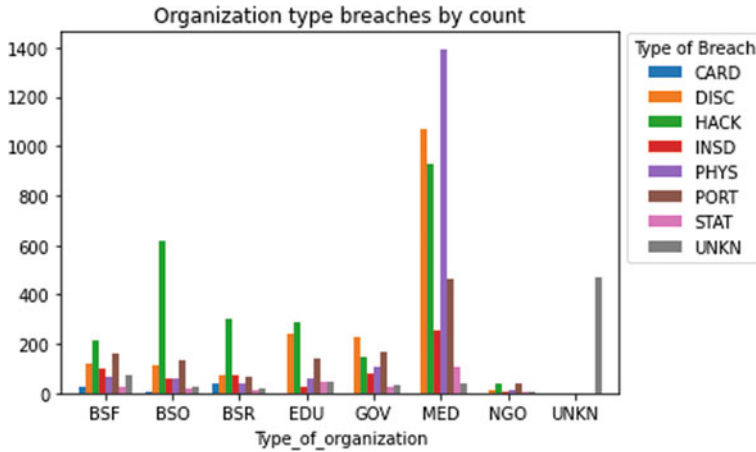


Fig. 5 Frequency on type of breaches per type of organization

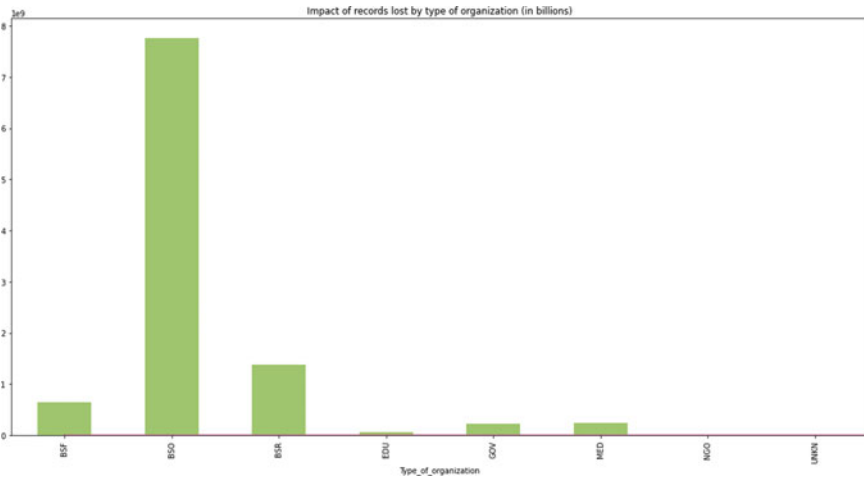


Fig. 6 Sum on total records lost by type of organization

4.3 Trends on States

To expand on the subject of impact versus frequency, we applied this concept to the state’s column in the PRC dataset. First, we had to clean the abbreviated states up as well as remove outliers outside of the US. Then we were able to plot them based on the number of occurrences.

Figure 7 shows California as a landslide past all other states, with New York about half of the way through. This is largely due to California being the most populated state with 39 million people [11], so more people naturally mean more records. The

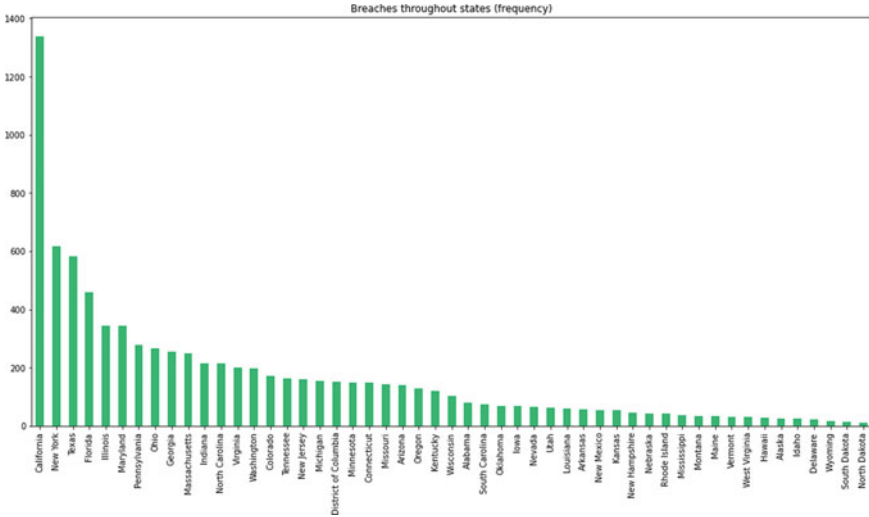


Fig. 7 Frequency on states

top five states here are California, New York, Texas, Florida and Illinois. These states are top breached because they either have more people or more businesses. Now, we compare the frequency to the impact on the total number of records lost. Figure 8 shows the reality of severity within those total counts of breaches. Here the top five states are California, Oregon, Florida, Georgia, and Texas.

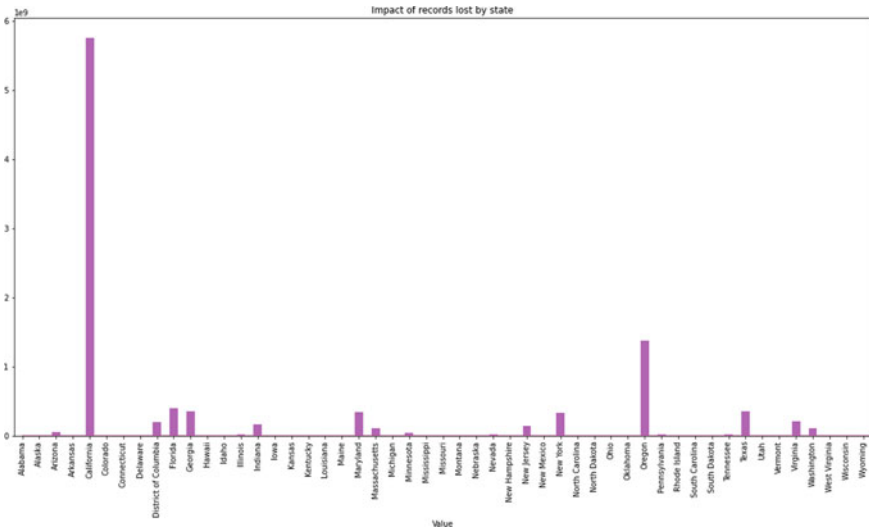


Fig. 8 Sum on total records lost by state

California remains the top state for frequency and impact, but Oregon is now in second place, replacing New York. Without taking a closer look at the impact, we would have missed this as in Fig. 7 Oregon is somewhere in the middle of the graph. After looking into the cause, we found that a huge disclosure for over 1 billion records happened to River City Media in Oregon. Similarly, the reason that California is a top state is because of certain outliers such as the Yahoo breach. When weighing impact side by side with the second dataset, the Anthem hack is another outlier for Indiana, making it the most impacted state. In order from most to least, after Indiana is New York, Florida, California, and Texas (Appendix 1). The top five states in both datasets are almost the same, sharing California, Florida, and Texas. These three states are the most populated states, and where there is a higher population there is a higher chance for human error. Political and economic conditions may also contribute to cyberattacks, since economically developed states may have access to more resources to quickly respond to a threat. An IBM report showed that the average amount of days it takes to identify and contain a data breach is 277 days [9]. So, companies who shorten the life cycle of a breach will save money as well as protect their reputation and remaining data.

5 Effect on the Business Industry

This section outlines the breaches in the business industry and the extent of the data loss. One of the most devastating business breaches was Yahoo for three billion records which happened sometime in 2013 and then another breach again for five hundred million records in 2014. The 2014 attack was not reported until September 22, 2016, and the 2013 attack was delayed even further to be disclosed on December 14, 2016 [12, p. 1278]. The severity of this breach was the result of poor timing [13] in incident response as well as poor password management. An attack on this level can be very dangerous, so such a delayed reaction is negligent of Yahoo. The second attack could have been prevented with an increased amount of security such as requiring users to reset their passwords. This is a perfect example of how human error played a part in the magnitude of these data breaches. Thankfully, Yahoo does not hold information such as social security numbers, so all that was stolen was emails, usernames, passwords, dates of birth, and security questions [14]. On the other hand, this can be quite dangerous if these credentials are used for other websites that do contain sensitive information. Nevertheless, to this day Yahoo has not quite fully regained its trust and reputation with the public.

A closer look at the impact throughout the years showed that 2016 had the greatest number of records lost with over 4 billion losses (Fig. 9). As Yahoo did not publicly disclose its breaches until 2016, its case is a pretty big outlier. Furthermore, the runner-up for most records lost was 2017 with almost 2 billion records. During this year, the spam email company River City Media in Oregon accidentally leaked 1.37 billion customers' records with information such as names, email addresses, physical

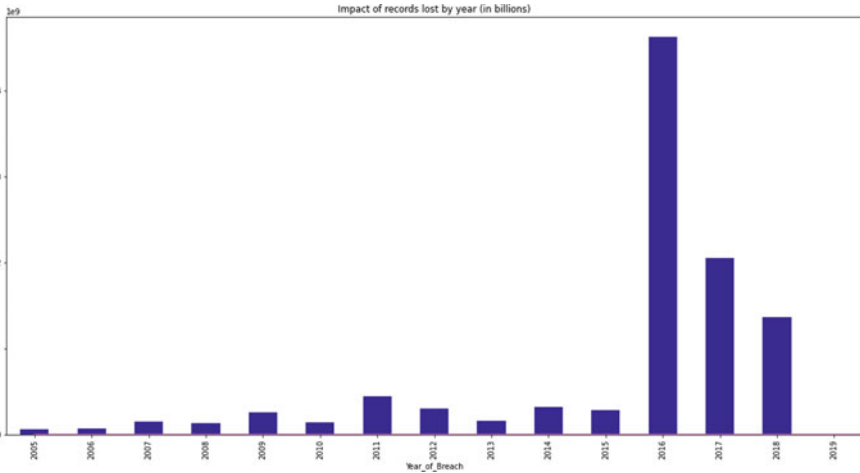


Fig. 9 Sum on total records lost by year

addresses, and IP addresses. We will later compare this to the effect on the medical industry with a similar graph.

Further focusing on River City Media’s unintended disclosure, [15] explains that the company left backup MySQL databases up online for at least three months as an oversight. These servers were not password protected, nor were they monitored as an employee from MacKeeper stumbled upon them. A case like this points to human error and negligence as there should be proper measures taken to ensure the data is protected.

To conclude this section, we can say that if companies face large-scale data breaches, then the customers will lose their trust. Individuals share their personal information and if an organization can’t keep their data secure, then they will lose their reputation and clients. This, in turn, results in a loss of revenue as new clientele will avoid companies with a poor reputation. While this is the indirect effect, the direct effect consists of customer lawsuits as well as hefty fines of up to \$43,792 per violation according to the Federal Trade Commission [16].

6 Effect on the Healthcare Industry

The objective of this section is to identify the cybersecurity trends in the medical industry and compare the effects to that of business. While businesses get hacked for more records, medical records are worth substantially more money on the dark web, especially when they contain social security numbers. Since a social security number never changes, it is more valuable than a credit card number as it can be used for things like identity theft and tax fraud. Moreover, the IRS pays billions of dollars

every year for tax refunds to victims who suffer identity theft. Thus, why the medical industry saw a 300% increase in daily ransomware attacks from 2015 to 2016 [17].

In fact, in IBM's cost of a data breach report, for the last twelve years, the medical industry has had the highest average total cost of a data breach with \$10.10 million as compared to the national average of \$9.44 million [9]. The average cost of a breached record is increasing which is linked to costs associated with paying ransoms, lawsuits, fines, and containing the data breach itself. Another indirect cost is lost sales from a reduction in customers due to loss of trust. This in turn affects the consumers as companies deliberately raise the prices of services and products to keep up with the costs of cybersecurity maintenance. IBM's report found that "60% of breaches led to an increase in prices passed onto customers" [9].

Through our analysis of the Health and Human Services dataset, we have found that the peak of attacks where individuals were affected was in 2015. This is attributed to the Anthem network server hack in Indiana which resulted in 78,800,000 individuals being affected.

Figure 10 shows the dramatic difference between 2015, with over one hundred million individuals impacted, to the next highest year 2021, with slightly over fifty million. The Anthem hack is one of the largest [18] healthcare industry attacks to ever happen. Correspondingly, Chinese hackers utilized a spear-phishing technique where they gained administrative access to Anthem's data warehousing infrastructure containing both customer and employee personal information. This information contained social security numbers, physical addresses, names, dates of birth, etc. However, other information such as test results and financial details remained safe. Moreover, spear-phishing is when emails are sent to target specific individuals or organizations ostensibly, or disguised as trusted companies, to gain confidential information. The aftermath of the Anthem attack included several hefty lawsuits totaling \$115 million [18] for insufficient procedures as well as identification and response to suspected or known security incidents. This is seen as a recurring theme in the largest data breach attacks, as typically, the company targeted is lacking a strong cybersecurity procedure.

Furthermore, after cleaning up the categories in the type of breach column, from the second dataset, we can see that Hacking/IT incidents to email and network servers are the top data breach reasons. In Fig. 11, network server attacks affect the highest number of individuals, so in the recommendations section, one issue we will focus on is network server protection. Since this dataset spans from 2009 to 2022, it is safe to say that the threat of ransomware is steadily increasing [10, Fig. 4]. As more ransomware is being created, a hacker has options whether it is to utilize malware through techniques such as phishing or Distributed Denial of Service attacks (DDoS). With an approach like phishing, it is easy for the hacker as all it takes is for one person to fall for a fake email and click something which will act as a vessel to deliver the ransomware. The ransomware will then gain access using the person's credentials. On the other hand, DDoS attacks are more focused on a network, so to achieve shutting it down they flood the target with traffic until it crashes. This causes the cybersecurity team to shift their attention, which provides an opening to steal information. DDoS attacks are among the most dangerous attacks [19, Fig. 17] as they disrupt internet

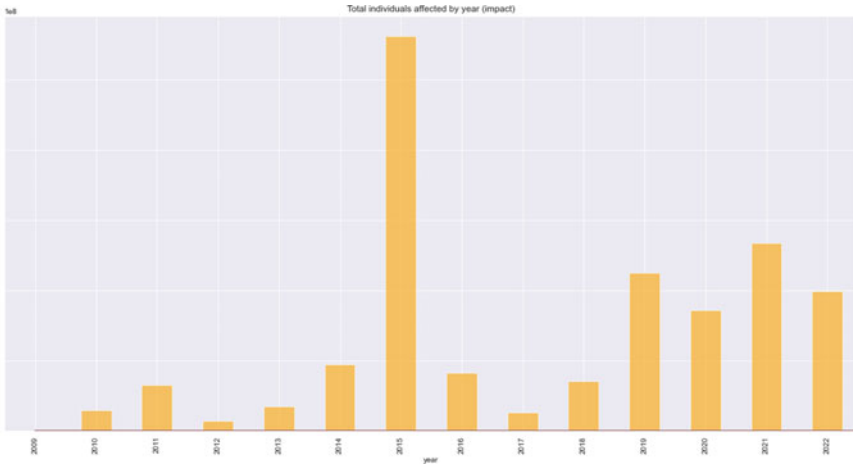


Fig. 10 Sum on total individuals affected by year

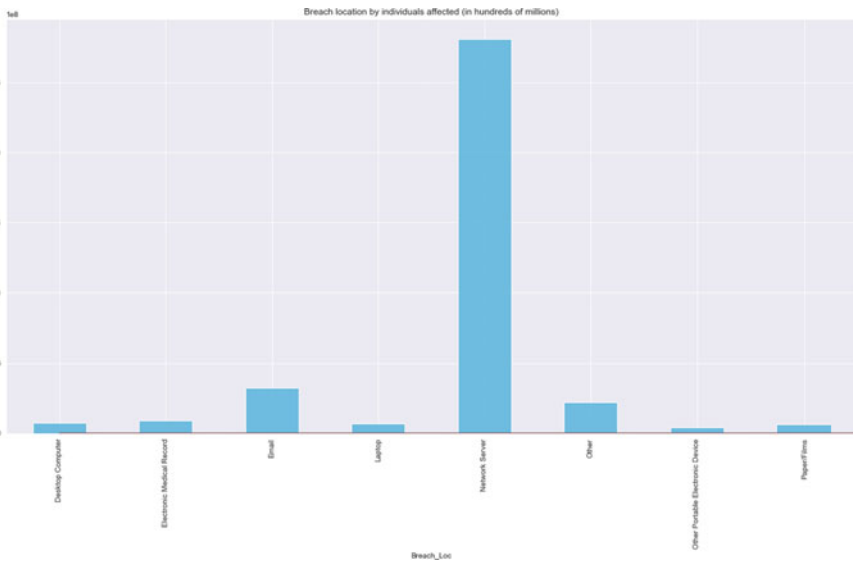


Fig. 11 Sum on total individuals affected by breach location

security [20]. The 2016 Dyn attack was a memorable DDoS attack, as the network was infected by a new malware named Mirai botnet with a strength of 1.1 TBps. With multiple malicious endpoints, it brought down hundreds of websites for several hours [21].

Although hacking and malware result in the most devastating breaches, there are other threats to the medical industry such as portable device loss. In particular, if

we look at the loss of 76,000,000 veterans' medical information in the PRC dataset, this is the result of purely human error. In this case, the hard drive was sent back to the manufacturer for repairs without wiping the memory clean first. With the right data safety precautions, as outlined in [22], a loss like this could easily be prevented. As mentioned before, medical information is very valuable and once stolen can take a while to show up and be hard to pinpoint. Once there is credit damage it is very tedious to undo and could take years.

To wrap up this section, we can deduce that the medical industry is at higher risk and the risk will grow over time if preventive methods are not applied. Not only are these healthcare institutions losing valuable patient data, but the cost associated with this loss is also increasing for both consumers and companies. Among the most compromised are network servers and phishing emails. This forecasts that stronger cybersecurity systems must be implemented, along with a reduction in human error through awareness and fewer opportunities to make mistakes. The effects on the medical industry are harrowing as not only do they lose integrity and confidentiality but there can also be a disruption of care to the patients. With damage to infrastructure, if a doctor cannot access medical records or a patient's health insurance, then this prevents them from providing necessary medical care. Proper response time is crucial here, and we can see that this is evident among responses to cyberattacks as well.

7 Recommendations/Preventative Measures

To start, one of the most vital recommendations we have is for companies to build a strong cybersecurity team that actively monitors, identifies, and contains threats. Reducing response time can minimize the extent of the data loss. As in the IBM report, it takes an average of 304 days for organizations to identify and contain a breach [9]. A great tool that companies can utilize to mitigate the threat more quickly is an artificial intelligence platform called Extended Detection and Response (XDR). AI such as XDR can reduce response time by about 74 days as well as save on average \$3 million on breach costs [9].

It is also crucial that companies who have experienced a data breach up their security measures as they are at higher risk due to methods such as hackers stealing cookies. Hiring a stronger cybersecurity team may be a helpful option for these companies. In fact, in our study (Fig. 12), it is found that when a business associate is present in the medical industry, then the probability of attacks is less or less severe. We see this example play out in the River City Media loss where the records were left up for three months. When there is a business associate present, they can ensure that data protection measures are in place and can identify any possible cybersecurity incidents before the problem exacerbates.

Another factor that is important to consider is that as we move forward with a remote workforce, there will be challenges with cybersecurity. This is why employees need to have strong antivirus and firewall systems as well as training. The more educated employees are, the better they can protect themselves and the data to that

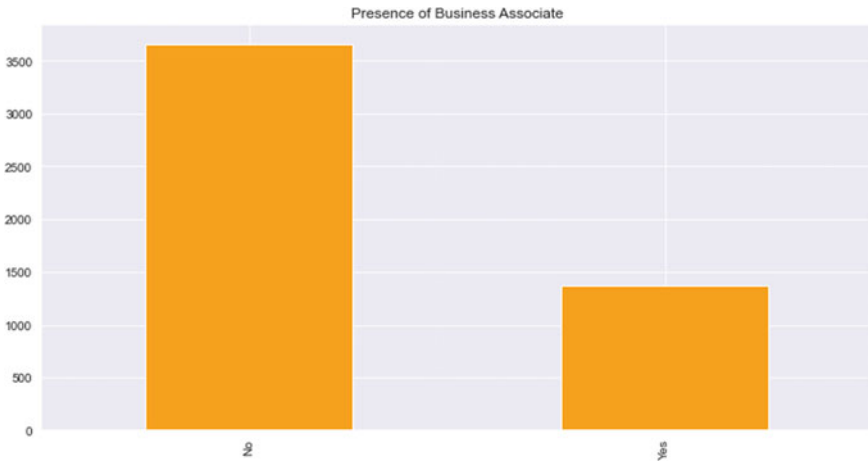


Fig. 12 Count on business associate present

they have access. When it comes to cybersecurity, awareness is half of the battle since many people fall for things like phishing emails, which continue to be the most expensive breach type at \$4.91 million on average [9]. Human error will always be an element, so although we cannot mitigate it, we can focus on minimizing the possibility of it occurring. In Verizon’s data breach report, the human element is involved in 82% of breaches and is often a gateway into a larger breach which is why a strong cybersecurity system is important [19]. A way to do this is through zero trust architecture, which is a cybersecurity design that follows the principle of “never trust, always verify” [23]. This approach includes more authentication methods as well as enforcing stricter lateral movement in the environment. Organizations that deployed this technique saw a decrease in the cost of breaches by an average of \$1 million [9].

Lastly, network server protection is major as we saw that it was a top trending target in the healthcare dataset. To mitigate server hacks there are several proposed protocols. One of them is called Heal the Breach “which randomly modifies the size of HTML responses before encryption is applied” [24]. This makes it a lot harder for hackers to figure out the secret guess token as it introduces randomness and constantly changes sizes. Another way to protect a network server is via the zero-trust architecture previously mentioned. Many companies have started switching over to things like dual authentication to increase security measures.

8 Results

Incidents based on cybersecurity have been increasing substantially leading into identity theft with unintended disclosure establishing loss of reputation in the businesses. Concisely, the prime reason behind the loss of information both in the medical periphery as well as the businesses serve to the hacking implicating cybersecurity concern. In addition, malware also serve to cause potential threat resulting into tampering of information worsening the situation of business or medical data management. On the other hand, the nation witnessing substantiate data breaches include California and Oregon on the basis of certain outliers present in sensitive and confidential data. Prominent incident of cyber security violations includes Yahoo and IBM data breach. Hence, data breach configures into adverse impact on both businesses and personal lives.

9 Discussion/Conclusion

9.1 Data Breach Trends in Businesses

Based on the findings, it has been observed that data breaches have increased at a considerable level owing to malicious attack and unauthorized access to sensitive data. The propensity of data breaches does not serve to be equal for all the industries in representation to the lucrative possibilities for the hackers as well as the cyber-attackers. The representation below showcases mean cost of data violation industry wise from the month of May 2020 to March 2022 [25]. However, most prominently, healthcare sector has been witnessed towards being adversely impacted by data breaches. As per the data of 2022, the mean cost of data breach for healthcare organizations serves to be 10.1 million US dollars (Fig. 13). In addition, as per the report of [26], within the initial quarter of 2023, over 6 million data reports were subjected to data breaches. The fourth quarter of 2020 have been into highest number id data breaches worth 125 million data sets (Fig. 14).

This paper studies the ongoing cybersecurity breach trends in the United States in the last fifteen years as well as proposes mitigation strategies for those trends. Although the medical industry is targeted more frequently, the business sector is further impacted as it loses thirty-nine times more records. Additionally, we compared types of breaches to find that hacking/malware is the top threat. With that said, the largest medical hack in the datasets was a 2015 phishing attack on Anthem in Indiana, and the largest business hack was in 2016 on Yahoo in California. The Anthem breach was the result of social engineering tactics used to gain access to the network server. Whereas Yahoo was breached in 2013 and 2014, and these incidents were not reported until 2016 in which their delayed response cost them their reputation. What these two incidents have in common is that human error played a big role, and stronger cybersecurity systems could have alleviated the extent

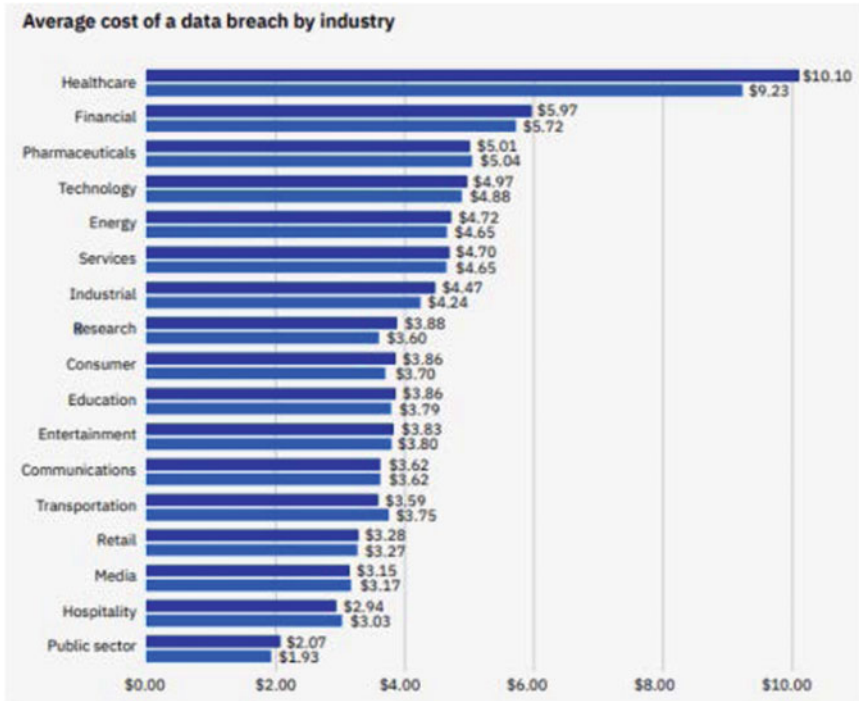


Fig. 13 Data breach cost by industry

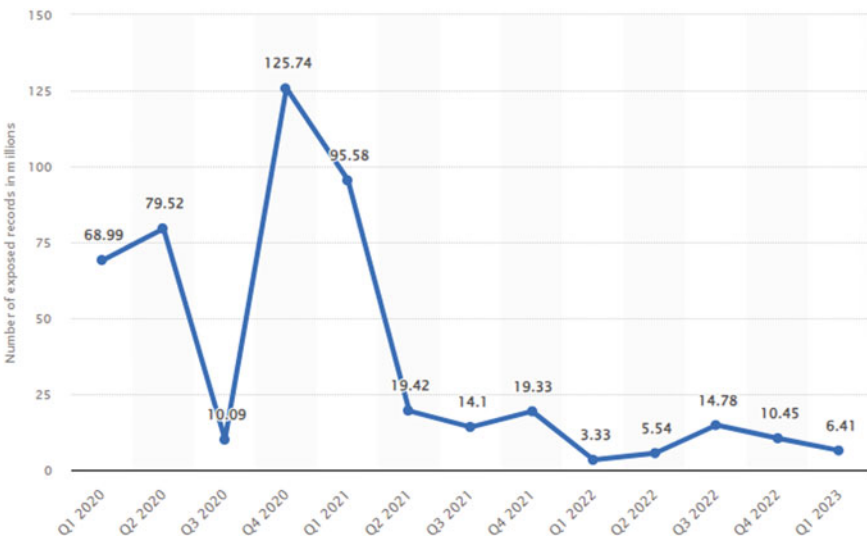


Fig. 14 Worldwide exposure of data records

of the loss. Moreover, our research analysis leads us to reject the null hypothesis, and accept the alternative that human error is one of the top reasons for data breaches. Essentially, we found that cyber security threats are dynamically increasing in cost and count, and organizations are under excessive pressure to have a faster response time. Certain companies are implementing artificial intelligence to help decrease the cost and time of a data breach. Other measures that could be taken are having a strong firewall system, using dual authentication or zero trust architecture, and having a diligent cybersecurity team that monitors any possible threats. The sooner a threat can be identified and contained, the less damage that is done.

Future studies may involve collecting more data on the business industry to collate trends. It is also possible to remove outliers to reduce the noise to better spot any underlying data direction. We can also use bins for the records column to normalize the values. Since the data has mostly been cleaned and analyzed, more features could be created which would allow us to move into the model development phase. If we are unable to create sufficient additional features, then we could always continue analyzing the data to extract information. All in all, there is a multitude of different ways to subset the data to take a closer look at certain features and their corresponding trends.

Appendix 1: Impact of Individuals by State

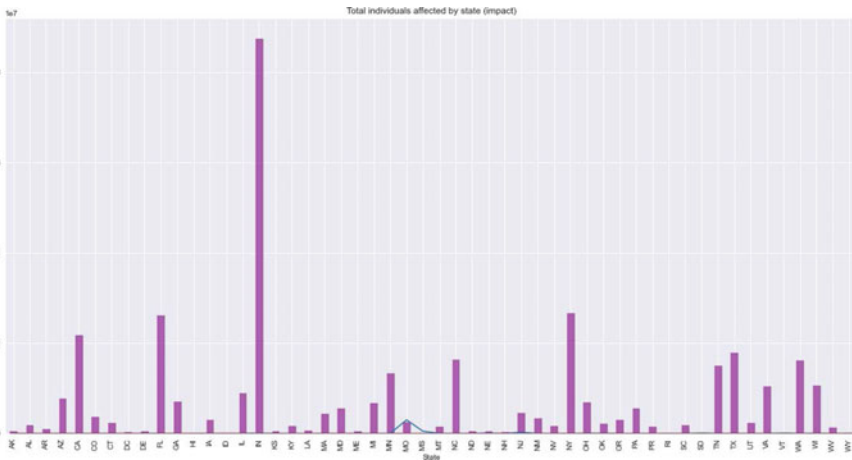


Fig. 15 Sum on total individuals affected by state (Health and Human Services)

References

1. Advisen (2020) Insurance data, media, and technology. Advisen Ltd. <https://www.advisenltd.com/>. Accessed 18 Dec 2022
2. What are the penalties for violating HIPAA? Penalties for Violating HIPAA | American Dental Association. <https://www.ada.org/resources/practice/legal-and-regulatory/hipaa/penalties-for-violating-hipaa>. Accessed 18 Dec 2022
3. Cremer F et al (2022) Cyber risk and cybersecurity: a systematic review of data availability. *Geneva Pap Risk Insur - Issues Pract* 47(3). <https://doi.org/10.1057/s41288-022-00266-6>
4. Ou CX, Zhang X, Angelopoulos S, Davison RM, Janse N (2022) Security breaches and organization response strategy: exploring consumers' threat and coping appraisals. *Int J Inf Manage* 65:102498. <https://doi.org/10.1016/j.ijinfomgt.2022.102498>
5. Yeo LH, Banfield J (2022) Human factors in electronic health records cybersecurity breach: an exploratory analysis. 19(Spring):1i
6. Seh AH et al (2020) Healthcare data breaches: insights and implications. *Healthcare* 8(2):133. <https://doi.org/10.3390/healthcare8020133>
7. Javaid DM, Haleem A, Singh DRP, Suman DR (2023) Towards insighting cybersecurity for healthcare domains: a comprehensive review of recent practices and trends. *Cyber Secur Appl* 1:100016. <https://doi.org/10.1016/j.csa.2023.100016>
8. Abbiati G, Ranise S, Schizzerotto A, Siena A (2021) Merging datasets of cybersecurity incidents for fun and insight. *Front Big Data* 3. <https://doi.org/10.3389/fdata.2020.521132>
9. IBM - United States. <https://www.ibm.com/downloads/cas/3R8N1DZJ>. Accessed 15 Dec 2022
10. Seh AH, Zarour M, Alenezi M, Sarkar AK, Agrawal A, Kumar R, Ahmad Khan R (2020) Healthcare data breaches: insights and implications. *Healthcare* 8(2):133. <https://doi.org/10.3390/healthcare8020133>
11. Population estimate for 2021. Rank List: States in Profile. https://www.statsamerica.org/sip/rank_list.aspx?rank_label=pop1. Accessed 18 Dec 2022
12. Trautman LJ, Ormerod PC (2017) Corporate directors' and officers' cybersecurity standard of care: the Yahoo data breach. *Am Univ Law Rev* 66(5):3. <https://digitalcommons.wcl.american.edu/aulr/vol66/iss5/3>
13. Wang P, Park S (2017) Communication in cybersecurity: a public communication model for business data breach incident handling. *Issues Inf Syst* 18(2):136–147. https://doi.org/10.48009/2_iis_2017_136-147
14. McCandless D (2022) World's biggest data breaches & hacks. Information is Beautiful. <https://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/>. Accessed 14 Dec 2022
15. River City Media. <https://privacyrights.org/data-breaches/river-city-media>. Accessed 14 Dec 2022
16. Fair L (2021) Latest FTC notice of penalty offenses tells 700+ national advertisers that deceptive endorsements can lead to financial penalties. Federal Trade Commission. <https://www.ftc.gov/business-guidance/blog/2021/10/latest-ftc-notice-penalty-offenses-tells-700-national-advertisers-deceptive-endorsements-can-lead>. Accessed 18 Dec 2022
17. Argaw ST, Bempong N-E, Eshaya-Chauvin B, Flahault A (2019) The state of research on cyberattacks against hospitals and available best practice recommendations: a scoping review. *BMC Med Inform Decis Mak* 19(1). <https://doi.org/10.1186/s12911-018-0724-5>
18. Mohammed Z (2021) Data breach recovery areas: an exploration of organization's recovery strategies for surviving data breaches. *Organ Cybersec J: Pract, Process People* 2(1):41–59. <https://doi.org/10.1108/oj-05-2021-0014>
19. 2022 data breach investigations report. Verizon Business. <https://www.verizon.com/business/resources/reports/dbir/>. Accessed 18 Dec 2022
20. Chakraborty S, Kumar P, Sinha DB (2019) A study on DDOS attacks, danger, and its prevention. *Int J Res Anal Rev* 6(2). <https://doi.org/10.1729/Journal.20847>
21. Kolias C, Kambourakis G, Stavrou A, Voas J (2017) DDoS in the IOT: Mirai and other botnets. *Computer* 50(7):80–84. <https://doi.org/10.1109/mc.2017.201>

22. Pesante L, King C, Silowash G (2012) Disposing of devices safely - CISA. US-CERT (United States Computer Emergency Readiness Team). <https://www.cisa.gov/uscert/sites/default/files/publications/DisposeDevicesSafely.pdf>. Accessed 18 Dec 2022
23. Wylde A (2021) Zero trust: never trust, always verify. In: 2021 international conference on cyber situational awareness, data analytics and assessment (CyberSA). <https://doi.org/10.1109/cyberesa52016.2021.9478244>
24. Palacios R, Fernandez-Portillo AF, Sanchez-Ubeda EF, Garcia-De-Zuniga P (2022) HTB: a very effective method to protect web servers against breach attack to HTTPS. IEEE Access 10:40381–40390. <https://doi.org/10.1109/access.2022.3166175>
25. Binns R (2023) Top 15 data breach statistics for 2023 - trends and insights. Website Builder Expert. <https://www.websitebuilderexpert.com/building-websites/data-breach-statistics/#:~:text=According%20to%20IBM%2C%20data%20breaches>. Accessed 16 Aug 2023
26. Petrosyan A (2023) Quarterly online data breaches 2022. Statista. <https://www.statista.com/statistics/1307426/number-of-data-breaches-worldwide/>

Identifying Daesh-Related Propaganda Using OSINT and Clustering Analysis



Gianluigi Me and Maria Felicita Mucci

Abstract The development of the digital society has substantially altered the conditions under which conflicts occur. Emerging threats are characterized by their asymmetry, diversity, and constant change; rapid transmission over the network; near-immediate nature; possibility for unrestricted access; and swift ability to alter the behaviour of individuals. This paradox is an example of cognitive warfare, which employs both traditional and novel information, cyber, and psychological warfare techniques. The self-proclaimed Islamic State engages in a unique type of disruptive cyber cognitive-intelligence activity utilizing cyberspace. We now refer to the Weaponization of Media Narratives: the struggle of narratives has overtaken the relevance of traditional military and physical Jihad. Jihadist activities consist of sending threatening messages to Western nations and promoting online propaganda in order to recruit new members and instil terror in individuals. Daesh's propaganda output is so extensive that it is practically impossible for humans to analyse it. Thus, it is crucial to establish and implement cyber defence strategies to prevent, identify, and deter jihadist Internet activity. Law Enforcement, Intelligence, and other organizations are constantly devising new tools to prevent, identify, and restrict terrorist operations over the Internet. The collection and analysis of information from a vast array of sources can give intelligence analysts with useful insights by revealing previously concealed but logically sound patterns and connections. Beginning with a review of Al-Naba's propaganda materials, this study seeks to construct an automated model that would aid in detecting and identifying the online locations of Daesh. We looked at Al-Naba' magazine instead of another newspaper because it has only been published in Arabic. Other magazines have been published in other languages and have been looked at in a lot of community identification and Social Network propaganda analysis studies in the past. Therefore, the purpose of our study was to discover if it is possible to employ computer assistance to evaluate Jihadist tales in order to identify

G. Me

Department of Economics and Finance, Luiss Guido Carli University, Rome, Italy
e-mail: gme@luiss.it

M. F. Mucci (✉)

S & A I Sistemi & Automazione S.p.A., Rome, Italy
e-mail: mfmucci@sealink.it

any (thematic) similarities across various propaganda sources. One of the specific goals was to evaluate whether or not there are tweets with a direct connection to Al-Naba' magazine. We wanted to make sure that the tweets were coded in a way that was consistent with the Twitter data—collected from Kaggle—we used as a training set. This was important because tweets could be put into different groups. This was done to see if the tweets were correctly put into their own groups based on information from Al-Naba's writings. So, the number of times each group shows up depends on how often it shows up in more than 1% of the texts in each cluster.

Keywords Clustering analysis · Cognitive warfare · Counterterrorism · ISIS · National security · Virtual Jihad

1 Introduction

The notion of security in the field of International Security studies poses a challenge due to the absence of a widely accepted definition for the term. In most cases, the traditional framework is portrayed as necessitating the safeguarding and advancement of the fundamental principles of nation-states, such as sovereignty and territorial integrity. The assurance of safety for the citizens of a nation is symbolic of the nation's security being in a state of excellence.

The concept of security has been significantly influenced by the dynamic historical context of global affairs and the corresponding power structures. Over the past half-century, there has been a significant shift in the global landscape. The aforementioned transition is characterized by a departure from regional governance structures and the emergence of novel concerns and requisites at the forefront of the worldwide and societal discourse. The discourse pertaining to security, conversely, has transcended the realm of purely military matters and encompassed a wider range of obstacles.

The aforementioned emerging concerns encompass a range of topics, including but not limited to economic matters, environmental issues, energy policies, health-care, gender dynamics, multiculturalism, migration patterns, and the realm of cyberspace. Each of these domains of operation has undergone securitization to some extent, as they are currently perceived as posing risks to the security of the nation. Furthermore, there exist novel participants and domains for the execution of activities. Presently, we inhabit a highly intricate global landscape that is marked by a period of transformation and differentiation, alongside megatrends that are fundamentally at odds with each other.

What was one of the most pivotal factors that facilitated this evolutionary process? The phenomenon of globalization is a prominent feature among various other factors. Globalization has been identified as the primary factor responsible for the widespread disruption of the established international security framework. The utilization of this tool is intended to discern a multitude of societal mechanisms that are perceived to be converting our present cultural actuality into a universal one. The phenomenon of globalization has resulted in a reduction of geographical and temporal barriers,

leading to the emergence of global issues and social benefits. This has posed a challenge for individual states to address global security threats and hazards independently. Furthermore, it has facilitated greater interconnectedness among societies and nations.

The merging of Defense and Security, hybrid threats and warfare, and a multi-dimensional and dynamic space are exerting an influence on the evolution of the Westphalian State framework and the concept of warfare. The emergence of the Internet has had extensive implications on various aspects of society, including politics, commerce, and everyday existence. The widespread adoption of information and communication technology has rendered temporal and spatial boundaries inconsequential. The emergence of the digital age has generated numerous prospects that engender a positive outlook for the future of the planet.

Nevertheless, the prevalence of cyber threats and susceptibilities persists as a concern. The emergence of the information society has significantly altered the circumstances in which conflicts are conducted. Contemporary hazards are characterized by their lack of symmetry, diversity, and constant evolution; their rapid dissemination through the network; their nearly instantaneous nature; their potential for unrestricted entry; and their ability to quickly modify individuals' attitudes.

The advent of sophisticated digital technologies and the proliferation of social media platforms have facilitated the process of artists reaching a larger and more targeted audience through customized messages that resonate with them. The expeditious dissemination of visual images and concepts through social media has rendered it a convenient tool for governments, terrorist groups, and insurgents seeking to alter the public's perception of wartime occurrences. Contemporary terrorists represent the initial cohort to have been raised amidst the widespread availability of digital communication technologies, including but not limited to the Internet and social media platforms. Hence, it is unsurprising that these networks play a crucial role in the process of radicalization and recruitment, which they employ to lure individuals who are susceptible to their messaging.

Erdem and Bilge [1] argue that the self-proclaimed Islamic State divides the world into two distinct camps, namely *Dar al-Islam* and *Dar al-harb*. The initial group is affiliated with the Sunni community, whereas the latter encompasses individuals who do not adhere to the Islamic faith ("apostates," "pagans," "infidels," "crusaders," among others). Daesh perceives all individuals who identify as Sunnis as prospective adherents or sympathizers who could be convinced to endorse its ideology and engage in combat on its behalf.

Conversely, the terrorist group maintains the belief that reconciliation with non-believers is unattainable and that the only viable solution is to employ violent means to conquer them (*Dar al-harb*). The partition is an indispensable element in the discourse of Information Systems (IS), as it conveys distinct messages to the two distinct groups of individuals.

To clarify, the terrorist organization known as Daesh aims to recruit individuals belonging to the former category, while those belonging to the latter must be intimidated in order to be overcome. The primary focus of Daesh is to target individuals who

belong to the Arab Sunni community, specifically those within the Sunni community. This is done for two distinct ideological reasons.

The initial rationale is that, in accordance with Islamic tradition, the Caliphate can solely be instituted on land that is possessed by individuals of Arab descent. The recruitment of Arabs by the organization can yield a second benefit in the form of increased respect, particularly in neighbouring territories, thereby enhancing the dissemination of its message. Daesh alludes to a symbolic framework that amalgamates historical, theological, and political motifs, with the intention of appealing to Muslims at large and the Arabic-speaking populace in particular.

Jihadism can be considered a contemporary phenomenon, characterized by novel objectives and tactics. The utilization of innovative technology has contributed to the reinforcement of contemporary and modern aspects of Islamism, which have been facilitated by the increased accessibility of such technology due to globalization. The emergence of novel methods for bringing a community to the forefront of people's minds can be attributed to technological advancements, the communications revolution, and enhancements in information storage and retrieval. Islamists have demonstrated proficiency in utilizing the resources of globalization to advance their objectives. In contemporary times, social media and online forums have emerged as a potent tool for propagandizing, inciting violence, and radicalizing a significantly larger audience than previously possible.

Daesh has utilized various social media platforms such as Facebook, Twitter, and blogs to disseminate propaganda, enlist new combatants, enhance the skills of current ones, acquire financial support, showcase their successes, and influence the public's perception of the ongoing conflict. The strategic significance of online media channels is evidenced by the adoption of these channels by terrorist organizations. In the contest to win the support of the Umma, the use of storytelling tactics has superseded the utilization of traditional weaponry and military power. The utilization of propaganda as a tool for offensive information warfare is a significant aspect of the Jihadists' campaign. Hence, the media and internet resources possess the potential to serve as a potent psychological tool, provided they are utilized judiciously. Daesh has adopted a narrative-driven and intensified form of terrorism as its primary asymmetric weapon.

Within this context, various institutions such as law enforcement and intelligence agencies are consistently devising novel tactics aimed at thwarting, identifying and impeding terrorist activities on the internet. The efficacy of traditional investigative techniques in promptly detecting possible terrorist risks has led to their resurgence in popularity. In recent decades, there has been a growing importance placed on terrorist informatics, which involves gathering and analysing data from diverse sources to offer pertinent insights to intelligence analysts. Undoubtedly, the application of text analysis to extensive datasets can uncover patterns and connections that were previously concealed but are logically valid.

Our empirical research and methodological approach started from the collection of the most relevant issues of *Al-Naba'* newspaper, published on a regular basis by both the Jihadists websites we had found. *Al-Naba'* features a variety of articles and material types, such as news, commentary, visualizations, religious writings

(including fatwas), and advertisements for various forms of media output. So, it totally reflects Daesh mindset and mirror events occurring on the ground, and social issues particularly relevant to the residents of the self-proclaimed Islamic State.

The development of methodologies and protocols for analysing extensive quantities of data presents a significant challenge due to the overwhelming volume of propaganda disseminated by Daesh, rendering manual analysis by humans nearly impracticable. A comprehensive understanding of the various taxonomies utilized in propaganda, the disparities in narrative across different media platforms, and the manner in which it evolves among users is imperative for the formulation of counter-narratives and counter-messaging tactics.

The reason why we analysed *Al-Naba'* magazine rather than another newspaper is because it has only been published in Arabic, whereas other journals have been released in other languages and have been the subject of numerous prior studies on community detection and Social Network propaganda analysis.

In the study conducted by Krebs [2], social network analysis was employed as a methodology to investigate terrorism. The analysis shed light on the connections between the individuals involved in the 9/11 hijackings and other terrorists. Notably, unstructured data from freely accessible sources such as newspapers was utilized in this research endeavor. Brams et al. [3] and Koschade [4] employed relational data pertaining to individuals in order to reconstruct terrorist networks and examine significant roles and actors within them. The advent of social media platforms has facilitated the accumulation of a vast reservoir of data pertaining to online interactions among individuals, with a particular emphasis on the ramifications of cybercrime and terrorism.

The application of cluster analysis has been employed in the classification of terrorist organizations through the utilization of multi-membership clustering techniques and similarity scores. The GPACT methodology, as formulated by Lautenschlager et al. [5], employs a multi-stage process that integrates the clustering of terrorist incidents to generate comprehensive profiles of criminal and terrorist organizations. There is a prevailing belief among experts that Social Media Intelligence (SOCMINT) will assume greater significance in the coming years as a network science-driven counterterrorism instrument.

Empirical research has been dedicated to examining the phenomenon of jihadist activities on the social media platform, Twitter. Research has indicated that there is a growing involvement of “sponsor” accounts associated with terrorist groups in areas affected by insurgency. Additionally, organizational accounts based in Europe that are affiliated with the proscribed militant network Al-Muhajiroun, which operates from Saudi Arabia, have also been observed to play a significant role. In their study, Berger and Morgan [6] conducted an analysis of Twitter data spanning from September 2014 to December 2015 to develop a comprehensive profile of individuals who sympathize with Daesh. Additionally, they examined the influence of Twitter account suspensions on the dissemination of Jihadist propaganda and the activities of ISIS supporters.

In their study, Bodine-Baron et al. discovered that the Arabic nomenclature for the group, as opposed to the commonly used abbreviation “ISIS,” served as a noteworthy distinguishing factor between individuals who supported ISIS and those who opposed

it, with the term “Daesh” being particularly significant. Frequent users of Daesh typically engage in the dissemination of antagonistic material, whereas regular users of “The Islamic State” tend to share content that exalts the group.

The comprehension of user behaviour within the jihadist context and the identification of potential new accounts necessitate the analysis and categorization of existing jihadist accounts, as well as the discernment of distinctive attributes that set them apart. According to the study conducted by Magdy et al. [7], it was observed that the tweeting patterns exhibited by accounts expressing support or opposition towards Daesh were closely aligned with significant news or events that took place around the time when the data was collected. In their study, Kaati et al. [8] devised a machine learning framework with the aim of discerning individuals who endorse terrorist organizations and engage in the dissemination of their propaganda through the social media platform, Twitter. In a study conducted by Rowe and Saif [9], an examination was carried out on the Twitter engagement patterns of individuals residing in Europe. The focus of the analysis was to investigate the impact of utilizing pro-Daesh language and disseminating propaganda on their online activities.

Machine learning has emerged as a superior approach for automated detection of online extremism, as it obviates the need for explicit programming to identify patterns. In their study, Xu and Lu [49] employed a seed strategy in order to augment their dataset for the purpose of identifying extremism. This study employed the Adaboost machine learning technique to autonomously identify and categorize individuals on candidates’ social media platforms as either sympathizers or non-supporters of Daesh.

Ashcroft et al. employed hashtags as a means of collecting and discerning Twitter accounts that express support for Daesh. This study utilized a machine learning-based approach to autonomously identify extremist content. The research community employs the Support Vector Machine, Naive Bayes, and Adaboost algorithms for the purpose of detecting extremist behaviour.

Kaati et al. [8] successfully collected information on Daesh by investigating the hashtags being utilized. They extracted both data-dependent and data-independent features, such as phrases, hashtags, word bigrams, letter bigrams, and temporal elements.

However, no study has been conducted on Al-Naba’. Therefore, the primary objective of this research is to present a novel knowledge-driven approach that can be employed to identify terrorists by analysing information obtained from social media platforms, specifically Twitter, and the primary publications of Daesh on Al-Naba’ magazine.

The process of manually identifying information is impractical owing to the substantial amount of unstructured data, also referred to as Big Data, that is accessible on social networking platforms. Tweets consist of textual content, often accompanied with hashtags, which have the potential to introduce ambiguity in the interpretation of data. The process of fully automatic identification poses significant challenges due to the huge amount of unstructured data involved. A semi-automated framework can be utilized to identify Daesh propaganda tweets, with the option of human intervention in cases of uncertainty to enhance the accuracy of identification.

The proposed methodology is capable of identifying the customary behaviours or “profile” of terrorists. This is achieved through the utilization of a data mining algorithm, specifically clustering analysis, to scrutinize digital content that is linked to terrorism. Subsequently, the algorithm utilizes the derived profile to facilitate real-time detection of individuals who have been flagged for suspected participation in terrorist activities.

2 Strategic Instability in the Digital Domain

The amalgamation of Defence and Security, Hybrid threats and warfare, and a dynamic, multi-domain space are contributing to changes in the Westphalian State structure and the fundamental notion of war. Consequently, there is a growing focus on safeguarding the Internet in its entirety, rather than solely protecting individual domains, particularly in the cyber domain [10].

The emergence of the Internet has had significant effects on various aspects of our society, economies, and international relations. The advent of the digital era is presenting a multitude of novel opportunities, thereby indicating a promising future for the world. In the current era of cybernetics, there has been a significant increase in the amount of information available to individuals. This has enabled people to communicate without being limited by administrative constraints or national boundaries. In contemporary times, due to prolonged technological advancements and progress, countries are more intricately interconnected than in the past. Notwithstanding, there exists a security deficit attributable to cyber threats and vulnerabilities. Disturbing trends in cyberspace, such as Cybercrime activities, can be identified and participated in, encompassing a range of malicious actions executed with the intention of committing a criminal offense, such as the dissemination of hostile disinformation and propaganda. Individuals involved are cognizant of the possibilities presented by this novel setting, and they are diligently utilizing this knowledge to subvert the fundamental veracity tenets it represents.

The current threat landscape is characterized by its asymmetrical nature, diversity, and constant evolution. It is propagated through network channels and has the ability to instantaneously disseminate information. Furthermore, it is readily accessible and has the potential to significantly influence human behaviour. The advent of novel digital technologies and the increasing prevalence of social media platforms have facilitated the ability of agents to efficiently and precisely disseminate tailored content to larger groups of individuals. In addition to the expanded availability of information, the Internet presents an augmented capacity for deceit. The phenomenon of Cognitive Warfare is currently emerging as a pressing concern and a pivotal determinant in the realm of global politics.

In contemporary times, a diverse range of domestic and foreign entities employ internet-based mass propaganda campaigns to disseminate distrust and intensify societal divisions, which could have severe consequences for our safety. Information warfare operations conducted by third countries and non-state actors can potentially

form part of hybrid threats to both domestic security and electoral processes, particularly when they are coupled with cyber-attacks. Consequently, the act of manipulating information poses a distinct threat that necessitates careful consideration when evaluating potential security hazards.

The task of guaranteeing national security is a formidable undertaking, given the presence of these various factors. A comprehensive and interdisciplinary approach is necessary to ensure a thorough understanding of both domestic and global issues at multiple levels. The main aim is to ascertain possible hazards, assess the susceptibilities of the nation, and devise a proficient plan of action. The initial phase in devising a successful approach to counter cognitive warfare involves the determination of the fundamental factors contributing to the problem.

There exist causal factors that are attributed to the individual, which are intrinsically linked to human nature and, consequently, to the fields of psychology and epistemology. The origins of these roots are diverse and their tracing can pose a challenging undertaking [11]. Human beings are vulnerable to the manipulation of information due to cognitive limitations and a deficiency in our knowledge. There exist supplementary societal considerations that warrant attention, including the erosion of trust in governmental and political institutions, the decline in public trust in media outlets, and disillusionment with technological progress. Despite the original intention of the Internet to promote freedom, it has been observed that it has actually resulted in confinement.

2.1 Social Network Warfare

The idea that the information age could have been sustained without the presence of social media is difficult to fathom. Social media platforms have become instrumental in shaping our private communication practices and facilitating the online presence of businesses and institutions.

Notwithstanding the benefits, the mere presence of social media platforms entails a range of hazards that could potentially present significant obstacles to national security and even the stability of the global community. The utilization of social media platforms facilitates the process of manipulating the general sentiment, disseminating erroneous data that may have detrimental effects on an individual's standing, and defaming commercial enterprises and establishments. Given the prevailing conditions, it is plausible that the Internet, specifically social media platforms, could serve as a battleground and a tool in a hypothetical conflict. In comparison to the rudimentary armaments employed in past conflicts, such as rudimentary clubs and projectiles, contemporary warfare is characterized by a sophisticated technological milieu, in which social media has emerged as an unforeseen yet potent instrument of combat.

The aforementioned phenomenon is primarily observed within the framework of a concept commonly known as hybrid warfare. The term "hybrid warfare" is a military concept that pertains to the amalgamation of diverse components of both military and non-military confrontations. In the context of hybrid warfare, it is common for

operations to be executed covertly and without a formal declaration of war. The implementation of covert cyberattacks, surreptitious physical attacks, and covert psychological, communicative, and social manoeuvres, including propaganda and disinformation, constitute a comprehensive strategy [12]. The conflict over information is characterized by actors who are engaged in a struggle to assert their respective political, economic, social, or cultural interests. This phenomenon is observed even in the context of major global powers. Social media platforms are utilized by nations to infiltrate hostile regions, engage in media and information warfare, and oversee and reinforce their cyberspace.

In recent years, the utilization of the Internet and social media has significantly enhanced the efficiency, depth, and accessibility of communication. In the realm of information warfare, words are often regarded as a potent weapon that possesses the capacity to influence the viewpoints of its target audience. Furthermore, they are deemed as a tool of mass disruption that can have tangible effects on real-world targets. The accessibility and affordability of social media tools have significantly enhanced the networking and organizational capabilities of users. The phenomenon has facilitated the reorganization of numerous terrorist organizations into a network-based configuration, thereby enabling independent cells and members to operate with a high degree of autonomy, particularly with respect to disseminating the group's propaganda.

As a result of the implemented restructuring, the organization has demonstrated increased levels of adaptability, responsiveness, resilience, and accessibility [13]. Social media has emerged as a strategic instrument for governments, terrorist organizations, and rebels seeking to shape public opinion due to its rapid dissemination of graphic images and ideas during conflicts. Contemporary terrorists represent the initial cohort to have been raised in an era of pervasive availability of digital communication technologies, such as the Internet and social media platforms. This has rendered them the inaugural cohort with the capacity to enlist in and assume leadership roles within terrorist groups.

Consequently, it is not unexpected that these networks play a pivotal role in the radicalization and recruitment tactics employed to entice vulnerable individuals. In contemporary times, social media and internet platforms have emerged as a highly effective tool for disseminating propaganda, instigating violent behaviour, and radicalizing a significantly larger audience than was previously feasible.

Daesh utilizes various social media platforms, such as Facebook, Twitter, YouTube, blogs, and others, to disseminate propaganda, engage in recruitment efforts, provide training to current members, solicit financial contributions, showcase their military successes, and shape public perception regarding the ongoing conflict [14].

3 Jihadist Cyber and Cognitive Warfare

Contrary to popular belief, the concept of terrorism predates the formation of contemporary nation-states as we currently understand them. The precise mechanics of how this concept is operationalized and the extent to which it is utilized by governments or individuals engaged in oppositional activities against specific nations remains ambiguous and subject to ongoing inquiry. Terrorism is a multifaceted phenomenon that can be driven by diverse political, religious, or philosophical ideologies and is expressed through a range of tactical methods, both in physical and virtual domains. There exist various factors that may motivate individuals to affiliate with terrorist groups or perpetrate acts of terrorism independently, which can be classified into distinct categories. Numerous analysts have attributed terrorism to economic and political factors. The proliferation of terrorist activity can be attributed to various factors, including poverty, unemployment, ethnic, religious, and territorial conflict, as well as political strife. The acquisition of media attention may be considered a primary objective for terrorists. Terrorists are driven by the aspiration to propagate their own extremist ideology and convictions onto others. Terrorist organizations attach significant importance to the media as it provides them with a platform to propagate their ideologies to a wider global audience, thereby instilling fear and apprehension among the masses [15].

The widespread availability of social media in the current information-rich era has facilitated the recruitment of new members and dissemination of propaganda material by terrorist organizations.

In the event of an attack, it is customary for news programs to allocate a substantial portion of their coverage to the incident, often with a focus on the non-state actor involved. Terrorist organizations have the capability to gather a significant number of adherents both domestically and globally, through the utilization of various tactics. According to certain specialists, terrorists have effectively influenced the media's coverage of their attacks. The phenomenon of terrorism is widely regarded as a manifestation of, if not a progenitor to, the media, and there is a consensus among stakeholders that the media has provided terrorist groups with a forum [16].

Online media has contributed to the increase in terrorist activities and their dissemination to the public, as terrorists utilize the media as a primary platform for message dissemination. When examining the operational methods of non-State entities, such as ISIS, it is imperative to give significant consideration to the utilization of digital resources, as they have emerged as a crucial component of terrorist operations [17].

Cyberspace is widely used by violent radical political organizations to engage in a particular form of disruptive cyber intelligence operations. These operations involve the dissemination of menacing messages to Western nations and the circulation of online propaganda materials with the objective of luring and enlisting fresh adherents.

Jihadists are well-known for their successful social media propaganda campaigns on Twitter and Facebook. Moreover, the rising popularity of communication applications such as Telegram and smartphones, which are predominantly impervious due to advanced encryption algorithms, is a significant cause for apprehension. It is

noteworthy that authentic cyber-attacks by the terrorist organization Daesh on crucial national systems have not been reported. The group primarily employs digital platforms to disseminate their propaganda, encourage recruitment, plan assaults, and finance their factions. Daesh has exhibited a lack of proficiency in the development of cyber weaponry or the targeting of critical infrastructure, opting instead to utilize cyberspace as a means of perpetrating acts of violence. The capacity of Daesh cyber operations to instigate individuals to employ violence against entities deemed legitimate by them is arguably the most detrimental and hazardous facet. Despite a perceived decline in the Jihadist threat, it is crucial to note that online terrorist organizations remain a tangible danger due to the persistent and expansive nature of their messaging and propaganda within the digital realm of the World Wide Web.

The utilization of digital platforms for the purpose of generating and disseminating propaganda offers several advantages to terrorist groups. These include enhanced anonymity, which reduces the level of scrutiny from law enforcement agencies, and the ease of use and ability to facilitate rapid communication among a large number of individuals. According to Timothy L. Thomas, the internet is being utilized as a tool for planning by terrorists. Terrorists are able to obtain anonymity, command and control resources, and a variety of other tools to facilitate the coordination and integration of attack options [18].

Terrorist organizations within the Caliphate have demonstrated the ability to disseminate operational directives via the Internet, with a particular emphasis on chat platforms such as Telegram. In contrast to alternative messaging platforms and applications, Telegram offers enhanced privacy and security features for messaging, as well as the option to store media content on the platform's website rather than on the user's device storage. Telegram is a commonly utilized platform by international combatants to engage in recruitment activities and extend invitations to potential recruits to join their respective movements. The Caliphate has employed a strategy, which has been rebranded as *virtual entrepreneurship*, to enlist individuals and sustain their messaging efforts while preserving their anonymity and operating within resource constraints [19].

In contemporary times, there has been a recurrent occurrence of political reactions leading to the obstruction or complete cessation of jihadist sources, websites, and accounts. Notwithstanding, these strategies could be construed as impeding investigations into the creators of the website, as well as the potential for profiles or blogs to be reinstated in the future under altered titles and characteristics. Certain scholars have underscored the significance of establishing a deterrent against ISIS insurgents, a goal that can be achieved through a range of measures. It is important to note, however, that addressing the intricacies and interdependencies of the situation necessitates a combined virtual and physical strategy.

To combat the misuse of Daesh's technology, it is necessary to undertake both individual and collective efforts. Individual efforts may involve making significant investments in innovations aimed at improving the detection and removal of terrorist and violent extremist content. On the other hand, collective efforts may entail collaborating and coordinating with enterprises, institutions, and non-governmental organizations to establish crisis procedures and regulations [20]. It is imperative for

nations to endeavour to formulate and execute cyber Defense tactics aimed at safeguarding their populace, with the ultimate goal of thwarting and vanquishing terrorist organizations. Consequently, the most efficacious approach would be to engage in cooperation with diplomatic networks in order to devise novel and viable strategies, while also adopting uniform tactics and directives to counter and vanquish a shared adversary.

4 Jihadist Operations in the Cyberspace

Non-state actors employ the Internet and diverse social media platforms not just to perpetrate acts of terrorism, but also to endorse a wide range of other objectives. This category encompasses a range of activities including incitement of violence, radicalization, recruitment of new members, planning of future attacks, information gathering, engagement with others, profit generation and funding acquisition, and dissemination of fear. Weimann has identified two distinct categories that can be used to classify terrorists' utilization of the internet, namely communicative and instrumental. The initial classification pertains to the dissemination of propaganda, the execution of psychological warfare campaigns, and the enlistment of fresh adherents. The subsequent classification pertains to the instruction in cyber skills, the formulation and synchronization of cyber operations, and the procurement of financial resources through digital means.

4.1 Cyber-Training

In recent decades, terrorist groups have demonstrated a growing dependence on the Internet as an alternative to conventional training facilities. In contemporary times, certain activities that were traditionally conducted in person have been replaced by their virtual equivalents. Virtual training camps [21] offer a platform for potential recruits to gain knowledge about and extend support to terrorist organizations, while simultaneously promoting engagement in direct actions by granting access to encryption tools and anonymization techniques. Instructional materials that are effective are frequently distributed in the format of concise videos, audio files, and digital handbooks. According to Weimann, the multimedia format that is easily accessible in multiple languages has transformed into a "terrorist university," serving as a platform for jihadists to acquire new tactics and skills that enhance their efficiency in carrying out attacks. Undoubtedly, a plethora of information is available on the internet regarding the process of joining terrorist organizations, devising and implementing terrorist operations, and fabricating explosive weaponry and other armaments (as depicted in Fig. 1). The Internet's interactive nature enables individuals worldwide to establish connections and create networks for the purpose of exchanging strategies and tactics.



Fig. 1 A tweet of a suspected Daesh account, who refers to the “amazing” military infographics of Al-Naba’

4.2 Cyber-Planning and Cyber-Execution

Over the past few decades, a vast majority of terrorist attacks have incorporated the utilization of the Internet and its associated advantages [22]. The orchestration of a terrorist attack commonly entails inter-party communication across extended geographical distances and employment of advanced techniques. The employment of anonymizing communication tools by terrorists for the purpose of planning attacks has advanced to a heightened degree of complexity. Data encryption tools and identity-masking software pose challenges in identifying the sender, recipient, and overall content of information [23]. The Internet can also serve as a means to initiate measures towards identifying potential targets and devising optimal strategies for carrying out attacks. Instances of pre-attack planning encompass procuring intelligence from publicly available sources and other channels, as well as obtaining guidance on the execution of a specific form of aggression. There is a possibility that terrorists could utilize the internet as a means to execute attacks that incorporate the aforementioned categories. The possibility of violence, particularly when firearms are involved, has the capacity to elicit extensive apprehension, fear, and alarm within

a community or a specific group. The acquisition of necessary tools for executing an attack may be facilitated by the prevalence of online activity. Online marketplaces can serve as a means for terrorists to procure the necessary components and services required to execute their attacks. The utilization of compromised security in credit cards or other electronic payment methods may facilitate the acquisition of such purchases [24].

4.3 Financing and Fundraising

The utilization of the Internet by terrorist groups and their sympathizers for the purpose of raising funds is a plausible occurrence. The financing of terrorism pertains to a broad range of activities that have been identified by the Financial Action Task Force (FATF) as means by which terrorist groups “raise, move, and use funds” [25].

The significance of this comprehension lies in its recognition that the financing of terrorism extends beyond the means by which organizations receive support. Instead, it includes a significantly broader spectrum of endeavours. A terrorist group necessitates diverse resources not only to fund its activities, but also to sustain its survival and expand its organizational capacity. Financial resources are necessary for a multitude of purposes, including but not limited to the provision of support for militants and their kin, the funding of travel-related expenditures, the recruitment and training of fresh members, the procurement of armaments and secure accommodations, the execution of operations, and the dissemination of the group’s ideology through social events or propaganda. The acquisition of financial resources is a critical factor in the success and sustainability of terrorist organizations, particularly during their formative stages when such resources are necessary to facilitate recruitment and garner support, as well as to establish significant material capacities [26].

4.4 Recruiting New Members

A form of cyber intelligence activity that has the potential to cause disruption, and is commonly employed by violent radical political factions, involves the dissemination of online propaganda with the aim of attracting a large audience and recruiting new members. Terrorist recruitment-oriented communication is typically crafted with the aim of appealing to vulnerable and marginalized subpopulations within a given society [27].

Thus, the efficacy of this propaganda with regards to enlistment and radicalization is contingent upon the presence of emotions such as injustice, alienation, or humiliation within an individual. Interactive involvement has the potential to cultivate a sense of unity and belonging, thereby establishing a virtual community between prospective terrorist recruits and current members of the terrorist organization.

4.5 Online Propaganda

Since 1984, jihadi media has been disseminated in four distinct phases, each of which corresponds to a novel medium for the distribution of information [28].

During Phase 1, which commenced in 1984, the dissemination of propaganda relied on conventional modes of communication such as oral and written means, including pamphlets, essays, printed magazines, and newsletters, which were disseminated by the mujahidin.

During Phase 2, which commenced in the mid-1990s, the internet was utilized for the development of jihadist websites that were created in a top-down manner.

Phase 3, which commenced in the mid-2000s, is characterized by the advent of virtual discussion forums where administrators disseminate and endorse information on behalf of jihadi factions, even in the absence of any established affiliations between these groups.

Phase 4, which commenced in the late 2000s, witnessed a significant shift in the utilization of social media platforms for jihadist online propaganda following the disintegration of the Syrian revolution in 2011. The utilization of social media platforms such as Facebook, Twitter, YouTube, and blogs by individuals has experienced a surge in recent times.

Terrorist organizations acknowledge the significance of online media platforms in their strategic planning, as evidenced by their endeavours to adopt the media prerequisite. The struggle for dominance over the *Umma's* hearts and minds has shifted towards a more prominent role for the war of narratives, as opposed to the conventional use of firearms and other weapons. The exploitation of the Internet by terrorists for the dissemination of their messages can yield several advantages, including the ability to maintain anonymity, the capacity to efficiently and expeditiously reach a broad spectrum of individuals, and the potential to foster discourse amongst their adherents.

5 A Brief Introduction to Online Counterterrorism

The term *counterterrorism* pertains to the governmental approach of thwarting and ultimately addressing terrorism through the amalgamation of theoretical and practical frameworks, strategies, and specific military methodologies (such as the employment of specialized forces). Likewise, counterterrorism methodologies may exhibit diversity in structure, akin to the multifarious nature of terrorist strategies.

The formulation of a counter-terrorism plan that is effective is predicated on the policy mandate of a government. Governments have the option of deploying armed troops, police, or paramilitary units, including intelligence personnel, to address the threat posed by both domestic and international terrorists.

The scope and magnitude of a state's military, paramilitary, and law enforcement activities exhibit significant variation. Terrorist disruption operations encompass a

range of tactics, such as surveillance, propaganda dissemination, arrest and detention, encroachment into areas frequented by terrorists, as well as large-scale troop deployments, targeted killings, and invasions of territories with high concentrations of terrorists. In certain circumstances, the utilization of force becomes necessary, and when it is employed, it is imperative that it be executed with the utmost precision and sensitivity. It is imperative for all stakeholders to adhere to the principles of international law, however, it is crucial to take into account the operational protocols that are unique to warfare.

Although terrorists have devised diverse techniques to utilize emerging technologies for unlawful activities, their exploitation of the internet creates motivations for intelligence gathering and other counterterrorism measures aimed at thwarting and combating terrorist acts. Virtual communication platforms such as websites and messaging applications serve as a substantial means of obtaining information pertaining to the operations, activities, authors, and at times, even the targets of terrorist organizations.

The heightened utilization of the internet by terrorist entities has led to an escalation in the accessibility of digital information that can be gathered and scrutinized for the purpose of countering terrorist operations. Undoubtedly, a conspicuous instance of governmental ingenuity in utilizing technology to counter terrorism is the emergence of extensive data collection and data analysis to facilitate the handling of information for inquiries stemming from the anti-terrorism campaign.

Various entities such as law enforcement, intelligence, and other organizations are continually devising novel approaches to proactively forestall, detect, and discourage terrorist actions on the Internet. Conventional investigative methods are gaining significance due to their effectiveness in promptly detecting terrorist menaces.

In recent decades, artificial intelligence (AI) has garnered significant attention as a means of analysing vast quantities of data in a highly efficient and effective manner. This is achieved by enabling the identification of concealed patterns and connections. The advancements in AI technology have resulted in its widespread utilization by various entities such as governments, corporations, and individuals, serving a broad spectrum of objectives ranging from financial management to extensive monitoring. The benefits of this extend to the domain of counter-terrorism as well. Consequently, there is a growing interest among law enforcement and counter-terrorism agencies worldwide in exploring the possibilities of harnessing the transformative capabilities of artificial intelligence.

The advanced AI capability is not limited to publicly accessible sources. AI systems possess the ability to effectively and lucratively analyse posts and communications on the Dark Web. Certain artificial intelligence programs have demonstrated notable proficiency in detecting buying habits on online platforms, utilizing parameters that encompass the types and quantities of items falling under categories frequently associated with terrorist activities. These often encompass weaponry, ammunition, and associated equipment, as well as transportation, chemicals with potential for use in explosive manufacturing, and commodities linked to human trafficking (which may serve as a source of terrorist financing) [29].

Within the realm of counterterrorism, the term terrorism informatics pertains to the utilization of diverse techniques for gathering and scrutinizing data from a vast array of origins, with the aim of providing intelligence analysts with practical and applicable information [30].

In order to achieve this objective, it is imperative to incorporate various fields such as computer science, business computing, statistics, machine learning, computational linguistics, and social network analysis. Data analysis and data mining are two prevalent techniques utilized to generate the necessary analysis for providing valuable insights. The data mining framework for anti-terrorism work comprises several constituent elements, including issue definition, data management, data mining, model assessment, and visualization or reporting of the results obtained.

Notwithstanding the widespread recognition of AI's predictive capabilities in the domain of counterterrorism, its implementation has been constrained. This assertion acknowledges the limitations of AI in terms of its ability to promptly and accurately address complex inquiries and generate precise prognostications.

5.1 Identifying Daesh-Related Propaganda with OSINT Tools

It is imperative that governments do not confine their efforts solely to the exploration of innovative approaches for acquiring and incorporating intelligence. It is crucial for individuals to engage in the critical examination of gathered intelligence to provide actionable insights in their endeavours to counteract acts of terrorism. By scrutinizing a restricted number of sources that hold noteworthy data, it is plausible to obtain a considerable volume of information concerning the adversary's capabilities and efficacy in diverse domains, along with their geographic location and strategies. The utilization of cyberspace enables Daesh to access a worldwide audience, which has the potential to expand their sway and the extent of their activities [31].

In the realm of digital enforcement, alongside with traditional intelligence gathering techniques, the designations SOCMINT (Social Media Intelligence) and OSINT (Open Source Intelligence) are at times employed interchangeably to amass relevant information that is pivotal in affording supplementary perspectives on specific hazards.

The online propaganda of Daesh is a highly complex operation that employs a range of platforms such as social media, forums, video-sharing sites, and encrypted messaging applications. The digital strategy employed by the group typically entails the distribution of professionally crafted video content, online publications, and infographics that convey messages with the intention of fomenting violence, radicalizing individuals, and attracting fresh recruits. The central motifs frequently focus on the establishment of religious credibility, portrayal of triumph in warfare, exhibition of administrative proficiencies, and accentuation of perceived adversaries.

Through the utilization of Open-Source Intelligence methodologies and tools, it is feasible to monitor and mitigate the dissemination of Daesh propaganda in the digital realm. By means of OSINT, researchers can discern the origins, scope, and modalities of jihadist propaganda.

Google Dorks are frequently utilized techniques for obtaining data from the public domain. The term refers to the utilization of sophisticated search techniques within the Google Search Engine for the purpose of locating specific information. The utilization of search operators is a prevalent technique in the process of enhancing search queries and limiting the extent of search outcomes.

For our research we have decided to use two similar investigation queries:

1. **intext:**“صحيفة النبا العدد” AND “**magazine**”: this search operator returns results where the expression “صحيفة النبا العدد” (which means “Al-Naba’ Newspaper Issue”) and the term “magazine” appear in the content of the webpage.
2. “**magazine**” AND “صحيفة النبا العدد”: the search engine will conduct an exact match pursuit for the text enclosed within quotation marks, without any alterations or modifications.

As a result, we have identified two noteworthy online platforms that disseminate propaganda affiliated with Daesh, as presented in the following Figures (Figs. 2, 3, 4, and 5).

As soon as we discovered these intriguing websites that might be connected to Daesh, we began to study them by taking a broad overview of the dashboards and the information they contained.

- The *I'lam Foundation* keeps a wide range of media content in numerous languages. The website also features constantly updated translations of fresh materials and offers access to an archive of previous print and video media items produced by Daesh. Most of the funds donated to the media center come from jihadists

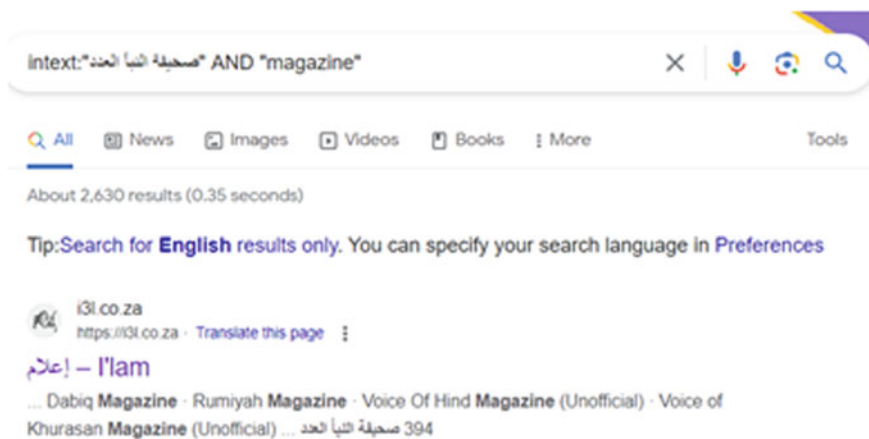


Fig. 2 First search query and result

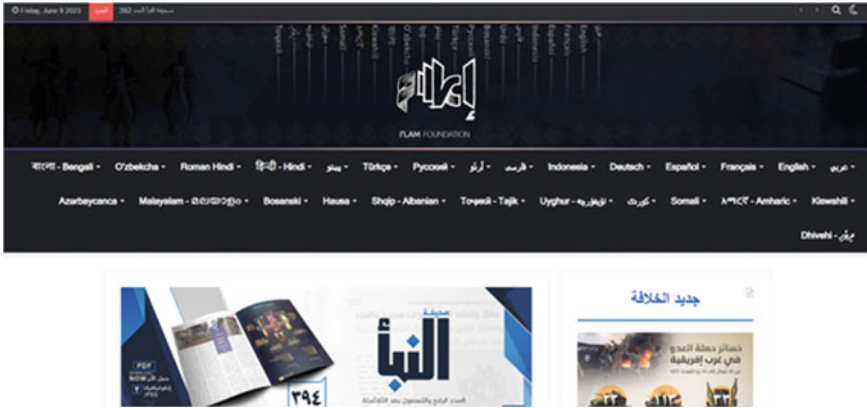


Fig. 3 I'lam Foundation in <https://i3l.co.za/>

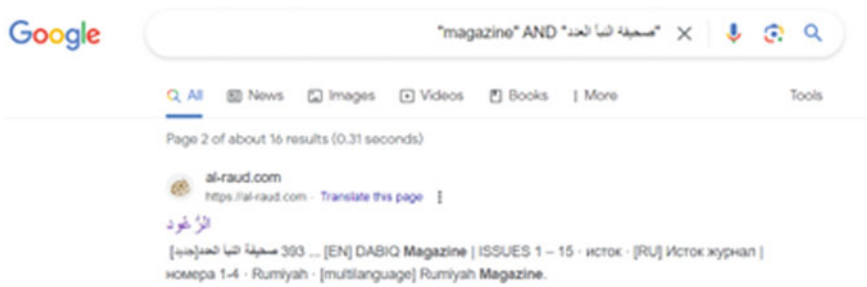


Fig. 4 Second search query and result



Fig. 5 Al-Raud in <https://al-raud.com/>

in western nations who support them with cryptocurrency. *I'lam* stands out by establishing an active dark web channel to serve as a hand-to-track backup if its surface websites go down and setting up a crypto wallet to accept donations anonymously. By visiting the provided URL or scanning the provided QRcode, it is possible to surface the website on Tor. The *I'lam* structure comprises a Telegram channel (accessible via the following URL: https://t.me/I3_3333) that is only dedicated to communication and dis-semination of English contents produced by the *Central Diwan of Media*.

- In contrast, *Al-Raud* maintains an extensive assortment of media materials, including functioning as a video streaming service, that are mainly published in the Arabic language. Supporters of Daesh additionally pro-mote the utilization of Virtual Private Networks (VPNs), proxies, and the TOR browser to access propaganda and facilitate donations.

Afterwards, the utilization of OSINT search engine methodologies for Domain Name analysis was carried out:

1. **Similarweb**, an analytics tool that offers comprehensive insights into website or mobile ranking, performance, source traffic, and other related metrics.
 - The global ranking of *I'lam* is 637,542, with a country ranking of 28,647 specifically in Algeria. A rise in the website's ranking was observed during the period spanning from March to April 2023. During the period extending from March 2023 to May 2023, Islam recorded an overall sum of 70,101 visits globally, with 29,775 of those visits being unique. During the past three months, most visitors were male, accounting for 62.19% of the total visitors. The age distribution indicates that the age group between 25 and 34 accounts for 31.26% of the total, followed by the age group between 18 and 24 at 23.53%, and the age group between 35 and 44 at 19.39%. The age group of 65+ represents only 5.47% of the total. The mobile device was the most frequently utilized means of accessing this website, accounting for 79% of consultations.
 - The global ranking of *Al-Raud* is 669,415, with a country ranking of 465,775 specifically in France. During the period spanning from March to May of 2023, *Al-Raud* garnered a cumulative sum of 23,340 visits on a global scale, with 11,998 of these visits being attributed to unique visitors. The mobile device was the most frequently utilized means of accessing this website, accounting for 93.52% of all consultations. A significant surge emerged from April to May. Iraq ac-counted for the majority of traffic share at 68.20%, with Germany and France following at 26.05% and 2.36%, respectively.
2. **Urlvoid**, which aids in the identification of possibly malicious websites. Additionally, it provides more details about the domain (IP address, DNS records, etc.) and compares it to established blacklists.

- The report summary for <https://i3l.co.za/> provided the following important information:
 - IP Address: 188.114.97.2
 - Server Location: Netherlands
 - Latitude/Longitude: 52.3716/4.8883
 - City: Amsterdam
 - Region: North Holland
- The report summary for <https://al-raud.com/> provided the following important information:
 - IP Address: 172.67.134.1
 - Server Location: United States
 - Latitude/Longitude: 37.751/−97.822

It is important to note that the actions mentioned above, while impactful, are best carried out by specialists with a thorough understanding of OSINT tools and the legal and ethical considerations involved. Inappropriate utilization of the aforementioned tools may lead to unintended outcomes, and even well-intentioned actions may sometimes produce unfavourable consequences. Hence, it is crucial to implement responsible and ethical approaches while employing OSINT to counteract jihadist propaganda.

5.2 Predictive Analytics and Online Propaganda

A fundamental comprehension of Artificial Intelligence (AI) and Machine Learning (ML) tools is necessary for law enforcement and counterterrorism entities to effectively evaluate its potential applications in the fight against online terrorism. Indeed, automated data analytics are being used to boost intelligence operations and security agencies, specifically through the implementation of data visualization techniques. Data can be automatically gathered from online forum or other social media platforms, and archived for subsequent examination with the aim of uncovering correlations and associations that may expose terrorist organizations or questionable behaviours.

Community detection is a fundamental aspect of network analysis that aims to reveal the latent structures of networks by partitioning them into smaller sub-networks. The investigation of community detection has been subject to extensive research and has been implemented in various practical network-related matters. Conventional approaches to community detection involve the deduction of community structures through the utilization of probabilistic graphical models and other pre-existing forms of information. Recent advancements in the field of network analysis have introduced novel techniques that leverage deep learning to transform complex networked data into a low-dimensional representation. These techniques have been developed to address the challenges posed by the increasingly intricate issues that

network methods aim to tackle, and the complex nature of the network data being analysed. Despite the recent advancements in network analysis, the absence of a comprehensive understanding of the theoretical and methodological underpinnings of community discovery is likely to impede future progress in this field.

Upon acquiring this knowledge, law enforcement and counterterrorism entities will be better equipped to comprehend the potential applications of AI and ML in identifying and mitigating virtual terrorism.

In the domain of Machine Learning, the systematic use of algorithms facilitates the extraction of underlying relationships within data and information. These algorithms, when used in conjunction with other tools, contribute to the construction of a machine learning model. This model employs mathematical operations and a designated set of example inputs, known as a training dataset, to generate predictions and assessments. This approach allows for the mapping of a range of outputs based on a given input, where the mapping is determined by a predefined set of rules or conditions. The primary aim is to develop significant models that can learn and perform tasks autonomously, without relying so much on human intervention or input. The efficacy of predictive analytics is enhanced by the presence of high-quality data.

6 Methodologies and Tools of Research

This section is going to describe how our experimental research was developed and the main tools used.

The first stage was collecting search documents in some ways linked to the scope of my work.

The next phase was to collect relevant data, after which we would have a better idea of which large studies would provide support for our research themes. Here, it was crucial to avoid distorting the results by including many papers with duplicate data in the synthesis of a systematic review. With the datasets collected then we trained and tested our Machine Learning Model.

6.1 Aim of Our Study

The objective of our study was to investigate the possibility of employing Machine Learning (ML) and Natural Language Processing (NLP) techniques for the purpose of analysing Jihadist narratives, with the aim of identifying potential commonalities across various propaganda sources. One of the objectives was to assess the existence of tweets that exhibit a direct correlation with *Al-Naba'* magazine. The magnitude of propaganda disseminated by Daesh is of such great proportions that its analysis by human agents is rendered nearly infeasible. Consequently, the development of methodologies and protocols that can be employed to scrutinize vast quantities of data

represents a pivotal obstacle. To effectively develop counter-narratives and counter-messaging strategies, it is imperative to possess an in-depth understanding of the diverse taxonomies employed in propaganda, the nuances in the narrative across media platforms, and the evolution of the narrative among users. The main objectives were pursued through the application of *Social Network Analysis* and Data Mining, with a particular focus on clustering algorithms. Ultimately, our data investigation focused on Tweets sourced from individuals who may be aligned with or supportive of Daesh.

6.2 Research Objectives

- i. Examining Daesh propaganda narrative and the most common taxonomies in *Al-Naba'* magazine.
- ii. Identifying Daesh affiliations to *Al-Naba'* through social media analysis of Twitter using Data Mining Techniques, in particular clustering.

6.3 Justification

The objective of this research was to construct a model that could facilitate the identification of *Al-Naba'* Daesh association by utilizing clustering algorithms. Consequently, the model that has been put forward would provide a structured approach to detecting covert jihadist propaganda with minimal human involvement. This endeavor aims to enhance the efficacy of security mechanisms by implementing a more accurate and expeditious method of detection compared to prior approaches.

6.4 Systematic Literature Review

The goal of my systematic review, conducted using an objective search strategy, was to locate as many of the relevant publications as possible that could address our research questions.

6.4.1 Questions to be Addressed

The literature search needed to address the following questions:

- i. To what degree has the scholarly community investigated and recorded the phenomenon of online extremism, radicalism, Islamism, and jihadism in academic publications?

- ii. What are the prevailing patterns in identifying online extremism, radicalism, Islamism, and jihadism, and how do these techniques compare to each other?
- iii. What categorization methods have been employed to address the issue of identifying and labelling instances of radicalism, Islamism, jihadism, and other forms of online extremism?
- iv. How did the authors validate the precision of the data employed in identifying and classifying online extremists, radicals, Islamists, and jihadists?

6.4.2 Keywords and Queries

Because no one source has all the main research, we had to rely on a variety of electronic sites. The criteria for selecting papers are meant to find those primary studies that offer direct evidence about the research question. For this reason, a selection criterion—based on a specific query—was applied. Keywords used to create the query were: ‘radicalization;’ ‘extremism;’ ‘jihadism;’ ‘Islamic State;’ ‘IS;’ ‘propaganda;’ ‘magazine;’ ‘Al-Naba;’ ‘social network;’ ‘Twitter;’ ‘detection;’ ‘classification;’ ‘network analysis;’ ‘machine learning;’ ‘clustering.’

6.4.3 Keywords and Queries

After collecting relevant primary studies, they had to be evaluated for their actual relevance. Inclusion and exclusion criteria were applied in order to narrow the scope of my analysis. As a result, in addition to community identification, only studies with extremism detection and classification employing computer approaches such as Network Analysis, Clustering, and Machine Learning were considered and examined.

The “quality” of main research needed to be evaluated alongside more standard inclusion/exclusion criteria. At this stage, we looked at more restrictive inclusion/exclusion criteria:

- **Extremism/Radicalism/Islamism/Jihadism:** Research should focus on either detection, categorization, or databases.
- **Types of Extremism (Online Radicalization, Online Recruitment, Online Propaganda):** Research must focus on radicalization, recruitment, and propaganda carried out via online sources such social media, websites, blogs, and other digital platforms, as these represent the fastest-growing forms of extremism today.
- **Detection and Community Identification:** Research efforts should be on figuring out how to use digital media and other methods to spot extremism and to pinpoint specific populations.
- **Datasets and Classification Algorithms:** Methodological approaches to the classification of datasets were a focus, as were the means by which such datasets may be collected or constructed.
- **Information Validation:** A number of different methods for validating the quality of the data were used in the study.

6.5 Literature Review

6.5.1 Community Detection

Krebs used social network analysis for the first time to the study of terrorism, and their findings were ground-breaking [32]. He tried to piece together the ties between the 9/11 hijackers and terrorists using unstructured data collected from free access sources like newspapers. Though shallow, the study did set the stage for more in-depth investigations of terrorism from this perspective.

Several researchers—including Koschade [4], and Belli et al. [33] and Brams et al. [3], used relational data on individuals to reconstruct terrorist networks and investigate important roles and actors. Conversely, the rise of social media has made available a treasure collection of data on how individuals interact with one another online. Academics have taken an interest in the effects of cybercrime and terrorism since they are as widespread online as they are offline. Specifically, there is new research focusing on the possibility of using social media network data to spot signs of radicalization or terrorist activity. Social media platforms enable us to derive the patterns and dynamics of extreme users by including geographical, chronological, and a great many additional profile elements beyond basic connection information.

Even though there has been some forward movement in the use of network approaches to model terrorism and in the testing and experimentation with techniques to address a wide range of academic research, it is essential to acknowledge that this forward movement has not resulted in the anticipated conjunction of network science with unsupervised learning and, more specifically, *Cluster Analysis*.

The study of Chenoweth and Lowham [34], which made use of information on terrorist organizations that plainly targeted American people, was one of the first efforts to use cluster analysis to the classification of terrorist organisations.

Using a multi-membership clustering method and similarity scores, Qi et al. categorized extremist websites into hierarchical groups [35]. The objectives of this technique were effectively accomplished, thanks in large part to the incorporation of aspects drawn from social network research as well as unsupervised learning. Group Profiling Automation for Crime and Terrorism (GPACT) is a methodology that was developed by Lautenschlager et al. [5] to construct profiles of criminal and terrorist organizations. Incorporating clustering of terrorist incidents into a multi-staged approach to classifying terrorist groups allowed for the rapid production of rich profiles of relevance and value to an analyst.

Furthermore, some experts believe that Social Media Intelligence (SOCMINT) will become more important in the future as a tool for network science-based counterterrorism. Here we have a fundamentally new challenge in the field of network science applied to counterterrorism. To put it simply, we now utilize networks to obtain insight into people as opposed to the other way around (where information about individuals was used to construct networks). More often than not, we are looking for a tiny, maybe hidden community inside a bigger network. As a result, new methods that can identify hidden networks inside social media are needed.

Many scholars have looked at the issue of community identification as it pertains to massive social networks [36]. Algorithms designed to find groups in networks look for clusters of nodes that are more closely interconnected than the rest of the network [37].

However, social networks collected from social media offer special issues owing to their scale and strong clustering coefficients. Additionally, it is well known that various sorts of associations may be represented by different types of ties in social media platforms such as Twitter [38].

In the field of network research, Newman's [39] and Blondel's [40] approaches are considered to be the gold standard for group and clustering detection. Their contributions include a far wider range of approaches. Algorithms for group optimization often divide the graph into k groups, where k is a variable that depends on the objective function being solved. Unexpectedly, Newman and Blondel both interpret this minimizing issue as a maximization problem by maximizing modularity.

Additional methods are needed to accomplish this objective. Over the course of the last several years, a new category of group identification algorithms has emerged as a result of the development of methods for finding communities across annotated networks. This research project's objective is to adequately include node-level information into clustering algorithms in order to find a solution to the problem of noise that is inherent to social media networks. Graph vertices are embedded in a vector space where pairwise, Euclidean distances may be determined using vertex clustering, which has its roots in conventional data clustering techniques. These techniques incorporate modern eigenspace graph representations with more conventional data clustering and classification techniques like *K-means* and *hierarchical agglomerative clustering*. The researcher has a lot of leeway in choosing the information to employ while working with these approaches. Vertex-based grouping and classification algorithms have been shown to perform well with social media because they take into account a wide range of vertex qualities, such as user account attributes, while still making use of the information intrinsic to the network [41]. By using vertex clustering with spectrally representing network characteristics and matched user account information, the study of Wang et al. [42] offers the *SocioDim* approach for detecting groups embedded in social media. The approaches given by Binkiewicz et al. [43] are essentially the same. To this binary split of the network, Tang et al. [41] article use *SocioDim* to classification. As Miller et al. [44] example shows, there is reason to believe that these techniques will be effective for detecting hidden networks on social media. Data shows that eigenspace approaches may accurately depict multiplex representations of different sorts of social links in social media, and preliminary research using simulated networks suggests that these methods may perform well in social media threat detection. When combined with user account data and node level information, eigenspace models of the multiplex social networks inherent in social media may prove to be a more successful tool for detecting extremist organizations.

6.5.2 Jihadists Activities on Twitter

Our empirical research started from the collection of the most relevant issues of *Al-Naba'* newspaper, published on a regular basis by both the Jihadists websites we had found.

Expert remarks and assessments on individual or collective jihadists, as well as the current situation of jihadist organizations and material on Twitter, are examples of qualitative descriptions. Quantitative insights provide context for the existence, activity, and interactions of the jihadist community on Twitter via the use of data and graphical representations. The purpose of most social network analysis conducted on jihadists is to either locate the most significant or key accounts in the network, or to determine the flows of information between nodes in the network.

Klausen investigated the network of 59 Twitter accounts of militants of Western origin alleged to be in Syria between January and March 2014 to get insight into the flow of information and the extent to which access and content of communications are restricted. According to the research, “sponsor” accounts linked to terrorist groups in insurgency zones are playing an increasing role, as are European-based organizational accounts affiliated with the proscribed Saudi Arabia-based militant network *Al-Muhajiroun*, which was banned by the British government in August 2005 [45]. Using Twitter data from September 2014 through December 2015, Berger and Morgan [6] constructed a profile of Daesh sympathizers. They suggested a process for tracking down and categorizing relevant Daesh accounts. They also investigated the impact of Twitter suspensions on ISIS supporter accounts and the spread of Jihadist propaganda [6]. Between June and October of 2015, a Twitter user going by the handle *Baqiya Shoutout* compiled a list of English-speaking accounts that supported Daesh. Berger and Perez [46] compiled and evaluated this list. Social network analysis revealed that the English-speaking Daesh community is relatively closed off. Suspending jihadist Twitter accounts seemed to have a severe impact, as seen by the precipitous drop in the number of accounts and the volume of pro-Daesh material in the networks. Particularly, the number of people who followed individuals who kept making new accounts after being banned from the service declined [46].

Bodine-Baron et al. found that the usage of the Arabic name for the group vs the abbreviation “ISIS” was a significant identifier of ISIS supporters and opponents (Daesh). According to the lexical analysis, the regular users of Daesh tend to post material that is very hostile of Daesh, whereas the regular users of “The Islamic State” tend to post content that glorifies the group. Jihadists detractors exceeded its supporters by a factor of six to one, yet the latter consistently out tweeted the former. Daesh mujahideen followers, and Sunni Muslims were identified via social network analysis as four metacommunities sharing common identities and content themes in discussions about Daesh on Twitter. Central communities inside the metacommunities and their interrelationship with one another became the focus of more research [47].

Important work that may be done to better understand jihadist user behaviour and uncover new potential jihadist accounts has to do with the characterization of jihadist accounts, or the identification of the qualities that distinguish jihadist Twitter

accounts. Magdy et al. [7] found that the tweeting patterns of accounts that supported or opposed Daesh corresponded to key news or events occurring toward the time the data was gathered. To do this, they examined the Arabic tweets to see if they included the complete term, such as Al-dawla Al-islamiya (“The Islamic State”), or the abbreviation “Islamic State.” After examining factors like hashtag usage and chronological patterns, they came to the conclusion that Daesh followers had joined Twitter to voice their support for the group. Further, a classifier was developed to determine if a user’s pre-Daesh tweets would indicate support for or opposition to Jihadists [7].

Supporters of terrorist organizations and those who share their propaganda on Twitter may be identified with the use of a machine learning model created by Kaati et al. [8]. The model was trained using data-dependent parameters like the most common phrases and hashtags and data-independent variables like the frequency of word length, characters, and emotion words. They demonstrated that their model achieves a high degree of accuracy while processing tweets in English but underperformed when processing tweets in Arabic [7].

For their 2016 study, Rowe and Saif [9] analysed the Twitter activity of Europeans before and after they began using pro-Daesh language and distributing propaganda. They presented strategies for determining if a user is engaged by analysing their material sharing habits and the pro- and anti-Daesh rhetoric they use. As an added perk, they looked at how users’ language and material sharing habits, as well as their interactions with others, changed before and after activation. Finally, they provided evidence that the degree of social cohesiveness in individuals’ respective Twitter networks is a significant predictor of whether they would adopt radicalized users’ pro-Daesh conduct [9].

6.5.3 Detecting Online Extremism

According to Chatfield et al., extremists use Twitter for propaganda, radicalization, and recruiting. The authors identified many sources that all agree that the Twitter account @shamiwitness is a proponent of the terrorist group Daesh. Propaganda, radicalization, and recruitment-related tweets were manually separated out. The communication network was analysed using multi-sided network graphs (showing interactions between @shamiwitness and other Twitter users). Thus, several subgraphs of interactions between @shamiwitness and other users emerged. Tweets were then manually sorted into three distinct categories based on their use of keywords, phrases, hashtags, and allusions to religion into the three main extremist communications [48].

Machine learning has surpassed other methods for automatically detecting online extremism since it does not involve any explicit programming for pattern identification. Xu and Lu applied a seed strategy to acquire more data for extremism identification [49]. Twitter may be mined for relevant information by searching using hashtags connected to certain beliefs. The total number of accounts used by Daesh members and others who are sympathetic to the group is 4,820. In the course of the examination into extremism, factors such as hashtag tokens, predicted hits, and harmonic closeness were taken into consideration. In the present

investigation, a method that is based on graphs was used to extract characteristics such as striking time prediction and harmonic proximity. Within the context of this inquiry, a machine learning-based method is utilized to carry out automated extremism detection. Adaboost is the name of the algorithm that the research team developed in order to automatically detect and classify users of candidates' social media accounts as either sympathizer or users who do not support Daesh.

Hashtags are used by Ashcroft et al. to gather and identify Twitter accounts that support Daesh [50]. In all, 6,729 user accounts were found throughout the research. Stylometric characteristics, temporal features, and emotion features are among those taken into account. Words with high frequency in the sample are used as stylometric characteristics. Hour of the day, time of day, day of the week, and day of the month are all examples of temporal characteristics. Extremely negative, negative, neutral, positive, and very positive labels are used to categorize sentiment-based characteristics. In this research, an ML-based method was employed to detect extremist content automatically. The Support Vector Machine, Naive Bayes, and Adaboost algorithms are used by the research community to identify extremist behaviour. Adaboost method, based on machine learning techniques, was then applied by many experts to automatically identify online extremism.

Kaati et al. [8] were successful in gathering information on ISIS by conducting an investigation of the hashtags that were being utilized. The tweets of around 6,729 people who have been designated as known ISIS supporters were collected. Both traits that are dependent on the data and those that are independent of the data are extracted by the authors. Examples of data-dependent features include phrases that are used frequently, hashtags that are used frequently, word bigrams, letter bigrams, and so on. Words that are able to convey feelings are another example of a feature that is not dependent on the data. Additional examples include temporal elements and characteristics of the stylistic presentation.

6.6 Research Phases [51]

6.6.1 Data Collection

In this section we outline how we collected data and provide an overview of our datasets. To ensure that primary research data is accurately recorded, this phase involves the creation of data extraction forms.

To reduce the likelihood of bias, data extraction forms should be created and piloted simultaneously with the study methodology. It is critical to avoid include several publications of the same data in a systematic review synthesis since duplicate reports would significantly bias any conclusions. That is why we didn't give them any thought. Since these are not data from an ongoing study, it is especially important to provide accurate details and, in certain situations, get the approval of the researchers. And there's always the chance that reports leave out crucial information. It is possible that they are poorly phrased and do not convey the intended idea. Once more, it is recommended that the necessary information be obtained by contacting the authors

themselves. Finally, only freely available and publicly accessible datasets from prior studies were assessed and analysed.

In this research and in our methodological approach, we have considered different digital sources of Daesh-related material and other extremist ideologies: tweets, magazines, websites, and blogs. The extent of the proposed automated classification aids in the detection of radical groups (and Daesh in particular), especially when these groups use the Internet for operations such as radicalization and propaganda. In general, this is a more expeditious and proactive strategy to dealing with the virtual Jihad's danger to national security.

However, several difficulties arise in this setting. Indeed, in order to keep ahead of Western surveillance, Jihadi warriors have been instructed to scrub all identifying information and metadata from their Twitter profiles, including their names, locations, and pictures [52]. Furthermore, efforts to restrict accounts on social media that promote terrorism meet another obstacle, which is the fact that when one account is blocked, users may simply establish another, and then another, and then another account. As a result, a proactive and coordinated response is required.

Data were collected for our investigation from several websites, forums, and online magazines. Daesh propaganda publications are released often and in a variety of languages. The journals are editorially crafted, complete with high-quality photographs and a well-designed layout, and are published in a number of different languages, English among them. This study uses the English versions of Dabiq (all the fifteen issues) and Rumiya (all the seven issues), taken from Kaggle dataset entitled "Religious Texts Used By ISIS" [53].

Our empirical research started from the collection of the most relevant issues of *Al-Naba'* newspaper, published on a regular basis by both the Jihadists websites we had found.

Al-Naba' features a variety of articles and material types, such as news, commentary, visualizations, religious writings (including fatwas), and advertisements for various forms of media output. So, it totally reflects Daesh mindset and mirror events occurring on the ground, and social issues particularly relevant to the residents of the self-proclaimed Islamic State. The articles were selected for their relevance as being particularly under the attention of the Western world's monitoring and inspection instruments (Table 1). The reason we analyzed *Al-Naba'* is because no prior studies have been conducted on it. Several articles of each issue were then selected, and their translation was refined.

6.6.2 An Overview of Al-Naba' Text Classification

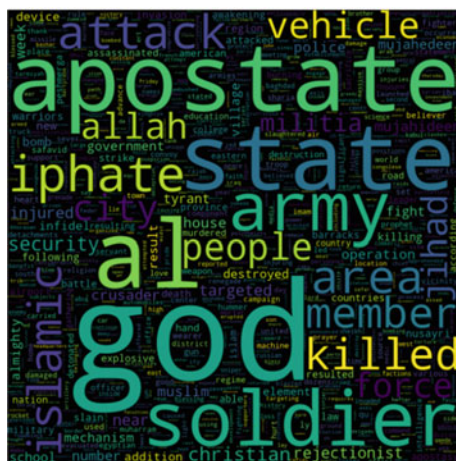
Our analysis started with the identification of the most prevalent terms present in our propaganda texts (Fig. 6). The word cloud technique is widely recognized as a highly effective and uncomplicated means of visualizing textual data. The size of words is contingent upon their frequency of occurrence, which is indicative of their significance within the given context.

The subsequent stage involved the identification of frequently occurring taxonomies to be extracted for our investigations. We covered Charles Winter's

Table 1 Selection of Al-Naba' magazine

Date (Gregorian)	Date (Islamic)	Issue
17 October 2015	3 Muharram 1437	# 1
24 October 2015	10 Muharram 1437	# 2
31 October 2015	17 Muharram 1437	# 3
7 November 2015	24 Muharram 1437	# 4
15 November 2015	2 Safar 1437	# 5
22 November 2015	9 Safar 1437	# 6
29 November 2015	16 Safar 1437	# 7
6 December 2015	23 Safar 1437	# 8
13 December 2015	30 Safar 1437	# 9
19 December 2015	7 Rabî Al-Awwal 1437	# 10
25 May 2018	10 Ramadan 1439	# 134
6 December 2019	8 Rabî Ath-Thâni 1441	# 211
12 March 2020	17 Rajab 1441	# 225
19 August 2021	10 Muharram 1443	# 300
26 August 2021	17 Muharram 1443	# 301
2 September 2021	24 Muharram 1443	# 302
10 September 2021	2 Safar 1443	# 303
4 November 2021	28 Rabî Al-Awwal 1443	# 311
31 December 2021	26 Jumâda Al-Awwal 1443	# 319
22 January 2022	18 Jumâda Ath-Thâni 1443	# 322
22 April 2022	20 Ramadan 1443	# 335
23 June 2022	23 Dhul-Qa'da 1443	# 344
8 July 2022	8 Dhul-hijja 1443	# 346

Fig. 6 Word cloud of the most frequent words in Al-Naba' dataset



study of the primary themes of jihadist propaganda, including other Victimhood, Utopia, and War, which reveal the organization’s tactics and influence. Specifically, *War* is identified as violence in general (armed attacks and acts of terror); *Utopia* as the idealization of the Proto-State; with Victimhood, by compiling evidence of the atrocities committed by Daesh’s adversaries, the group bolstered the justification for its role as a guardian force. ISIS has been able to cover up military losses and changing fortunes with religious ideals; it is aware that in order to make its warriors, supporters, and future recruits feel better about themselves and to encourage them to join, it must provide an explanation for and justify its losses. Winter’s categorization was used to identify the three primary groups to which each of the texts under consideration refer. The aforementioned statements were integrated into the *Cogito Intelligence API* of Expert.AI, which is an online semantic-based system that is specifically designed for crime and intelligence purposes. It includes specialized classification and offers complete semantic processing patterns such as *Categorization*, *Text Mining*, and *Fact Mining*.

Based on this assessment, subgroups were empirically established (Table 2) and then attributed to each text.

A dataset was then created with the following columns: Issue; Gregorian Date; Islamic Date; Text; Main Category; Sub-Category; Other Sub-Category, as shown in Fig. 7.

In the following lines (Table 3) we offer some examples of our classification considering only the columns ‘text,’ ‘main category,’ ‘sub-category,’ and ‘other sub-category.’

Table 2 Empirical categorization adopted in Al-Naba’ texts

War	Utopia	Victimhood
Act of crime against the West	Religiously inspired terrorism	Act of crime against the Caliphate
Armed attack	Legitimize the group	Offensive operation
Act of terror	Governance and state education	Martyrdom legitimization
Dar al-Islam versus Dar al-Harb		Counterterrorism

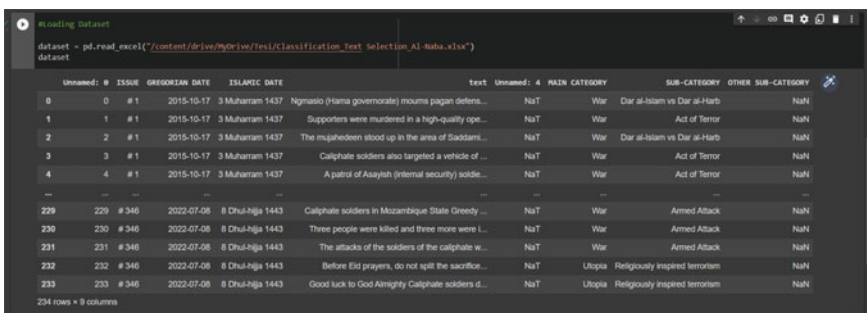


Fig. 7 An overview of Al-Naba classification dataset

Table 3 Few examples of the first Al-Naba' taxonomy classification

Text	Main category	Sub-category	Other sub-category
Caliphate soldiers in Mozambique State Greedy Christians killed five people and injured five more this week. They also assassinated an army member. Greedy Mozambican injured another and set fire to a barracks. With new attacks that occurred in five villages in the Cabo Delgado area	War	Armed attack	NaN
Several members of the Safavid Army and Shiite Popular Mobilization Forces were killed or injured, and some of their vehicles were destroyed, in a series of attacks on their positions in Arab Jabour by Caliphate forces using improvised explosive devices and light and medium guns. Among those targeted mechanisms was a rejectionist Popular Mobilization Forces (PMF) leader's mechanism, which resulted in his injury and the injury of his fellow members	War	Armed attack	NaN
Two explosive devices were detonated by Caliphate soldiers	War	Act of terror	NaN
Caliphate soldiers also targeted a vehicle of one of the Officials of the Popular Mobilization Forces (PMF) in the village of (Al-Tha'er) belonging to the area of (Tarmiyah) with an explosive device. So he was injured	War	Act of terror	NaN
Small meetings of Houthi polytheists in Sana'a state have been revealed	War	Dar al-Islam versus Dar al-Harb	NaN
23 crusaders and apostates were assassinated	War	Dar al-Islam versus Dar al-Harb	NaN
The Islamic State decides on an official statement regarding the attacks on the French capital (Paris) on Friday night, November 13th, which resulted in deaths and injuries	War	Act of crime against the West	NaN
There were over 500 crusaders. Instilling fear and panic in the minds of unbelievers	War	Act of crime against the West	Dar al-Islam versus Dar al-Harb
And those who push for deception are the Mujahideen Muslims	Utopia	Legitimize the group	NaN
Fighting these armies is a source of pride	Utopia	Legitimize the group	NaN

(continued)

Table 3 (continued)

Text	Main category	Sub-category	Other sub-category
On the topic of who it is required to combine judgements on people, "Sharia says that the house of disbelief, headed by unbelief's rules, is distinct from the house of Islam, which is followed by Islamic provisions and administers what Allah has revealed"	Utopia	Religiously inspired terrorism	NaN
But it is the Qur'an and the teachings of our Prophet al-Adnan that guide us (peace and blessings of Allaah be upon him). We shall battle and fight and fight until we find someone who is completely dedicated to God. This is God's will. We shall reject and renounce the apostates. And we oppose the infidels and reject them	Utopia	Religiously inspired terrorism	NaN
Christians are greatly tormented in their faith disorder, they have not settled on a God and have even deified three gods. It is a condition. It is a sign that their hearts are broken and their tongues are broken	Utopia	Religiously inspired terrorism	Dar al-Islam versus Dar al-Harb
But it is significant because of the circumstances that Islam and Jihad are facing in those countries. Apostates and Crusaders professed to be Christians	Utopia	Religiously inspired terrorism	Dar al-Islam versus Dar al-Harb
Establishment of a Sharia Sciences College in Anbar Province. The Islamic State has not shied away from creating the proper and effective foundations for brain development, healing, and eradication of any contaminants through attentiveness in the discipline of forensic science	Utopia	Governance and state education	NaN
The love of one's country is one sort of transgression, the destructive notion of "patriotism" that has arisen among Muslim sons	Utopia	Governance and state education	NaN
A failed American landing near (Hawija) and repelling Shiite advances into (Artis) and (Ajil)	Victimhood	Offensive operation	NaN
Two different operations targeted the mujahedeem in the Sadr al-Youssoufia area	Victimhood	Offensive operation	NaN
The French government issued an emergency declaration. French security personnel were deployed	Victimhood	Counterterrorism	NaN

(continued)

Table 3 (continued)

Text	Main category	Sub-category	Other sub-category
After years of compromise and deception, the crusading US Secretary of State, John Kerry, finally revealed the reality of the American vision for the future of the situation in the Levant, which boils down to the fact that the tyrant Bashar al-Assad remains in power despite killing and abandoning millions of Levant residents in a fierce war against them	Victimhood	Counterterrorism	NaN
The imbecile “Trump” announced the defeat of the Islamic State in eastern Afghanistan	Victimhood	Crime against the Caliphate	NaN
The Islamic State was founded in on a martyrdom attack	Victimhood	Crime against the Caliphate	NaN
The Caliphate’s warriors in Sinai had been slaughtered	Victimhood	Crime against the Caliphate	NaN
The martyr brother Abd al-Rahman al-Logari—may Allah accept him—managed to breach the security barriers surrounding Kabul airport and then indulge in the midst of a massive gathering of US troops, spies, and their accomplices	Victimhood	Martyrdom legitimization	NaN

Data Preparation and Pre-processing Text

The data must first be compiled, then the quality must be evaluated, and finally, if required, the data must be cleaned. Only after these steps can the analysis begin. Finding and fixing anomalies in the data that we utilize is an essential step in the process of improving the overall quality of the information that we make use of. Raw data are indeed transformed into a suitable format for experimental purposes, without misleading the classifier. To sum up the idea that a poor classifier is the result of using noisy, unrealistically biased, or low-quality data, the expression “garbage in, garbage out” is often used. The pre-processing phase may, at its most basic level, involve activities such as removing identical records, normalizing the data, removing unnecessary and redundant data, removing misspellings or invalid or incomplete data, converting words to lowercase, removing punctuation, special numbers or characters, stop words, Count Vectorization or Tokenizing, Stemming, or Lemmatizing, and so on:

```

import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('omw-1.4')
from nltk.stem import SnowballStemmer
from nltk.stem.wordnet import WordNetLemmatizer
import re

def preprocess(text, stem=False):
    stop_words = stopwords.words('english')
    stemmer = SnowballStemmer('english')
    lemmatizer = WordNetLemmatizer()

    text = text.lower() # lowercase

    text = re.sub(r'![!]+', '!', text)
    text = re.sub(r'[?]+', '?', text)
    text = re.sub(r'[.]+', '.', text)
    text = re.sub(r'"', '"', text)
    text = re.sub(r'\s+', ' ', text).strip() # Remove and double spaces
    text = re.sub(r'&?;', r'and', text) # replace & -> and
    text = re.sub(r"https?:\/\/\/t.co\/[A-Za-z0-9]+", "", text) # Remove URLs
    # remove some puncts (except . ! # ?)
    text = re.sub(r'[:"$%&^*+,-/;:<=>@\^_`{|}~]+', '', text)
    emoji_pattern = re.compile("[
        u"\U0001F600-\U0001F64F" # emoticons
        u"\U0001F300-
\U0001F5FF" # symbols & pictographs
        u"\U0001F680-
\U0001F6FF" # transport & map symbols
        u"\U0001F1E0-\U0001F1FF" # flags (iOS)
        u"\U00002702-\U000027B0"
        u"\U000024C2-\U0001F251"
        "]" + , flags=re.UNICODE)
    text = emoji_pattern.sub(r'EMOJI', text)

    tokens = []
    for token in text.split():
        if token not in stop_words:
            tokens.append(lemmatizer.lemmatize(token))

    return " ".join(tokens)

```

First Clustering Analysis

To accomplish some simple text clustering, we used Sklearn's TFIDF vectorizer and Minibatch means. The IDF score was intended to extract unique words that might be employed for clustering. Clustering is an unsupervised technique, and K-Means needs that the number of clusters be defined.

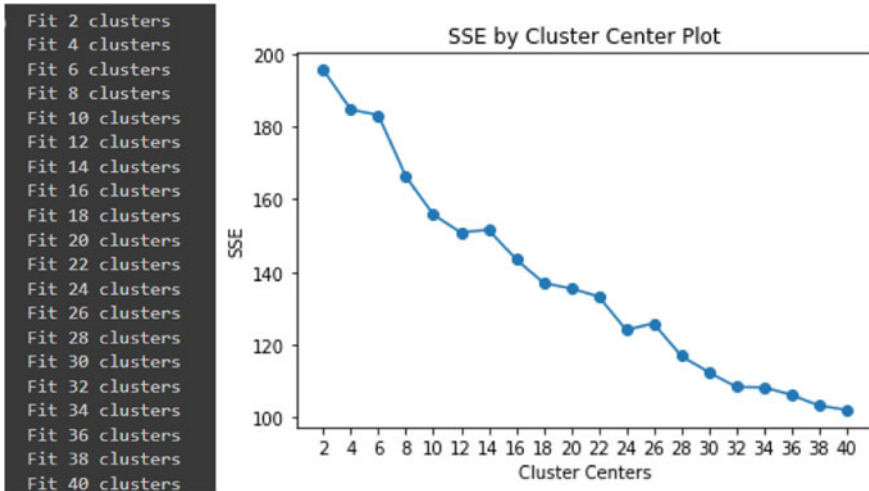


Fig. 8 Optimal cluster representation

A straightforward method is to plot the Error Sum of Squares (SSE) for a variety of cluster sizes.

```
def find_optimal_clusters(data, max_k):  
    iters = range(2, max_k+1, 2)  
  
    sse = []  
    for k in iters:  
        sse.append(MiniBatchKMeans(n_clusters=k, init_size=1024, batch_size=2048, random_state=20).fit(data).inertia_)  
        print('Fit {} clusters'.format(k))  
  
    f, ax = plt.subplots(1, 1)  
    ax.plot(iters, sse, marker='o')  
    ax.set_xlabel('Cluster Centers')  
    ax.set_xticks(iters)  
    ax.set_xticklabels(iters)  
    ax.set_ylabel('SSE')  
    ax.set_title('SSE by Cluster Center Plot')  
    find_optimal_clusters(clean_tweets_tfidf, 40)
```

We seek the “elbow” when the SSE begin to level off. MiniBatchKMeans generates some noise, therefore we increased the batch and init sizes. Nonetheless, the standard implementation of K-Means is too sluggish. Indeed, we will observe that distinct random states yield distinct plots (Fig. 8).

Then, we depicted the clusters that our K-Means algorithm created. One plot employs Principal Component Analysis (PCA), which captures the global structure of

```
[ ] clusters = MiniBatchKMeans(n_clusters=3, init_size=1024, batch_size=2048, random_state=20).fit_predict(clean_tweets_tfidf)

[ ] from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
import matplotlib.cm as cm

def plot_tsne_pca(data, labels):
    max_label = max(labels)
    max_items = np.random.choice(range(data.shape[0]), size=3000, replace=True)

    pca = PCA(n_components=2).fit_transform(data[max_items,:].todense())
    tsne = TSNE().fit_transform(PCA(n_components=50).fit_transform(data[max_items,:].todense()))

    idx = np.random.choice(range(pca.shape[0]), size=300, replace=False)
    label_subset = labels[max_items]
    label_subset = [cm.hsv(i/max_label) for i in label_subset[idx]]

    f, ax = plt.subplots(1, 2, figsize=(14, 6))

    ax[0].scatter(pca[idx, 0], pca[idx, 1], c=label_subset)
    ax[0].set_title('PCA Cluster Plot')

    ax[1].scatter(tsne[idx, 0], tsne[idx, 1], c=label_subset)
    ax[1].set_title('TSNE Cluster Plot')

plot_tsne_pca(clean_tweets_tfidf, clusters)
```

Fig. 9 Plotting clusters algorithm

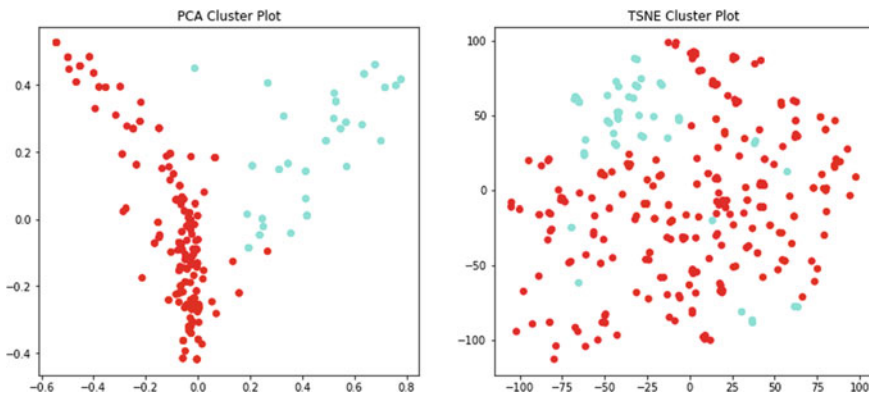


Fig. 10 Plotting clusters graphic representation

the data more accurately. The alternative employs t-SNE, a Nonlinear dimensionality reduction technique which is higher at capturing adjacent connections. In order to accelerate the t-SNE process, we sampled 3,000 documents and applied a PCA 50-dimension reduction to the data (Figs. 9 and 10).

Finally, we revised over the clusters and printed the top terms based on their TFIDF score to determine whether there were any evident patterns and trends. This may be accomplished in Pandas by computing an average value across all dimensions, organized by cluster label. Finding the top words in NumPy is as simple as sorting the average values for each row and picking the top 15 words (Fig. 11) in our case.

Based on the findings of the Keywords algorithms, the terms found for each cluster type appear to be relatively similar, with the exception of the last cluster, which tends

```

Cluster 0
city,warrior,area,explosive,device,week,rejectionist,police,apostate,vehicle,member,killed,army,soldier,caliphate

Cluster 1
apostate,security,infidel,number,fight,muslim,operation,jihad,province,islam,attack,soldier,army,islamic,state

Cluster 2
force,house,attack,mujahedeen,country,muslim,apostate,christian,mujahideen,people,almighty,security,jihad,allah,god
    
```

Fig. 11 Top keywords found

```

[ ] print(len(dataset[dataset["MAIN CATEGORY "] == "War"]))
print(len(dataset[dataset["MAIN CATEGORY "] == "Utopia"]))
print(len(dataset[dataset["MAIN CATEGORY "] == "Victimhood"]))

116
71
37

[ ] dataset["cluster"] = clusters

[ ] dataset.to_excel("Classification_Text_Selection_Al-Naba_con_cluster.xlsx")
    
```

Fig. 12 Creation of Al-Naba’ dataset based on the identified clusters

```
[24] dataset["cluster"] = clusters
```

```
[25] dataset.to_excel("Classification_Text_Selection_Al-Naba_con_cluster.xlsx")
```

Unnamed: 0	ISSUE	GREGORIAN DATE	ISLAMIC DATE	text	Unnamed: 4	MAIN CATEGORY	SUB-CATEGORY	OTHER SUB-CATEGORY	cluster	
0	0	#1	2015-10-17	3 Muharram 1437	Ngrnesio (Hama governorate) mourns pagan defens...	NaT	War	Dar al-hlam vs Dar al-Harb	NaN	0
1	1	#1	2015-10-17	3 Muharram 1437	Supporters were murdered in a high-quality ope...	NaT	War	Act of Terror	NaN	1
2	2	#1	2015-10-17	3 Muharram 1437	The mujahedeen stood up in the area of Sackani...	NaT	War	Dar al-hlam vs Dar al-Harb	NaN	0
3	3	#1	2015-10-17	3 Muharram 1437	Caliphate soldiers also targeted a vehicle of...	NaT	War	Act of Terror	NaN	0
4	4	#1	2015-10-17	3 Muharram 1437	A patrol of Asayish (internal security) solde...	NaT	War	Act of Terror	NaN	0
...
229	229	#346	2022-07-08	8 Dhu'l-hija 1443	Caliphate soldiers in Mozambique State Greedy...	NaT	War	Armed Attack	NaN	0
230	230	#346	2022-07-08	8 Dhu'l-hija 1443	Three people were killed and three more were i...	NaT	War	Armed Attack	NaN	2
231	231	#346	2022-07-08	8 Dhu'l-hija 1443	The attacks of the soldiers of the caliphate w...	NaT	War	Armed Attack	NaN	0
232	232	#346	2022-07-08	8 Dhu'l-hija 1443	Before Eid prayers, do not spill the sacrifice...	NaT	Utopia	Religiously inspired terrorism	NaN	2
233	233	#346	2022-07-08	8 Dhu'l-hija 1443	Good luck to God Almighty Caliphate soldiers d...	NaT	Utopia	Religiously inspired terrorism	NaN	0

234 rows x 10 columns

Fig. 13 An overview of Al-Naba’ dataset classification with clusters

to be more religion-focused than violence-focused. Then, each detected cluster was allocated to its corresponding text, and a new dataset was generated (Fig. 12).

From our first research, it was clear that the first cluster was dominated by references to *War*, while the other two deal both with *Utopia* and *Victimization* (Fig. 13). Therefore, we could not consider the acquired findings to be accurate and use them to identify associated terrorist supporters on the Twitter social network.

6.6.3 Second Identification of Al-Naba' Taxonomies

In our previously exploratory analysis, we didn't consider that clustering methodology accepts a list of documents as input; however, each of these documents must be converted from text to a list of numeric values before being processed by the algorithm. There are different ways to convert text into a list of numerical values to represent it, however for our work, we choose to employ once more the Cogito Intelligence API of Expert System.

The relevant metadata was identified by arranging the *Al-Naba'* extracts in a sequential manner for analysis. Thus, we commenced the development of our dataset within a shared Excel spreadsheet.

Its line-column chart is a mix of a line graph (containing Al-Naba texts) and a set of columns that correspond to the extracted Cogito system parameters, with as few empty cells as feasible. We also attempted to maintain duplicate values across documents in the same columns. For each row, we had to place the first value with whitespace to its left, locate the first available whitespace to its left, ensure that there are no other values in that column occupying a space that would be occupied by another row of the value under consideration, and copy and paste all the values into the new leftmost column.

The new dataset created contains the following columns:

```
['text', 'cat1', 'cat2', 'cat3', 'cat4', 'cat5', 'cat6', 'cat7',
'cat8', 'cat9', 'Organizations 1', 'Organizations 2', 'Designated
Terrorist Organizations 1', 'Designated Terrorist Organizations
2', 'People', 'Places 1', 'Places 2', 'Places 3', 'Military Forces
1', 'Military Forces 2', 'Building', 'Military Equipment 1', 'Mili-
tary Equipment 2', 'Military actions 1', 'Military actions 2',
'Inferential entities 1', 'Inferential entities 2', 'Inferential
entities 3', 'fact1', 'fact2', 'fact3', 'fact4', 'fact5', 'fact6',
'fact7', 'fact8']
```

The line-column chart is a composite visualization consisting of a line graph that displays *Al-Naba'* texts and a series of columns that represent the extracted Cogito system parameters. The chart has been designed to minimize the number of empty cells present. A dataset was constructed through the use of a JSON score, whereby each text was represented by a vector (a list) of scores derived from all categories that were present in all documents. To balance the dataset, the scores were transformed to a range of 0 to 1. This was achieved by a conversion process, whereby the original scores, which spanned from 0 to infinity, were adjusted accordingly. To accomplish this task, the `sklearn.preprocessing.MinMaxScaler` was applied. The clustering algorithm was applied to the dataset with a predetermined number of five clusters.

Consequently, a dataset was produced, featuring a column situated at its conclusion. Our research aimed to identify the associations between the assigned clusters for each text and the corresponding activated categories. Upon conclusion, it was observed that two clusters were identified with a focus on the categories of *explosion/explosives* and *terrorism/act of terror*.

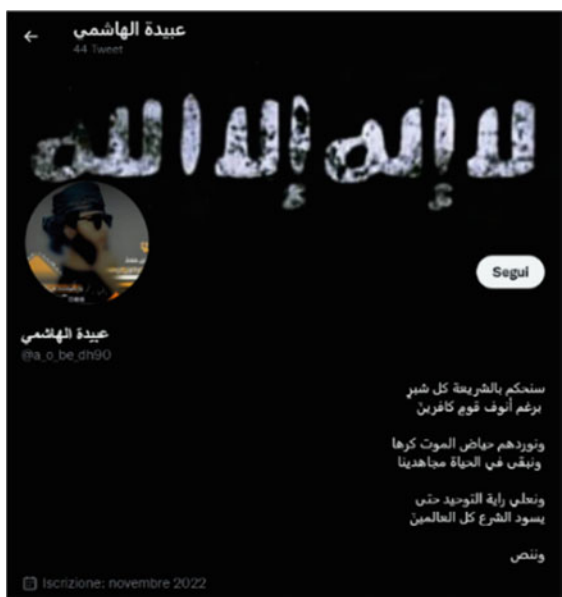
After carrying out training on the clustering model using n text categories, we proceeded to examine whether the incorporation of Tweets was significantly linked to the aforementioned categories. Various open-source intelligence techniques were employed to extract tweets from users who may have a specific association with Daesh and its jihadist propaganda. Initially, the Twitter Advanced Search function was utilized with the subsequent query:

(الدولة OR الإسلامية OR داعش OR ISIS OR عالمية جهاد حركه OR صحيفة النبأ) lang:ar
until:2022-11-09 since:2022-01-01

One of the search results, namely the account with the handle @a_o_be_dh90, piqued our interest. However, it was observed that the account had been suspended within a short span of a few days (Fig. 14).

Our preliminary inquiry encompassed not only the substance of disseminated posts, but also the individuals who follow and are followed by the account in question. To enhance research efficiency and accuracy, our team has opted to incorporate an additional open-source intelligence (OSINT) tool, specifically [TweetTopic](#). This specific instrument provides a useful function that has proven to be beneficial in our research. Upon provision of the Twitter handle, the tool proceeds to collect the latest 3,000 Tweets and subsequently generates a word cloud. This method identifies the prevalent terms utilized within the postings of the target. Upon clicking a particular phrase from the search results, solely those Tweets that contain the designated

Fig. 14 A Twitter account with Daesh affiliation (already suspended)



term are exhibited. Upon selecting any of the textual circles, the associated posts are promptly displayed. The utility of the tool was significant in instances where a large volume of posts needed to be reviewed within a limited timeframe. Tweet-Topic facilitates expedient identification of the subject matter of the tweets of our intended recipient and expedites our inquiry into any pertinent topics. Furthermore, the word cloud exhibits the highest occurring retweets or identification accounts linked to the assessed tweet. Consequently, a collection of one hundred Twitter accounts, ten of which are presented in Table 4, and conceivably associated with Daesh, were established. Of the aforementioned accounts, ninety-three were generated between February 2022 and November 8th, 2022. Certain tweets made by them include references to articles sourced from Al-Naba' magazine (Fig. 15).

Table 4 Suspected Daesh accounts on Twitter

Account	Link	Registration time
Account-name_1	Twitter-url-account-name_1	September 2022
Account-name_2	Twitter-url-account-name_2	September 2022
Account-name_3	Twitter-url-account-name_3	September 2021
Account-name_4	Twitter-url-account-name_4	February 2018
Account-name_5	Twitter-url-account-name_5	September 2012
Account-name_6	Twitter-url-account-name_6	October 2022
Account-name_7	Twitter-url-account-name_7	September 2022
Account-name_8	Twitter-url-account-name_8	November 2022
Account-name_9	Twitter-url-account-name_9	July 2022
Account-name_10	Twitter-url-account-name_10	October 2022

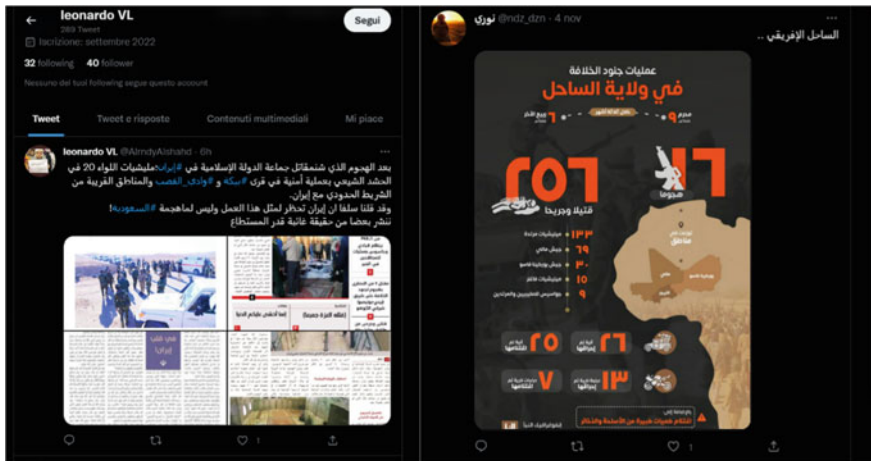


Fig. 15 Tweets, on Al-Naba' magazine, of suspected Daesh-related accounts

Given the possibility of categorizing tweets into the identified clusters, our endeavor was to code them in a manner consistent with documents collected from Twitter, as a training dataset, in order to determine whether they were appropriately classified within their expected cluster.

The dataset used in our study consists of 17,000 tweets, which were obtained from a publicly available dataset provided by Kaggle, a renowned community in the field of data science. These tweets originate from diverse accounts that have actively expressed their endorsement of Daesh. The tweets encompass a time span starting from January 6, 2015, and concluding on May 13, 2016 [54].

We used Seaborn, a Python library that facilitates the creation of visually enhanced charts through its heatmap() function (Fig. 16). The application of clustering techniques on a heatmap is a common practice. The concept involves categorizing items that exhibit similar patterns in their numerical variables.

Afterwards, we named our clusters. The emergence of each category has been contingent upon its presence in more than 1% of the texts within each cluster, as follows:

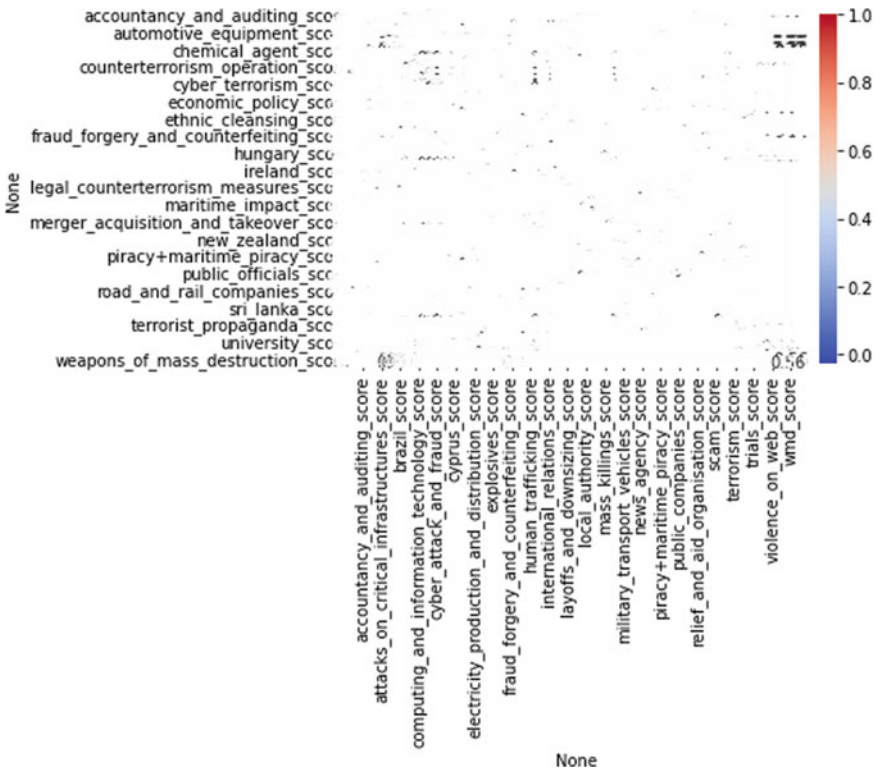


Fig. 16 Heatmap of the training clustered dataset

```

Cluster n: 0
|- 0 testi, 0 categorie
---
Cluster n: 1
|- 475 testi, 1 categorie
|-- 475 testi di categoria 'islam'
---
Cluster n: 2
|- 2465 testi, 12 categorie
|-- 162 testi di categoria 'act_of_terror'
|-- 178 testi di categoria 'armed_forces'
|-- 152 testi di categoria 'counter_terrorist_attack'
|-- 152 testi di categoria 'cyber_security'
|-- 242 testi di categoria 'iraq'
|-- 2391 testi di categoria 'key_groups'
|-- 181 testi di categoria 'politics'
|-- 505 testi di categoria 'religiously_inspired_terrorism'
|-- 461 testi di categoria 'syria'
|-- 2448 testi di categoria 'terrorism'
|--361 testi di categoria 'terrorist_attack_on_critical_
infrastructures'
|-- 156 testi di categoria 'united_states_of_america'
---
Cluster n: 3
|- 398 testi, 3 categorie
|-- 398 testi di categoria 'iraq'
|-- 168 testi di categoria 'key_groups'
|-- 201 testi di categoria 'terrorism'
---
Cluster n: 4
|- 5567 testi, 0 categorie
---
Cluster n: 5
|- 800 testi, 5 categorie
|-- 266 testi di categoria 'cyber_security'
|-- 316 testi di categoria 'key_groups'
|-- 212 testi di categoria 'russia'
|-- 786 testi di categoria 'syria'
|-- 332 testi di categoria 'terrorism'
---
Cluster n: 6
|- 980 testi, 0 categorie
---
Cluster n: 7
|- 590 testi, 6 categorie
|-- 152 testi di categoria 'act_of_terror'
|-- 587 testi di categoria 'key_groups'
|-- 230 testi di categoria 'religiously_inspired_terrorism'
|-- 195 testi di categoria 'syria'
|-- 589 testi di categoria 'terrorism'
|--184 testi di categoria 'terrorist_attack_on_critical_
infrastructures'
---
Cluster n: 8

```

```

|- 974 testi, 0 categorie
---
Cluster n: 9
|- 123 testi, 0 categorie
---
[{}, {'islam_score': 475}, {'act_of_terror_score': 162, 'armed_
forces_score': 178, 'counter_terrorist_attack_score': 152,
'cyber_security_score': 152, 'iraq_score': 242, 'key_groups_
score': 2391, 'politics_score': 181, 'religiously_inspired_
terrorism_score': 505, 'syria_score': 461, 'terrorism_score':
2448, 'terrorist_attack_on_critical_infrastructures_score':
361, 'united_states_of_america_score': 156}, {'iraq_score': 398,
'key_groups_score': 168, 'terrorism_score': 201}, {}, {'cyber_
security_score': 266, 'key_groups_score': 316, 'russia_score':
212, 'syria_score': 786, 'terrorism_score': 332}, {}, {'act_
of_terror_score': 152, 'key_groups_score': 587, 'religiously_
inspired_terrorism_score': 230, 'syria_score': 195, 'terrorism_
score': 589, 'terrorist_attack_on_critical_infrastructures_
score': 184}, {}, {}]
0
1 islam_score
2 act_of_terror_score / armed_forces_score / counter_terrorist_
attack_score / cyber_security_score / iraq_score / key_groups_
score / politics_score / religiously_inspired_terrorism_score
/ syria_score / terrorism_score / terrorist_attack_on_critical_
infrastructures_score / united_states_of_america_score
3 iraq_score / key_groups_score / terrorism_score
4
5 cyber_security_score / key_groups_score / russia_score / syria_
score / terrorism_score
6
7 act_of_terror_score / key_groups_score / religiously_inspired_
terrorism_score / syria_score / terrorism_score / terrorist_
attack_on_critical_infrastructures_score

```

Subsequently, we provided an enhanced rendition of cluster dictionaries that exhibits improved readability. The transformation of dictionaries and sets into lists was performed:

Since K-means clustering is an unsupervised technique clustering, it cannot be evaluated in comparison to prior knowledge. Therefore, in unsupervised clustering, internal evaluation, which confirms the homogeneity of the objects (tweets) inside each cluster and the isolation of each cluster from the others, will be one of the approaches to assess cluster quality. An alternative option is to change the value of the parameter and then see the effects. It is important to note that the clustering procedure itself will not change.

6.7 Research Scope and Limitations

Undoubtedly, our research is subject to limitations, as there exist numerous unknown factors and potential challenges that may arise. Both the work and the analysis are

insufficient to conduct a comprehensive investigation. Despite the aforementioned constraints, our objective was to demonstrate the potential of intricate networks in uncovering concealed trends within the worldwide landscape of terrorism, with the aim of encouraging the application of novel analytical frameworks to the examination of Daesh media narratives. The identification of concealed patterns across multiple social and digital platforms can lead to the discovery of commonalities in propaganda. It is our assessment that there exists an urgent requirement to develop and validate novel methodological approaches which, if efficacious, could be extrapolated to other settings with more accurate data, thereby facilitating more beneficial and conclusive outcomes.

7 Conclusions

Daesh keeps on using several technologies to carry out its deeds and plans. Virtual Jihad is becoming more and more an issue of concern that must be addressed and depends largely on offensive propaganda-based information warfare. This narrative-driven, heightened form of terrorism has become Daesh major asymmetric weapon. Online terrorism, specifically, has long-lasting effects on the psyche (“Cognitive Warfare”) of impacted communities and virtual users and is capable of causing significant harm.

Over the past two decades, efforts to combat the menace of terrorism have yielded some positive outcomes. However, in the forthcoming years, it will be imperative to factor in the risks associated with advancements on the global network of interconnected computer systems, commonly referred to as the World Wide Web. Consequently, it is imperative for law enforcement, intelligence agencies, and other relevant entities to continually devise novel and distinctive methodologies aimed at thwarting, detecting, and mitigating terrorist operations conducted via the Internet. The gathering and evaluation of intelligence can be instrumental in the assessment of potential threats, the development of appropriate responses, and the direction of policy. The impact of emerging technologies and societal changes on the functioning of law enforcement necessitates careful consideration in this procedure. Furthermore, the cultivation of cultural and digital literacy is imperative in promoting intergovernmental cooperation and alliance in combatting cyberterrorism.

The information collected from online sources has the potential to assist law enforcement in obstructing the group’s channels of communication and furnish insights for devising strategies to refute their accounts. Concealed patterns across various social and digital platforms may unveil similarities in propaganda. OSINT is an essential tool in the fight against Daesh’s online propaganda. While not a panacea, it can prove advantageous in identifying and disrupting the collective’s dissemination of propaganda.

Effective utilization of Open Source Intelligence (OSINT) while adhering to legal and privacy regulations necessitates collaboration among governments, technology corporations, and civil society. The employment of OSINT for such purposes

necessitates adherence to ethical and legal guidelines, and any dubious conduct must be brought to the attention of the appropriate law enforcement or security agencies.

The objective of our study was to showcase the efficacy of integrating basic data mining tools with social network analysis capabilities, alongside conventional data mining techniques and practical semantic analysis of online propaganda, in order to effectively identify terrorist affiliations with Al-Naba' magazine. Our investigation commenced by posing the following inquiry: can Machine Learning and Natural Language Processing algorithms be utilized to evaluate Daesh narratives with the aim of detecting potential resemblances among various propaganda sources?

In light of the potential for classifying tweets into the designated clusters, our objective was to encode them in a manner that aligns with the Twitter data collected as a training dataset. This was done to assess whether the tweets were accurately categorized within their respective clusters extracted from Al-Naba' publications. As a result, the occurrence of each category has been dependent on its prevalence in more than 1% of the texts within each cluster.

This particular case study presents a distinct chance to make inferences about positively labelled cases by utilizing Twitter suspensions and clustering techniques. Nevertheless, the consistent availability of a substantial quantity of labelled cases is improbable. Hence, the inclusion of semi-supervised algorithms in our implementations would enhance the generalizability of our approach, thereby warranting further investigation in future research endeavours.

The identification of extremist communities holds significant importance in the realm of social media analysis. It is imperative to address the influence exerted by groups such as Daesh, with the aim of mitigating their impact in the foreseeable future. Based on the employed methodology, there is no basis for reservations regarding the possibility of exploring additional viable research avenues beyond the point that has been reached. The primary objective is to render the internet an inhospitable milieu for terrorist propaganda, thereby depriving entities such as Daesh of their most potent instrument for radicalization and recruitment.

Disclaimer The present document is merely informative in nature and refrains from endorsing any unsanctioned individuals or entities to directly engage with Daesh propaganda. It is recommended that such activities be exclusively carried out by individuals who have received adequate training and are operating within the confines of appropriate legal regulations. It is imperative to promptly notify the appropriate law enforcement authorities of any dubious actions.

References

1. Erdem BK, Bilge R (2017) The announcement of dar al-harb in cyber media in context of the theological policy of jihad: Reading the Cyber-Jihad and ISIS based on the Pharmakon Characteristic of the Cyber Media. *Int J Islamic Thought* 11:17
2. Krebs VE (2002) Mapping networks of terrorist cells. *Connections* 24(3):43–52
3. Brams SJ, Mutlu H, Ramirez SL (2006) Influence in terrorist networks: From undirected to directed graphs. *Stud Conflict Terrorism* 29(7):703–718. <https://doi.org/10.1080/10576100600701982>

4. Koschade S (2006) A social network analysis of Jemaah Islamiyah: The applications to counterterrorism and intelligence. *Stud Conflict Terrorism* 29(6):559–575. <https://doi.org/10.1080/10576100600798418>
5. Lautenschlager J et al (2015) Group profiling automation for crime and terrorism (GPACT). *Procedia Manuf* 3:3933–3940
6. Berger JM, Morgan J (2015) The ISIS Twitter census: Defining and describing the population of ISIS supporters on Twitter
7. Magdy W, Darwish K, Weber I (2015) #FailedRevolutions: Using Twitter to study the antecedents of ISIS support. arXiv preprint [arXiv:1503.02401](https://arxiv.org/abs/1503.02401)
8. Kaati L et al (2015) Detecting multipliers of Jihadism on Twitter. In: 2015 IEEE international conference on data mining workshop (ICDMW), pp 954–960. IEEE. https://www.foi.se/download/18.7fd35d7f166c56ebe0b1000c/1542623725677/Detecting-multipliers-of-jihadism_FOI-S--5656--SE.pdf
9. Rowe M, Saif H (2016) Mining pro-ISIS radicalisation signals from social media users. In: Proceedings of the international AAAI conference on web and social media, vol 10, no 1, pp 329–338. <https://doi.org/10.1609/icwsm.v10i1.14716>
10. Cantalapiedra DG (2019) Toward a new security concept in a multi-domain space: complexity, war and cross-domain security. *ieee.es | Instituto Español de Estudios Estratégicos*. https://www.ieee.es/Galerias/fichero/docs_opinion/2019/DIEEEO85_2019DAVGAR_seguridad_ENG.pdf. Accessed 8 Sept 2022
11. Le Guyader H (2022) Cognitive domain: a sixth domain of operations. *Hal Open Science*, p 3-1. <https://hal.science/hal-03635898/document>. Accessed 26 June 2023
12. Caliskan M (2019) Hybrid warfare and strategic theory. Beyond the Horizon ISSG. <https://behorizon.org/hybrid-warfare-through-the-lens-of-strategic-theory/>. Accessed 27 Sept 2022
13. Zanini M, Sean JAE (2001) The networking of terror in the information age. In: *Networks and netwars: the future of terror, crime, and militancy*, p 32
14. Theohary CA (2015) Information warfare: the role of social media in conflict. Congressional Research Service, the Library of Congress. <https://sgp.fas.org/crs/misc/IN10240.pdf>. Accessed 19 Sept 2022
15. Zeiger S, Gyte J (2020) Prevention of radicalization on social media and the Internet. International Centre for Counter-Terrorism (ICCT). <https://icct.nl/app/uploads/2021/10/Chapter-12-Handbook.pdf>. Accessed 28 Sept 2022
16. Gartenstein-Ross D (2015) Radicalization: social media and the rise of terrorism. House Testimony, Hearing before the US House of Representatives Committee on Oversight and Government Reform, Subcommittee on National Security, p 28. <https://www.govinfo.gov/content/pkg/CHRG-114hhrg97977/html/CHRG-114hhrg97977.htm>. Accessed 28 Sept 2022
17. Ingelevič-Citak M, Przystlak Z (2020) Jihadist, far-right and far-left terrorism in cyberspace—same threats and same countermeasures?. *Int Comp Jurisprud* 6:2. <https://doi.org/10.13165/J.ICJ.2020.12.005>
18. Thomas TL (2003) Al Qaeda and the Internet: the danger of *cyberplanning*. Foreign Military Studies Office (ARMY). Fort Leavenworth Ks. <https://apps.dtic.mil/sti/pdfs/ADA485810.pdf>. Accessed 30 Sept 2022
19. Meleagrou-Hitchens A, Hughes S (2017) The threat to the United States from the Islamic State’s virtual entrepreneurs. *Combating Terrorism Center at West Point*, vol 10. Mar 2017. <https://ctc.usma.edu/the-threat-to-the-united-states-from-the-islamic-states-virtual-entrepreneurs/>. Accessed 2 Sept 2022
20. Bishop P, Macdonald S (2019) Digital Jihad: online communication and violent extremism. ISPI, Milano, pp 135–136
21. Stenersen A (2008) *The Internet: a virtual training camp?* Taylor & Francis Group
22. Todorovic B, Trifunovic D (2020) Prevention of (ab-) use of the Internet for terrorist plotting and related purposes. International Centre for Counterterrorism (ICCT). <https://icct.nl/app/uploads/2021/10/Chapter-19-Handbook.pdf>. Accessed 24 Oct 2022
23. Trifunović D (2015) Digital steganography in terrorist networks. In: *Proceedings (ZBORNİK RADOVA)*

24. Ruge F (ed) (2018) Cybercrime as a threat to international security. ISPI Dossier. https://www.ispionline.it/sites/default/files/pubblicazioni/dossier_cyber_ruge_july2018.pdf
25. Terrorist financing. Financial Action Task Force. Report 2008. <https://www.fatf-gafi.org/media/fatf/documents/reports/FATF%20Terrorist%20Financing%20Typologies%20Report.pdf>. Accessed 24 Oct 2022
26. Adams J (1986) The financing of terror: behind the PLO, IRA, Red Brigades, and M-19 stand the paymasters: how the groups that are terrorizing the world get the money to do it. Simon and Schuster
27. Borum R (2007) Psychology of terrorism. University of South Florida Tampa Dept of Mental Health Law and Policy. <https://www.ojp.gov/pdffiles1/nij/grants/208552.pdf>. Accessed 20 Oct 2022
28. Zelin A (2013) The state of global Jihad online. New America Foundation. <https://www.washingtoninstitute.org/media/3122>. Accessed 25 Oct 2022
29. Leveraging artificial intelligence to counter terrorism. Voyager Labs. Sept 2021. <https://www.voyager-labs.com/leveraging-artificial-intelligence-to-counter-terrorism/>. Accessed 27 Oct 2022
30. Chau M et al (2016) Intelligence and security informatics: 11th Pacific Asia workshop. PAISI 2016. Auckland, New Zealand: 19 April 2016, proceedings, vol 9650. Springer
31. Bottazzi G, Me G (2014) The botnet revenue model. In: Proceedings of the 7th international conference on security of information and networks, Glasgow, UK, 9–11 Nov 2014. ACM, New York, NY, USA, pp 459–465
32. Perliger A, Pedahzur A (2011) Social network analysis in the study of terrorism and political violence. *Polit Sci Polit* 44(1):45–50
33. Belli R et al (2015) Exploring the crime-terror nexus in the United States: a social network analysis of a Hezbollah network involved in trade diversion. *Dyn Asymmetric Confl* 8(3):263–281
34. Chenoweth E, Lowham E (2007) On classifying terrorism: a potential contribution of cluster analysis for academics and policy-makers. *Defence Secur Anal* 23(4):345–357
35. Qi X et al (2010) A hierarchical algorithm for clustering extremist web pages. In: 2010 international conference on advances in social networks analysis and mining. IEEE
36. Papadopoulos S et al (2012) Community detection in social media. *Data Min Knowl Disc* 24(3):515–554
37. Boccaletti S et al (2006) Complex networks: structure and dynamics. *Phys Rep* 424(4–5):175–308
38. Joseph K, Carley K (2015) Culture, networks, twitter and foursquare: testing a model of cultural conversion with social media data. In: Proceedings of the international AAAI conference on web and social media, vol 9, no 1
39. Newman ME (2006) Modularity and community structure in networks. *Proc Nat Acad Sci* 103(23):8577–8582. 3.1
40. Blondel VD et al (2008) Fast unfolding of communities in large networks. *J Stat Mech: Theory Exp* 2008(10):P10008
41. Tang L, Liu H (2011) Leveraging social media networks for classification. In: Data mining and knowledge discovery, vol 23, pp 447–478
42. Wang X et al (2010) Discovering overlapping groups in social media. In: 2010 IEEE 10th international conference on data mining (ICDM). IEEE, pp 569–578
43. Binkiewicz N, Vogelstein JT, Rohe K (2014) Covariate assisted spectral clustering. arXiv preprint [arXiv:1411.2158](https://arxiv.org/pdf/1411.2158.pdf). 2.2.1, 3.1. <https://arxiv.org/pdf/1411.2158.pdf>
44. Miller BA, Beard MS, Bliss NT (2011) Eigenspace analysis for threat detection in social networks. In: 14th international conference on information fusion. IEEE, pp 1–7
45. Klausen J (2015) Tweeting the Jihad: social media networks of Western foreign fighters in Syria and Iraq. *Stud Confl Terror* 38(1):1–22
46. Berger JM, Perez H (2016) The Islamic State’s diminishing returns on Twitter: how suspensions are limiting the social networks of English-speaking ISIS supporters. George Washington University Program on Extremism

47. Bodine-Baron et al (2016) Examining ISIS support and opposition networks on Twitter. RAND, Santa Monica, CA. https://www.rand.org/pubs/research_reports/RR1328.html
48. Chatfield AT, Reddick CG, Brajawidagda U (2015) Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided Twitter networks. In: Proceedings of the 16th annual international conference on digital government research
49. Xu J, Lu T-C (2016) Automated classification of extremist Twitter accounts using content-based and network-based features. In: 2016 IEEE international conference on big data (big data). IEEE
50. Ashcroft M et al (2015) Detecting jihadist messages on Twitter. In: 2015 European intelligence and security informatics conference. IEEE
51. Me G, Mucci MF (2023) Countering Daesh cognitive and cyber warfare with OSINT and basic data mining tools. In: IC3 2023, vol 10, pp 71–80
52. Awford J (2014) Jihadist social media policy: ISIS look to stay one step ahead. Mail Online. <https://www.dailymail.co.uk/news/article-2798185/isis-look-stay-one-step-ahead-western-spies-teaching-fighters-remove-metadata-tweets.html>. Accessed 12 Nov 2022
53. Religious texts used by ISIS. Kaggle.com. 2014. <https://www.kaggle.com/datasets/fifthtribe/isis-religious-texts>
54. How ISIS uses Twitter. Kaggle.com. 2019. <https://www.kaggle.com/datasets/fifthtribe/how-isis-uses-twitter>

Cyberbullying Detection in Twitter Using Deep Learning Model Techniques



Anu Ranjana Seetharaman and Hamid Jahankhani

Abstract The aim of the research is to develop a combined deep machine learning model (DEA_RNN) that will detect cyber bullying on Twitter. The suggested approach merges Elman-type Recurrent Neural Networks (RNNs) with a refined Dolphin Echolocation Algorithm (DEA) to improve parameter optimization and reduce the training duration. The proposed approach can handle the energetic nature of brief writings, adapt with point models for compelling extraction of trending themes and is proficient in recognizing and classifying cyberbullying tweets. The four modules in the proposed system are Dataset Collection, Data Cleaning and Preprocessing, Algorithm Implementation, and Prediction. The data cleaning and preprocessing phase involve noise removal, out-of-vocabulary cleansing, and tweet transformations to improve feature extraction and classification accuracy. The model algorithm uses DEA and RNN to generate a kind of echo similar to the behavior of dolphins during the hunting process. This research attempts to address the issue of recognizing cyberbullying on Twitter, which could be a challenging errand due to the energetic nature of brief writings and the changing nature of cyberbullying. The proposed approach beats models such as Simple Recurrent Neural Network (RNN), in terms of execution.

Keywords Deep machine learning · Twitter · Recurrent neural networks · Social media · Cyberbullying · Convolutional neural network

A. R. Seetharaman · H. Jahankhani (✉)
Northumbria University, London, UK
e-mail: Hamid.jahankhani@northumbria.ac.uk

A. R. Seetharaman
e-mail: anu.seetharaman@northumbria.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_7

1 Introduction

Social media platforms have gained widespread popularity and usage across individuals of all ages. However, with this popularity comes the increasing prevalence of cyberbullying. Cyberbullying (CB) refers to hostile conduct that is purposeful, recurrent, and entails a power imbalance or asymmetry, and which occurs via electronic means or the internet. Cyberbullying on Twitter is particularly challenging to detect due to the ever-changing characteristic of brief textual content, and the changing nature of cyberbullying. It is important to detect cyberbullying on Twitter to make social media safe. Content alignment based on administered computerized learning models is commonly utilized for putting tweets into offensive and non-offensive tweets. The directed classifiers have poor execution. We cannot change the course names and it is very noteworthy to use scenarios. Too, they may be appropriate as it were for a foreordained group pertaining occasions but cannot deal with tweets that change over time. Topic modeling ways have moreover been used to extricate crucial points from a set of information to make designs or classes within the whole dataset. Whereas the concept is comparable, point modeling has impediments in taking care of the energetic nature of tweets and in recognizing unused occasions and scenarios [11].

The present study recommends the usage of a hybrid deep learning approach termed as DEA-RNN to surmount the challenges in identifying cyberbullying on the Twitter social network. The proposed model is an amalgamation of Elman-based Recurrent Neural Networks (RNNs) and a fine-tuned Dolphin Echolocation Algorithm (DEA) that addresses the problems of parameter optimization and training time reduction. This approach offers numerous benefits over existing methods as it can handle the ever-changing characteristic of short texts, extract popular topics through topic modeling, and classify cyberbullying tweets with high accuracy. Cyberbullying affects to use technology and social medias to bully and hurt other's mental health. This mainly affects the youngsters, and the research are made on its impact especially on the children [2]. Cyberbullying can have genuine impacts on their passionate and mental wellbeing, counting expanding their levels of self-destructive ideation, sadness, and uneasiness.

In expansion, cyberbullying can also have long-term impacts on an individual's mental wellbeing. It is vital to note that cyberbullying not as it were influencing the casualties but moreover the culprits, who may encounter negative mental results such as uneasiness, disgrace, and blame. Cyberbullying may be a genuine issue that can have obliterating impacts on youthful casualties and their families. It is crucial to address this issue and provide support for those who have been affected by cyberbullying [1].

The contributions of this thesis include the proposed DEA-RNN model for cyberbullying detection, a new Twitter dataset collection for evaluation, and thorough experimental results that reveal its superior performance over other competing models. The proposed DEA-RNN model is a promising approach for detecting

cyberbullying on Twitter and has the potential to be extended to other social media platforms as well.

2 Cyberbullying and Its Impact

Among the youths, cyberbullying is one of the growing concerns, and its effect on adolescents and children's well-being is of specific interest to the researchers. It can be defined as any aggressive behaviour which is also carried out through electronic devices or the internet [3]. Cyberbullying has serious impact on the mental health as well as on the emotional health of youngsters and children. Because of this, their suicidal thoughts will be developed, depression, and anxiety level will be increased.

According to the research outcome, it is shown that the victims of cyberbullying experience high depression and anxiety compared to the non-anxiety ones. Also, its impact is more dangerous compared to the traditional bullying. Its victims experience feelings of isolation, shame, and helplessness that lead to mental health problems.

As per the study it has also been seen that cyberbullying has significant effects on an individual's academic performance that result in low grades and desired not to go to school. In addition, their social relationships get affected in a negative way that led to low self-esteem and high social isolation.

Cyberbullying also shows long-term effects on a person's mental health that includes PTSD (Post-Traumatic Stress Disorder), depression, and anxiety [9]. The victims of cyberbullying also experience physical symptoms like sleep disturbances, stomach pains, and headaches. Its impact is not limited to the victims only, perpetrators also experience negative psychological consequences like anxiety, shame, and guilty.

Cyberbullying's effects could be damaging individuals physically. Cyberbullying's physical effects include sleeping issues, headaches, and stomach aches. In addition, all psychological effects involve suicidal thoughts, self-harm, depression, and anxiety. Cyberbullying's victims might have also trouble to form relationships with all other individuals. Such victims might be afraid of trusting others and mightn't want in socializing with all others. It could lead to them to feel isolated as well as lonely. Parents must be aware of all such effects along with taking appropriate actions if they sense the teen is cyberbullied. Cyberbullying could have some serious consequences for all bullies as well as the victims. Hence, both teenagers along with parents must know all effects of preventing cyberbullying from taking place or dealing with this if this happens.

3 Neural Network

A neural network is an AI-based algorithm which can process natural language, recognise speech and image. This algorithm is effective because of its ability to generalize and learn from the examples. This ability helps it to make predictions even in noisy and complex environments [14].

As per the study, the advancement in neural network algorithms has led to development of deep learning which mainly refers to the neural network having several layers. The use of deep learning is successful in many applications like speech and image recognition, autonomous driving, and natural language processing [7].

Neural networks are machine learning's subset and depend upon training data for learning and improving their accuracy with time. although once every learning algorithm is fine-tuned, each of these algorithms is a powerful tool in artificial intelligence and computer science that allow for classifying and clustering data at higher velocity. In addition, tasks in image recognition as well as speech recognition could take hours versus minutes if compared to manual identification [6]. Neural networks could be divided into different types that are utilized for different purposes. Multi-layer perceptron (MLP) includes output layer, input layer along with one hidden layer. Such neural networks are developed of sigmoid neurons, in place of perceptron, as most issues are non-linear.

In Natural Language Processing (NLP), the trained data from the models are used. This trained data is also used for computer vision and other similar neural networks. The MLP and the Convolutional Neural Network (CNN) are similar, but the difference is the CNNs are used for image and pattern recognition and computer vision. Principles of linear algebra, especially matrix multiplication, are harnessed by such networks for identifying patterns in any image. Also, a recurrent neural network (RNN) can be identified using feedback loops.

Neural networks work quite the same as the neural network of any human brain. Any neuron within any neural network is the mathematical function which collects as well as classifies information as per a particular architecture. A strong and efficient resemblance is bear by the network to different statistical methods: regression analysis as well as curve fitting. Any neural network includes all interconnected nodes' layers. In addition, all nodes are termed perception and are the same as any multiple linear regression. The signal is fed by the perception created by the multiple linear regression into the activation function [15, 16]. Neural networks are used widely, with a different application for many financial operations, product maintenance, business analytics, enterprise planning, and trading. Neural networks evaluate price data along with unearthing opportunities to make trade decisions depending upon data analysis.

3.1 *Neural Network Algorithm to Detect Cyberbullying*

As per the study, the detection of cyberbullying is said to be a challenging task because of the generation of a huge amount of data on social media platforms. However, neural network algorithms have shown promise in identifying cyberbullying behaviour. CNN (Convolutional Neural Network) is one of the algorithms which achieved success in the detection of cyberbullying in social media.

CNNs work by processing input data in a series of convolutional layers, where each layer learns specific features from the data. Such features are then used to make predictions about the input data. In the context of cyberbullying detection, CNNs can learn to identify specific patterns of language, sentiment, and context that are associated with cyberbullying behaviour [15].

CNN is the algorithm which advances all current work to detect cyberbullying with the adoption of deep learning's principles in place of classical machine learning. The architecture of CNN includes four layers, such as embedding, max pooling, convolutional and dense. It can be achieved by generating word embeddings for all words in any tweet along with feeding these to the CNN directly. Word embedding is techniques' class used for generating textual material's numerical representation. Word embedding's striking feature is that these generate the same type of representations for similar semantic words. Such a remarkable feature also enables the machine in understanding what any text means in place of dealing with this as random numbers' string.

Another neural network algorithm which has been used to detect cyberbullying is long short-term memory (LSTM) network. Each of these LSTM networks is the kind of recurrent neural network which can learn and remember long-term dependencies in sequential data [6]. This makes them well-suited for detecting cyberbullying in text-based social media data, where the context of a post may be spread out over multiple sentences or comments.

LSTM network can be used as the ensemble model of deep learning with different layers that can outperform single-layer models of neural networks. For further optimizing these models, two-word embeddings' stacking could be done for enhancing the performance of these models. It works on all data in real-time and it assesses if any text contains content of cyberbullying by running this through the model. This LSTM network is utilized within different activities of deep learning and artificial intelligence that include text mining, sequence mining, image processing, and NLP.

As per the study [10] ensemble of CNN and LSTM models is used to detect cyberbullying in tweets. Their results showed also that this ensemble model has outperformed individual models, achieving 0.79 score. Authors noted that the use of neural network algorithms for cyberbullying detection is a promising approach, but more research is needed to develop models that are robust to variations in language and context.

These models make the detection of cyberbullying one fully automated approach along without any human involvement or expertise while guaranteeing also better results. In addition, comprehensive experiments have also proved that algorithms of

deep learning have outperformed classical approaches of machine learning within the issue of cyberbullying. These models can improve the detection of cyberbullying in future and could be used for analysing along with rooting out online bullying's instances perpetrated by users of social media. Classifiers of deep learning to detect hateful texts as well as discussion contents that precede cyberbullying are developed along with might be applied in different platforms [13]. These models have demonstrated some promising performance and will improve much more in the future. Solving the issue is vital to control materials of social media in different languages along with protecting all users from any negative impact of any toxic comment such as offensive language or verbal assaults.

3.2 Naïve Bayes-Based Algorithm to Perform Content-Based Analysis

Naive Bayes (NB) is a probabilistic machine learning algorithm that has been successfully applied in text classification tasks. In the context of cyberbullying detection, NB-based algorithms have shown promise in performing content-based analysis to measure the presence and severity of cyberbullying in social media data [4].

NB algorithms work by computing the probability of each feature in the input data belonging to each class. The class with highest probability then gets assigned to the input data (Andrade et al. 2013). In the context of cyberbullying detection, the input data may be a post, comment, or message, and the classes may be cyberbullying or non-cyberbullying.

To use NB-based algorithms for cyberbullying detection, researchers have developed various features that can capture the linguistic and semantic characteristics of cyberbullying behaviour. These features may include word frequency, sentiment, topic modelling, and other language-based features [12].

The algorithm of Naïve Bayes is the supervised algorithm that is based upon Bayes theorem along with utilized to solve classification issues. This algorithm is used within text classification that includes one high-dimensional set of data to train. Naïve Bayes classifier is an efficient algorithm of classification that helps develop fast models using machine learning which could make some fast predictions. This algorithm assumes that a specific feature's occurrence is independent of other features' occurrence. Each feature of this algorithm contributes individually to identification. Any Naïve Bayes model is easy to develop and quite useful for huge data sets. This algorithm is also known for outperforming every highly sophisticated method of classification. This algorithm performs better for categorical input variables in place of numerical variables.

There exists training data for training the model along with making this functional. Then, there is a need in validating all data to evaluate the model as well as make new predictions. With the use of conditional probability, the evidence's probability can be calculated from all given outputs. The primary goal is computing the

output's probability-based upon all current attributes. This algorithm of Naïve Bayes is utilized as the technique of probabilistic learning for classification of texts. It is a well-known algorithm used for classification of documents of different classes. This algorithm is also used for analysing feelings or sentiments, whether positive, negative, or neutral.

Although the algorithm of Naïve Bayes has numerous limitations, still this is the most selected algorithm to solve issues of classification due to its simplicity. This algorithm works properly upon spam filtering as well as document classification. This algorithm possesses the highest success rate when compared to all other algorithms due to its efficiency as well as speed. In this Naïve Bayes classifier, all features are independently used, and each feature is independent of one another.

3.3 Content-Based Analysis Using Icons and Hieroglyphs

The use of icons and hieroglyphs as selection features for training and testing machine learning algorithms for cyberbullying detection has gained attention in recent years. These features can capture the visual and symbolic characteristics of cyberbullying behaviour that may not be adequately captured by language-based features.

To use icons and hieroglyphs as selection features, researchers have developed various techniques to extract and represent them as numerical features that can be used in machine learning algorithms. These techniques may include colour-based feature, texture-based feature, shape-based feature, and other visual features.

Studies have shown that using icons and hieroglyphs as selection features can improve the accuracy of machine-learning algorithms in cyberbullying detection. For example, as per the study [5] Support Vector Machine (SVM) algorithms are used with visual features to detect cyberbullying in Arabic tweets. Their results showed that the SVM-based algorithm achieved an accuracy of 85.71% in detecting cyberbullying, which is much higher when compared to accuracy achieved by language-based features alone [8].

Visual information, like hieroglyphs, images, and icons, capture imagination. Several hieroglyphs analysis needs epigraphers in spending much time to browse existing catalogues for identifying individual graphs. Technical advances within digitization, information management, and automatic analysis of images are enabling this possibility in analysing, organizing, and visualizing huge cultural datasets. As a visual cue, shapes are used in different tasks of image analysis. Techniques of data visualization that organize as well as present data within one structured graphical format, could be applied for visualizing glyph images along with enabling efficient browsing as well as a search of hieroglyph datasets. These systems could enable mining along with the discovery of semantic as well as visual patterns of hieroglyphs. However, in multimedia computing systems' design to realize cross-cultural computing, an issue is searching and analysing modern data.

Algorithms of computer vision have shown some potential in providing new insights into the digital humanities' realm. Different systems are proposed for aiding

the analysis of artistic, historical, and cultural materials that could efficiently facilitate scholars' daily life in this field. Visual data retrieval is used to search technique drawings, clip-arts, hand-drawn images, 3D objects, trademark images, and natural image datasets. All traditional retrieval systems involve grid-based matching, curve fitting, and point-to-point matching. Such methods either don't scale well upon huge datasets or offer only limited flexibility upon shape variations [17]. However, local shape descriptors have been proposed recently and used for retrieval based on shapes. These methods could scale sub-linearly with proper search structures.

Human speech's emergence can be considered as a coincidental which largely depends upon both non-verbal as well as verbal forms. However, every writing system is different, as writing could not be classified to be a language; this is simply the way to record spoken languages using visible signages. Writing's pictorial style is accepted as the means to communicate. Pictographic writing systems' study has introduced many implicate to develop cultures as well as languages. Such systems include ideas, objects, images, and symbols. Such icons also function as the noun as they explain one word which means one point; however, to construct the meaning, this pictographic work needs the syllables' hierographic structure. Syllabifications are quite like constructing linguistic words as units' combinations within a language. In this new millennium, every emoji icon goes through a similar construction.

3.4 Decision Tree Algorithms to Detect Harmful Words

In the context of cyberbullying detection, decision tree algorithms can be used to classify harmful words into different categories based on their severity and context. Researchers have developed various decision tree-based algorithms for this purpose, such as C4.5, CART, and ID3. These algorithms use different criteria for feature selection and splitting nodes in the decision tree.

The social media platform has become public expression's modern channel for all individuals irrespective of every socio-economic boundary. Due to this pandemic situation, the common people also have begun looking to platforms of social media as the formal manner for being connected with different masses. Several researchers have also published their work upon the automated detection of offensive content as well as hate speech.

Sentiment analysis is a process to classify entities into neutral, negative, or positive categories with the use of NLP. Several models of deep learning and machine learning are used for tackling the issue of the detection of hate speech. Random Forrest Classifier is the ensemble learning approach which uses multiple decision trees for regression along with classification. Such classifiers aggregate decision trees' results for improving the whole accuracy. For any provided data input, there is the creation of one decision tree for all samples within the dataset. Each tree's prediction is voted upon as well as the prediction having the highest votes is also determined as end prediction. Sentiment analysis is performed on the dataset for obtaining each

sentence's polarity, along with testing if the analysis will be enough for classifying all meanings.

Accomplishing the detection of harmful words is a tough task as this includes processing text along with understanding that context. The datasets of harmful words are not clean usually; hence, they should be pre-processed before algorithms of classification could detect harmful words in them. In addition, different models of machine learning have also different strengths which make them better than all others for specific tasks like detecting harmful words. A few models are much more accurate while the remaining others are much more efficient. Also, it is crucial in using different models along with comparing their performance for finding the most appropriate one for the detection of harmful words. Some pre-training approaches have also become popular in the last couple of years, and this is crucial in testing if they work properly with detection algorithms of harmful words.

4 Research Methods

Social media stages have picked up broad notoriety and utilization over people of all ages. Be that as it may, with this notoriety comes the expanding predominance of cyberbullying. Cyberbullying (CB) is characterized as forceful behavior that's purposefulness, rehashed, and includes a lopsidedness of control or quality, which takes put through electronic gadgets or the web. Cyberbullying on Twitter is especially challenging to identify due to the energetic nature of brief writings and the changing nature of cyberbullying. Topic modeling techniques have also been employed to extricate crucial subjects from a set of information to make designs or classes within the whole dataset. Whereas the concept is comparative, point modeling has confinements in dealing with the energetic nature of tweets and in recognizing unused occasions and scenarios [11].

The aim of this investigation is to create a hybrid deep learning framework, named DEA-RNN, for identifying instances of cyberbullying on social media platform Twitter. The research is aimed to assess the efficiency of the DEA-RNN model in comparison to other established algorithms.

The proposed model will be a good way to detect bullying on Twitter and has the potential to be extended to other social media platforms. Moreover, as with any machine learning model, there are potential limitations and ethical considerations that need to be addressed. Like the proposed model might fail to detect minor cyberbullying or might consider something like hard management as cyberbullying. This may cause consequences and it is very important to propose a transparent model and to investigate privacy very seriously while developing the model to avoid potential negative impacts. The proposed DEA-RNN model has the capability to aid in the identification and prevention of cyberbullying on social media platforms. However, further research and careful consideration of ethical principles are necessary to ensure its effectiveness and minimize any unintended consequences.

4.1 *DEA-RNN Model*

The DEA-RNN model is a combination of two main components: a Deep Embedding Attention (DEA) model and RNN. The DEA model is used to get information from the text data. It consists of multiple layers of BiLSTM and Attention Mechanism. The BiLSTM layers help to capture the sequential dependencies in the text data, while the Attention Mechanism helps to highlight the most relevant parts of the text. The output of the DEA model is a fixed-length vector that summarizes the input text.

The RNN component is used to model the temporal dependencies in the Twitter data. Specifically, the model uses a Bi-GRU to capture the temporal dynamics of the Twitter data. The output of the RNN component is also a fixed-length vector that summarizes the temporal dynamics of the Twitter data. Finally, the outputs of the DEA and RNN components are fed in the model.

To train the DEA-RNN model, a large, annotated dataset of cyberbullying tweets is required. Supervised learning technique is used to train the DEA model on the annotated dataset and the unsupervised machine learning techniques is used to train RNN on the unannotated dataset. The DEA-RNN model combines deep learning and recurrent neural networks. So, it is a powerful approach for detecting cyberbullying on Twitter and it captures both the textual and temporal aspects of the Twitter data.

Current worldwide consideration to the challenges of natural assurance is constraining firms and governments to assess, rank, and select eco-efficient innovations. Innovations may devour inputs to create both alluring and undesirable yields. It appears that the data envelopment analysis (DEA) could be an appropriate strategy to assess eco-efficient advances. There are a few DEA expansions for dealing with undesirable yield, and now and then it is troublesome to select an appropriate demonstrate to assess the innovations. The challenge gets to be much more complex when the results of models are not comparable. In such a condition, subjective determination of elective DEA models may lead to deviation from an ideal choice.

4.2 *RNN*

RNN is a recurrent neural network, which is an artificial neural network that was developed to explicitly handle sequential data, such as time-series data or jobs involving natural language processing. These types of data can be processed in different methods. In RNN, a feedback loop is included. This feature can take the previous input and utilize it to inform the processing of the current input in a network. This is achieved by re-introducing the network's output from a time step earlier in the process as an additional input at the time step that is now being processed. Because of this, the RNN can construct a memory of the preceding inputs and the order in which they occurred. Because RNNs can capture long-term dependencies between the inputs, they are particularly useful for simulating the progression of sequential

data. Because of this, they are ideally suited for activities such as music composition, speech recognition, and the processing of natural languages.

However, RNNs are susceptible to the vanishing gradient problem, which manifests itself when the gradients in the network become very insignificant as they propagate back through the feedback loop. This issue can only be solved by avoiding feedback loops entirely. Because of this, it may be challenging for the network to understand its long-term dependencies. In order to solve this issue, a variety of RNNs, such as the LSTM and (GRU networks, have been developed. These networks make use of gating methods to restrict the flow of information throughout the network and avoid the vanishing gradient problem.

In general, RNNs have shown themselves to be an effective tool in the field of machine learning, and this area of research and development remains very active.

In neural networks, recurrent systems have been considered difficult to study because the maximization process is NP-Hard, which applies to all neural networks. The problem is particularly challenging for recurrent systems because of the overall reliance issue. This issue refers to the problems of disappearing and bursting gradients that arise when propagating errors over numerous temporal phases.

Bidirectional RNNs (BRNN), which use data from both the past and the future to determine the outcome at any given time t , were introduced in the second article by Schuster and Paliwal. This differs from earlier networks where the outcome was only influenced by the input. BRNN has been successfully applied, among other things, to series labeling tasks in natural language processing.

One approach to further enhance the capabilities of RNNs is by combining them with Data Envelopment Analysis (DEA) for machine learning. DEA is a technique that is used to evaluate the relative efficiency of decision-making units, which can be adapted to work with RNNs in a number of ways. In addition, DEA can also help to identify the most important input variables, which can be useful for feature selection and data visualization. Another way of combining RNNs with DEA is to use DEA as a post-processing step, where the outputs of the RNN are evaluated using DEA to identify the most efficient outputs. This can be useful for problems where there are multiple output variables, as DEA can help to identify which outputs are most important for decision-making.

In general, the combination of RNNs and DEA has the potential to improve the performance of RNNs in learning solutions, by helping to minimize the dimensionality of the input space, identify important input and output variables, and evaluate the efficiency and performance of the RNN.

5 Data Analysis and Critical Discussion

5.1 Data Gathering

The input dataset consists of tweets that were spitting approximately 32 cyberbullying catchphrases and were obtained through the Twitter API. Some of the catchphrases recommended in brain research writing include bonehead, LGBTQ (gender ambiguous, transgender, and odd), prostitute, pussy, faggot, shit, sucker, skank, ass, live, on edge, simpleton, extortion, ambush, fuck, fucking, disgusting, bitch, ass, whale, etc. Other watchwords like “boycott,” “kill,” “kick the bucket,” “beastly,” “loathe,” and “attack,” “fear-based oppressor,” “peril,” “biased,” and “dim” were suggested inside. The initial collection consists of 430,000 records, with watchwords based on extremism, indignation, swearing, and sexism contributing about 125,000 tweets each. The tweets in this dataset combine different exemptions. Because tweets in the English language are necessary, tweets including other tongue-in-cheek phrases are deleted, and retweets are modified. All of these shapes are created as part of the subsequent pre-processing system. The remaining essential pre-processing activities are then carried out. Information collection from Twitter is the method of recovering information from the social media stage Twitter. Twitter gives an endless source of information that can be utilized for different purposes, counting inquire about, assumption investigation, and promoting examination. There are a few ways to gather information from Twitter, and the strategy utilized depends on the particular needs of the analyst or investigator. One way to gather information from Twitter is to utilize the Twitter API. To utilize the Twitter API, one ought to apply for get to and get an API key, which permits get to the Twitter API. Once get to be allowed, the engineer can use the API to recover information such as tweets, retweets, likes, adherents, and more.

Another way to gather information from Twitter is to utilize web scratching instruments. Web scratching apparatuses are software programs that can consequently extricate information from web pages. To gather information from Twitter using web scratching apparatuses, one should recognize the net page containing the information and utilize the net scratching device to extricate the information from the internet page. This strategy requires a few programming aptitudes and may damage Twitter’s terms of benefit, so caution is prompted. Information can too be collected from Twitter utilizing third-party devices and services. These apparatuses and administrations give get to Twitter information through an API or web interface, making it less demanding for analysts and investigators to gather and analyze Twitter information. A few well-known third-party devices and administrations for collecting Twitter information incorporate Hootsuite, Grow Social, and TweetDeck. Ready to conclude that information collection from Twitter requires an understanding of the accessible strategies and instruments and the needs of the analyst or investigator. With appropriate information collection methods, Twitter can be a fabulous source of information for a wide run of applications.

5.2 *Data Annotation*

This section devotes the majority of its attention to the process of annotating and classifying the tweets that were chosen from the initial dataset obtained from Twitter. To get things rolling, a random selection of the gathered tweets yielded the first 10,000 tweets. Following the selection of these tweets, a group of three human annotators difficultly doled out labels to each of them by hand over the course of one and a half months. The tweets were classified as either non-cyberbullying (category “0”) or cyberbullying (category “1”) depending on their substance. The human annotators made their labeling choices by carefully considering the point by point rules that were given to them.

The creators arbitrarily chosen 10,000 tweets for this reason, and each tweet was physically labeled into one of two categories: non-cyberbullying (“0”) or cyberbullying (“1”). Three human annotators were included within the labeling prepare, which endured for one and a half months. The choice on whether a tweet included cyberbullying was based on particular rules is surveyed. Any errors found between the two annotators’ names were settled by a third annotator. After settling the inconsistencies and cleaning up the information, the creators gotten a last dataset of 10,000 labeled tweets, out of which 6,508 (65%) were non-cyberbullying, and 3,492 (35%) were cyberbullying tweets. Be that as it may, the creators watched that the labeled Twitter dataset was imbalanced, with an altogether bigger number of non-cyberbullying tweets compared to cyberbullying tweets. To address this awkwardness, the creators utilized adjusting procedures such as oversampling or under-sampling.

DEAR_RNN (Deep Evolutionary Attention-based Recurrent Neural Network) is a machine learning model that is designed to identify cyberbullying in tweets. Data annotation is a crucial step in training the DEAR_RNN model. In data annotation, human annotators manually label the selected tweets as cyberbullying or non-cyberbullying. In the case of DEAR_RNN, the annotators were instructed to label the tweets into two categories, “0” for non-cyberbullying and “1” for cyberbullying. The process of data annotation for DEAR_RNN began with randomly selecting 10,000 tweets from the Twitter dataset. These tweets were then labeled manually by three human annotators over a period of one and a half months.

A third annotator was brought in to reconcile any inconsistencies that were found during the initial phase of the annotation process. Following the resolution of the inconsistencies, the final dataset had a total of 10,000 tweets with associated labels. Out of these, 6,508 (65%) were categorized as tweets which did not include cyberbullying, while 3,492 (35%) were classified as tweets that included cyberbullying. The labeled tweets from Twitter had an uneven conveyance, with a more noteworthy number of tweets that did not include cyberbullying than tweets that did include cyberbullying. Oversampling was utilized in arrange to address the issue of unequal representation of classes, and the Engineered Minority Oversampling Strategy (Destroyed) was connected in arrange to oversample the minority course (cyberbullying tweets). In order to ensure that the dataset was representative of the

whole, the oversampling method consisted of repeatedly reproducing the cyberbullying samples. Following the application of the oversampling technique, the total number of tweets that were examined was 13,016. After that, the DEAR_RNN model was trained with the help of this balanced dataset.

5.3 *Data Pre-processing*

After the information collection and comment stages, the following step is to pre-process and clean the information to get ready it for include extraction and determination. Pre-processing and cleaning are basic steps in any machine learning assignment as they guarantee that the information is in a reasonable organize and free from mistakes, irregularities, and commotion. The DEAR_RNN data pre-processing and cleaning phase involve several sub-tasks, including:

Tokenization—The first step in data pre-processing is to split the tweets into individual words or tokens. The tokens can be obtained using various methods such as whitespace separation, punctuation separation, and regular expressions.

Spell checking—This is important because misspelled words can negatively affect the accuracy of the subsequent tasks.

Removal of URLs, user mentions, and hashtags—URLs, user mentions, and hashtags are not relevant to the analysis of the text and can be removed to reduce noise in the dataset.

Cleaning of special characters and punctuation—Special characters and punctuation such as emojis, quotation marks, and brackets can also be removed to simplify the data and reduce noise.

Encoding—Finally, the pre-processed and cleaned data is encoded into a numerical format that can be used for feature extraction and selection. This is typically done using techniques such as one-hot encoding, word embedding, or bag-of-words.

In general, the pre-processing and cleaning phase in DEAR_RNN is crucial for obtaining accurate and reliable results in subsequent tasks.

In DEAR_RNN, the data pre-processing and cleaning phase can be split into three sub-phases to make sure that the raw tweet dataset is cleaned up and turned into a final dataset. In the first sub-phase, noise like URLs, hashtags, mentions, punctuation, symbols, and emoticons are taken out. OOV words are cleaned up by doing things like spell checking, expanding acronyms, changing slang, and removing characters that are too long. The final sub-phase is made up of changes to tweets, such as changing the text to lowercase, stemming, word segmentation (tokenization), and taking out stop words. These smaller steps are needed to improve the quality of the tweets, get more accurate feature extraction and classification, and improve the quality of the tweets. By cleaning and pre-processing the data in this way, the DEAR_RNN model can work with it better, which leads to more accurate predictions.

5.4 Feature Extraction and Selection

Feature extraction and selection is a crucial step in the DEAR_RNN model as it aims to extract relevant features from the pre-processed tweet data that will aid in the classification of cyberbullying and non-cyberbullying tweets. During this phase of the process, which is known as the feature extraction phase, several kinds of features are retrieved from the twitter data that has previously been processed. Word-level features, character-level features, and semantic-level features are all included in these features. The use of n-grams and character embeddings are both utilized in the process of character feature extraction. Techniques such as topic modeling and word embeddings are utilized in the process of extracting information on a semantic level.

After the features have been retrieved, the next step is to use feature selection techniques in order to determine which characteristics are the most important and how they contribute to the classification process. Techniques for selecting features have the goals of reducing the dimensionality of the feature space, cutting down on noise. The method known as Data Envelopment Analysis (DEA) is employed inside the framework of the DEAR_RNN model in order to carry out the process of feature selection. In the DEAR_RNN model, each tweet is treated as a DMU, and the features are treated as inputs and outputs.

Extracting features in a Twitter dataset involves transforming the raw text data into numerical features that can be used for machine learning models. This process involves several steps:

Tokenization—This is the process of splitting the text into individual words, called tokens. This is usually done by splitting the text on whitespace, punctuation, and other delimiters. For example, the sentence “Hello, world!” would be tokenized into two tokens, “Hello” and “world”.

Stop word removal—Stop words are common words that do not carry much meaning, such as “the”, “and”, and “of”. These words are often removed to reduce the dimensionality of the feature space and improve performance.

Vectorization—This is the process of converting the tokenized and preprocessed text into a numerical representation that can be used for machine learning. There are several methods for vectorizing text, including bag-of-words, TF-IDF, and word embeddings.

Feature selection—This is the process of selecting a subset of the extracted features that are most relevant for the machine learning task. This can be done using various techniques, such as chi-squared test, mutual information, and feature importance scores. In a Twitter dataset, the above steps are applied to the tweets in the dataset to extract relevant features for machine learning models. The features can include various aspects of the tweets, such as the presence of certain words or hashtags, the sentiment of the tweet, the user who posted the tweet, and so on. These features can be used to train machine learning models to perform tasks such as sentiment analysis, topic classification, and cyberbullying detection.

In the context of a Twitter dataset, feature extraction refers to the process of identifying and selecting relevant pieces of information that can be used to classify tweets as cyberbullying or non-cyberbullying.

In addition to the above, POS tags can be used to identify specific types of words (e.g., nouns, verbs, adjectives) and their relationships within a sentence. There are various feature selection methods available in the literature, and one of the most commonly used methods is the Information Gain (IG) method. This method ranks the features based on their ability to provide the most significant amount of information about the classification problem. The most prominent features are then selected and fed into the DEA-RNN classifier.

Generally, feature extraction in a Twitter dataset involves identifying and selecting relevant pieces of information from text data using NLP tools. The selected features can be further enhanced by considering POS tags, function words, and content word features. Finally, feature selection methods such as Information Gain can be used to identify the most significant features for the classification task.

5.5 *Data Analysis*

In this section, we will detail how we evaluated the efficacy of the DEA-RNN model by using datasets acquired from Twitter. Specifically, we will focus on the evaluation process. We utilized a number of different models.

We additionally detail the methods that were utilized in order to annotate the data that was included in the input dataset. We chose two deep learning-based models (RNN and Bi-LSTM) and three other models from the current research on cyberbullying detection in social media to compare the performance of our proposed model with that of existing models. We found that our proposed model performed better than the existing models. We used the configuration settings of these baseline models exactly as they were described in the original papers they came from.

Python version 3.7.4 and the Pycharm IDE version 2020.2.3 were utilized for the purposes of our studies, in addition to a number of different libraries.

A number of criteria are used to assess the proposed DEA-RNN model, and it is compared to other deep learning and machine learning-based models. On a personal computer, the tests are carried out using common libraries and tools. Preprocessing operations are performed on the input dataset using the NLTK Python module in accordance with the methodology proposed in this study. Following that, the dataset is divided into training and testing subsets. The dataset is divided into three scenarios: 65:435% (Scenario 1), 72:28% (Scenario 2), and 93:7% (Scenario 3), in order to assess the effectiveness of tweet classification using each approach. The best performance of each approach is shown by the evaluation measures that are chosen. Each approach is used 20 times, the results are averaged, and then each evaluation metric is given a more accurate average value. The fivefold cross-validation method is also used.

An overview of the assessment measures that were utilized to gauge the effectiveness of the proposed model is given in this section. Each approach in this study is tried 20 times across all tests to ensure accurate results, and the average of the outcomes for each assessment metric is calculated.

In this segment, the DEA-RNN classifier’s experimental findings are appeared, at the side comparisons to other profound learning models just like the Bi-LSTM and RNN, as well as machine learning models just like the MNB, RF, and SVM. In the tests, three alternative input dataset scenarios—Scenario 1 (65:35%), Scenario 2 (72:28%), and Scenario 3 (93:7%)—were used to assess the classifier’s capacity to predict cyberbullying. In order to generate an average of the results, experiments were ran 20 times for each classifier and dataset input scenario. Each model’s performance was assessed using a variety of criteria.

The study compared the DEA-RNN model’s accuracy to that of other current models. The authors calculated the average accuracy for each scenario and discovered that Scenario 3’s DEA-RNN had the highest average accuracy, 90%. The accuracy ratings for the other models were, respectively, 83.45%, 80%, 77%, 64%, and 75%. In scenario 1, the DEA-RNN demonstration achieved an accuracy rate of 82%, outperforming the taken into consideration current models. Among all the models, Bi-LSTM had the second-highest score, however MNB had the execution result that was the worst possible.

Therefore, the accuracy of the proposed model and alternative approaches was best in Scenario 3.

Figure 1 shows the performance evaluation based on accuracy.

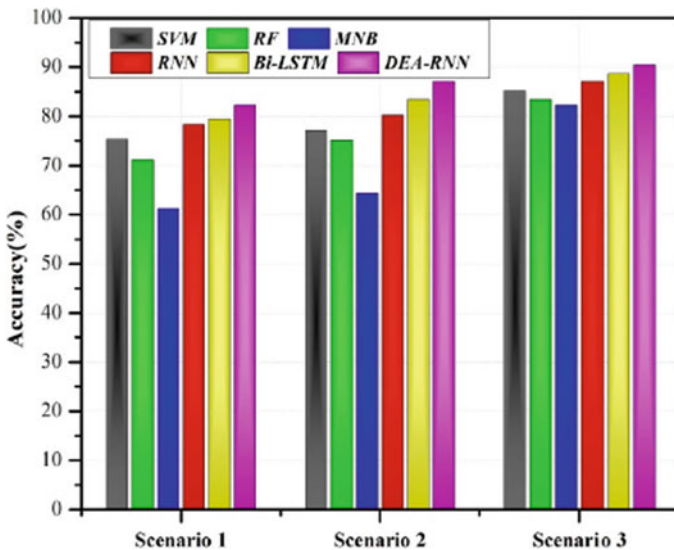


Fig. 1 Accuracy based evaluation

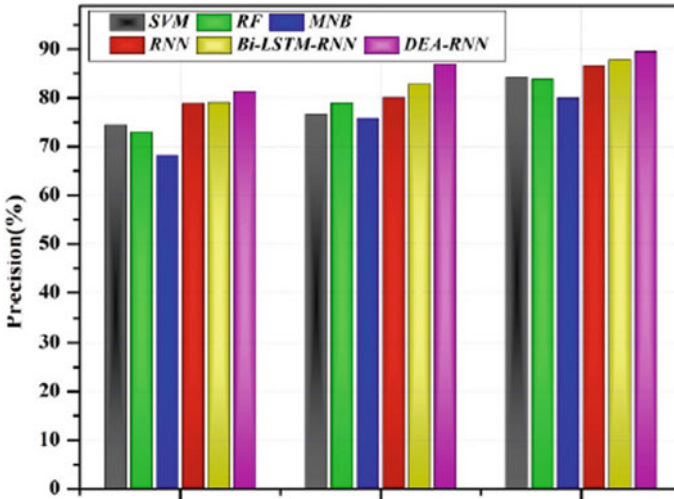


Fig. 2 Precision based evaluation

Comparing the DEA-RNN model's performance against that of other models already in use. In scenario 3, the accuracy of the DEA-RNN was 89.52%, compared to 87.9%, 86.62%, 84.25%, 80.01%, and 83.87% for others. In scenario 2, the DEA-RNN outperformed the other existing models, which attained precision values of 82.88%, 80.09%, 76.6%, 75.78%, and 78.96%, respectively. The DEA-RNN's best precision value was 87.02%. Bi-LSTM achieved the second-highest precision score among all the models in scenarios 2 and 3, while MNB had the lowest performance outcomes (Fig. 2).

In this part we tested a new model called DEA-RNN to see if it could do a better job than other computer programs at detecting cyberbullying on Twitter. They compared it to other models and tested it with different sets of data, using measures like accuracy, recall, precision, F-measure, and specificity to see how well it did. They found that the DEA-RNN model was the most accurate, with an average accuracy of 90.45% in Scenario 3, and it did better than other models like Bi-LSTM and RNN. The deep learning models worked better than the machine learning models, and the MNB model did the worst. Overall, the DEA-RNN model looked really good for detecting cyberbullying on Twitter.

6 Conclusion and Future Work

In light of the discoveries of this consideration, there are a number of distinctive suggestions and regions for future investigate that may well be taken into thought. The investigate of designs of cyberbullying on other social media stages, such as

Instagram, Facebook, YouTube, and Flickr, is one conceivable road for advance request that might be sought after. This will provide a more comprehensive picture of the patterns of cyberbullying over an assortment of stages, and it'll help find commonalities and refinements.

Utilizing information from various sources in arrange to distinguish occurrences of cyberbullying is another curiously range of request. In expansion, the substance of tweets isn't the as it were thing that future inquire about might center on; it can too examine the behavior of Twitter clients. This will give a more comprehensive understanding of the marvel of cyberbullying on Twitter and permit for the creation of cures that are more successful. In expansion, the inquire about found that the DEA-RNN demonstrate was fruitful in distinguishing occurrences of cyberbullying on Twitter based exclusively on the substance of the tweets themselves. Be that as it may, cyberbullying can moreover take put through other shapes of media, such as sound and video recordings and still photos. As a result, the recognizable proof of cyberbullying in these distinctive media sorts should be the subject of encourage investigate within the future.

The investigate proposes that real-time checking and intercession may be a strategy that has potential within the battle against cyberbullying. Tweets that constitute cyberbullying can be categorized and distinguished in genuine time with the assistance of the show that was proposed, which clears the way for mediations that are incite and effective. The experiments had been performed on distinctive input dataset situations and evaluated the use of several metrics such as accuracy, don't forget, precision, F-degree, and specificity. The consequences showed that the DEA-RNN model outperformed other fashions in terms of accuracy, attaining the best average accuracy of ninety.45% in scenario 3, followed by Bi-LSTM and RNN. The deep studying fashions completed higher than the gadget getting to know fashions, and MNB had the worst overall performance among all the fashions. The proposed version and different methods finished optimally in state of affairs three in terms of accuracy. General, the DEA-RNN model showed promising results for cyberbullying detection on Twitter.

From this study we can see that, Twitter clients can annoy one another, call one other names, spread rumors, or by and large lies, and send messages of terrorizing. These are all illustrations of cyberbullying. Cyberbullying can be coordinated at anybody, counting open figures, writers, activists, lawmakers, and indeed normal clients and celebrities.

The usage of mysterious accounts or imposter profiles may be a drift that has been watched in connection to cyberbullying on Twitter. These accounts are made in arrange to conceal the user's personality, which makes it much less difficult to irritate and bully other individuals without fear of the results of one's activities. Since of this namelessness, there's a more prominent potential for the spread of despise discourse, prejudice, and other sorts of bias.

Another later advancement is the utilize of hashtags, or "hashtags," to spread despise discourse or single out certain people or organizations. Individuals have utilized hashtags like #CancelCulture and #MeToo to irritate and scare others online, which has driven to a poisonous climate on Twitter.

In expansion, the utilize of bots and other shapes of mechanized account can intensify the impacts of cyberbullying that happens on Twitter. These accounts are pre-set to spread messages of contempt and amplify the reach of cyberbullying, both of which make it more troublesome to screen and combat the marvel. Twitter has, in response to these advancements, made an assortment of arrangements and instruments to avoid cyberbullying. These approaches and devices incorporate channels for announcing injurious substance, channels, and calculations implied to distinguish and expel such substance. However, cyberbullying proceeds to be an unavoidable issue on the stage, and on the off chance that it is to be successfully tended to, it'll require a concerted exertion effort exertion on the portion of clients, administrators, and social media firms.

References

1. Al Duhayyim M, Mohamed HG, Alotaibi SS, Mahgoub H, Mohamed A, Motwakel A, Eldesouki M (2022) Hyperparameter tuned deep learning enabled cyberbullying classification in social media. *Comput Mater Contin* 73:5011–5024
2. Al-Garadi MA, Hussain MR, Khan N, Murtaza G, Nweke HF, Ali I, Mujtaba G, Chiroma H, Khattak HA, Gani A (2019) Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges. *IEEE Access* 7: 70701–70718
3. Altay EV, Alatas B (2018) Detection of cyberbullying in social networks using machine learning methods. In: 2018 international congress on big data, deep learning and fighting cyber terrorism (IBIGDELFT). IEEE, pp 87–91
4. Dhyani M, Kumar R (2021) An intelligent chatbot using deep learning with bidirectional RNN and attention model. *Mater Today: Proc* 34:817–824
5. Haidar B, Chamoun M, Serhrouchni A (2019) Arabic cyberbullying detection: Enhancing performance by using ensemble machine learning. In: 2019 international conference on internet of things (things) and IEEE green computing and communications (greecom) and IEEE cyber, physical and social computing (cpscom) and IEEE smart data (smartdata). IEEE, pp 323–327
6. Jeevan RK, Venu Madhava Rao SP, Kumar PS, Srivikas M (2019) EEG-based emotion recognition using LSTM-RNN machine learning algorithm. In: 2019 1st international conference on innovations in information and communication technology (ICIICT). IEEE, pp 1–4
7. Liu H, Lang B, Liu M, Yan H (2019) CNN and RNN based payload classification methods for attack detection. *Knowl-Based Syst* 163:332–341
8. Muaad AY, Kumar GH, Hanumanthappa J, Benifa JB, Mourya MN, Chola C, Pramodha M, Bhairava R (2022) An effective approach for Arabic document classification using machine learning. *Global Trans Proc* 3(1): 267–271
9. Muller HA, Zavolokina L (2021) Artificial intelligence for combating cyberbullying on social media platforms: a systematic review. *IEEE Trans Comput Soc Syst* 8(4):984–999
10. Murnion S, Buchanan WJ, Smales A, Russell G (2018) Machine learning and semantic analysis of in-game chat for cyberbullying. *Comput Secur* 76:197–213
11. Murshed BAH, Abawajy J, Mallappa S, Saif MAN, Al-Ariki HDE (2022) DEA-RNN: a hybrid deep learning approach for cyberbullying detection in Twitter social media platform. *IEEE Access* 10:25857–25871
12. Paullada A, Raji ID, Bender EM, Denton E, Hanna A (2021) Data and its (dis) contents: a survey of dataset development and use in machine learning research. *Patterns* 2(11):100336
13. Perez C, Karmakar S (2023) An NLP-assisted Bayesian time-series analysis for prevalence of Twitter cyberbullying during the COVID-19 pandemic. *Soc Netw Anal Min* 13(1):51

14. Raiyani K, Gonçalves T, Quaresma P, Nogueira V (2018) Fully connected neural network with advance preprocessor to identify aggression over Facebook and Twitter
15. Yang C, Jiang W, Guo Z (2019) Time series data classification based on dual path CNN-RNN cascade network. *IEEE Access* 7:155304–155312
16. Yang L, Shami A (2020) On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing* 415:295–316
17. Zeng Z, Li Y, Li Y, Luo Y (2022) Statistical and machine learning methods for spatially resolved transcriptomics data analysis. *Genome Biol* 23(1):1–23

A Signcryption Approach to Address Security and Privacy Issues in Online Gaming Platforms



Mahmut Ahmet Unal  and Tasmina Islam 

Abstract Online game has become increasingly popular with people of all ages. But it also poses a number of security and privacy risks and children are particularly vulnerable to these risks. In order to protect children and other online gamers, it is important to address these risks and implement appropriate security measures. This paper reviews a cryptographic method, Signcryption that processes signing message by the digital signature and encrypting it by the encryption algorithm simultaneously, providing a high level of security. In this paper, the existing Signcryption schemes have been reviewed and a new scheme is proposed that is based on an extended version of Meshram's encryption algorithm. The analysis of proposed scheme shows that it is comparable to existing schemes and provides a number of additional security properties, such as unforgeability and forward secrecy.

Keywords Security · Cryptography · Signcryption · Online games

1 Introduction

In the modern era, people spend more time online. Concurrently, the incidence of cybercrime is on the rise, with its threats being regarded as a worldwide concern for every individual or group [26]. Notably, children have been spending an increasing number of hours online, more than 83% of children aged 12–15 in the United Kingdom hold a smartphone and devote more than 20 h per week using online applications and services [20]. Additionally, gaming is the third most popular activity on Microsoft's platform according to its data [31].

This usage prompts worries in public and government organisations in regards to the commercial misuse of individuals' data [16]. Mobile applications have been

M. A. Unal · T. Islam (✉)

Department of Informatics, King's College London, London, UK
e-mail: tasmina.islam@kcl.ac.uk

M. A. Unal

e-mail: mahmut.unal@kcl.ac.uk

demonstrated to be especially dangerous to people's privacy [6]. Although utilizing online platforms or websites makes people's life simpler, it requires certain measures to secure their data. Children become increasingly vulnerable to long-term privacy problems as they spend more time online [6], due to these platforms frequently sharing people's data with a variety of data trackers and ad networks. Furthermore, the private sector typically owns these data traces [19]. Data collectors' existence within third-party libraries in the services children use on a daily basis is a particularly significant, though frequently underestimated, concern for children today [22]. However, the most important reason for taking a new stand against data collection is the harm caused by long-term risks to children's reputations and chances [6].

Children, on the other hand, are unable to comprehend why their data is important to other parties [13], as they are incapable of understanding the privacy-related situations despite their presence. When there is a deficiency of either fundamental information security or cybersecurity knowledge among children and/or parents, privacy controls and parental controls emerge as beneficial [26]. However, in the study done by [28], a questionnaire was conducted with 12 children aged 8–11 years old. According to the findings of this study, kids prefer to solve problems and concerns on their own rather than express them to their parents. Children also believe that the owners of social media applications or corporations should prioritise investing more in enhancing their internet experience, ensuring greater safety and security [26].

Therefore, it appears that children could perceive the digital environment as a "personal space" for asserting their personalities and seem to be frequently concerned about paternal or maternal interference in their rights without their consent [28].

According to this research, there is still more to be addressed to secure the data of children on these platforms. Application and game developers are being given additional duties to produce suitable applications for children because of recent privacy regulations. It is shown that developers continue to rely on private information advertising networks in their development practises [6].

In essence, there are several reasons why online gaming security is necessary. The first is the client–server architecture, which has various problems and is widely distributed (most notably in massive multiplayer online games). Then there's the economy, which in many of these games can be used to make money. Also, there are the millions of online gamers who are willing to pay to play a game for an extended period [31].

Taking the client–server communication on online games as an example, cyber-attackers (also known as hackers) could abuse the security vulnerabilities on these platforms, which is dubbed 'cybercrime' in the cyber world, and children could be considered an easy target due to a lack of awareness of the security or privacy issues related to holding smartphones and other technical tools. These potential vulnerabilities have been relatively ignored by not only children, but also their parents [26]. Furthermore, cybercrime has been determined by [8] as a crime that is involved with the operating of computers and the internet, particularly computer networks, such as illegal access or interfering with information, which is kept in a database or a computer itself [31].

Despite the fact that there has been cybersecurity research dedicated to comprehending security weaknesses in the security techniques aimed at addressing such problems and computer systems [7, 29] indicated that the motivation to handle security flaws in such systems has been on enterprise-level cybersecurity functionalities or the developing technology of the IOT (Internet Of Things) [31]. However, the literature lacks sufficient research about cyber risks that are caused by online gaming [21].

Additionally, the uncertainty of what type of data is being transferred back and forth to a game service, or is being stored on database servers, might be detrimental. Thus, such connections bring privacy and security problems to the personal information and safety of gamers [31]. Furthermore, [23] have discovered that during the communication between user and server, all the data being transferred could be intercepted and used by hackers or attackers without getting any permission from them, as some automated connections between pairs can be transparent [31].

As a result, users may demand safe and secure communications for a variety of reasons, including protecting confidential information from adversaries such as hackers, possibly requiring the use of confidentiality and data integrity to prevent unauthorised access and usage. Distributing those requirements, cryptography is the most prominent approach for securing communications [24].

Factually, cryptography emerged in order to allow all parties involved in communication to maintain confidentiality of the messages that pairs send to each other even if an intruder is capable of interfering with the channel of communication. Even though ensuring privacy remains a main intention, it has evolved to offer many more advanced purposes, such as assuring communication integrity and authenticity, along with not just conventional communication security goals. Cryptography, which was exclusively operated by the military, has been extensively employed by the public and is vital to network and computer security [4].

Additionally, there are lots of different paradigms in which cryptography is involved. One of those techniques in cryptography is signcryption, by which both signing a message with a digital signature and encrypting it with public key encryption are satisfied simultaneously. Thus, signcryption accomplishes not only message confidentiality and authenticity but also dwindled computation costs compared to the signature-then-encryption algorithm [35].

Fundamentally, not only security requirements but also lower computational costs, which are essential for internet communication, are delivered by signcryption. This paper proposes a novel signcryption scheme, an extended and modified version of [18]'s encryption scheme, to address computational efficiency, security, and privacy requirements of gaming platforms involving children, with the aim to mitigate these challenges and emphasises the importance of preventive measures for individuals with limited security knowledge.

The remainder of the paper is organised as follows: Sect. 2 provides a brief literature review, Sect. 3 discusses the proposed scheme. Section 4 analyses the results and finally, Sect. 5 concludes the paper.

2 Literature Review

Zheng [34] proposed the first signcryption system by which 85% of communication overhead and 50% computational time was saved in contrast to the method of signature-then-encryption whilst also ensuring unforgeability, confidentiality, and non-repudiation. The DLP and modular exponentiation were used in developing this method [1].

Bao and Deng [3] enhanced Zheng's signcryption system, which enables any third party or judge to validate a signature in the case of a disagreement, without possessing the recipient's private key. However, the external key exchange ought to be used for the process of verification by this scheme [1].

Zheng and Imai [36] introduced the first Elliptic-Curve-Cryptography-based signcryption system that included all the essential cryptographical properties. This method is 40 percent less complex to communicate and 58 percent faster to compute than the prior signature-then-encryption scheme. This method has all the essential security properties, however, it misses forward secrecy [1].

Malone-Lee and Mao [17]'s RSA-based signcryption provides non-repudiation in a coherent way [1]. Li and Chen [14] developed an innovative proxy signature-based signcryption technique. The fundamental concept behind this strategy is to be signed the message by someone else, who is known as "proxy signer", in place of the original signer. In a variety of applications might be included proxy signature structure, such as an electronic payment system [1].

A Diffie-Hellman-based signcryption was designed by [15] that performs the original signcryption and enables key invisibility [1].

A new signcryption approach based on different digital signatures scheme called hybrid that is safe against forgery attacks performed by the person impersonated to be the original sender is proposed by [5]. The fundamental goal of this scheme is integrating a symmetric data encapsulation mechanism and an asymmetric key encapsulation mechanism. This concept is known as insider security [1].

Vanover [32] determined and built an efficient scheme based on the unique model of security for ID-based signcryption. The scheme they proposed does not necessitate employing any MAC (Message Authentication Code) function [32].

Schnoor signcryption, which is more efficient in terms of computing cost than the other signcryption schemes, is presented by [25]. Basically, he attempted to enhance earlier signcryption techniques in order to reduce communication time and overcome security flaws [1].

Amounas and Kinani [2] introduced an Elliptic Curve based signcryption scheme that can be believed to be much more effective than the other same mathematical method based techniques proposed before 2013 [1]. As most of the security attributes are offered by this scheme, such as unforgeability, confidentiality and authentication of messages, public verifiability, forward secrecy, and non-repudiation, with efficient computational cost in comparison to the traditional scheme signature-then-encryption. Furthermore, all those benefits make this method more suitable for environments that have limited computational control [2].

A nonce Elliptic curve cryptography based signcryption technique, which delivers all the security requirements, has been generated by [12] in order to reduce communication overhead for a pair, who sends the message first [9].

Many signcryption methods have been proposed and applied earlier, yet they differ in terms of the security features they offer and the computational cost they impose. In other words, every prior signcryption approaches achieve the same idea yet different algorithms so as to accomplish the fundamental goals of dwindling computational cost and communication overhead [1]. The complexity of solving particular mathematical assumptions, which are more likely to be DLP, IFP, Bilinear Diffie Hellman Problem, and ECDLP (Elliptic Curve Discrete Logarithmic Problem), assigns the ability of these methods [27]. Hence, so as to better understand the signcryption concept, signature-then-encryption scheme, signcryption and El-Gamal based Signcryption have been discussed respectively, which provides a broader view to compare the scheme proposed to the original schemes that relied on El-Gamal digital signature.

2.1 Signature-Then-Encryption

The message is signed by employing a digital signature technique with the private key of the sender, and then it is encrypted by the use of a symmetric encryption method with the secret key picked at random by the sender in order to keep the communication confidential and unforged. A public key encryption method is applied with the recipient's public key by the sender so as to encrypt this secret key in the form of an envelope. Both the envelope and encrypted text are then sent to the intended recipient by the sender [10]. Private key is used to decipher the envelope in order to get the secret key by the recipient after receiving the encrypted message and envelop, and then he uses this secret key so as to decrypt the cipher text to attain the signature and the plain text. Finally, the communication is verified by the receiver using this signature [10]. This method is dubbed signature-then-encryption scheme.

El-gamal encryption and signature-then-encryption based on DSS (Digital Signature Standard) or RSA (Rivest-Shamir-Adleman) digital signatures are only a few of the schemes that rely on signature-then-encryption [1].

2.2 Signcryption

Signcryption, a revolutionary cryptographic technology that the security concerns of digital signature, which are fundamentally integrity, unforgeability, and nonrepudiation, is applied with public key encryption by this scheme [30], was introduced for authentication and confidentiality [10]. In addition to these concerns, most signcryption techniques accomplish public verifiability and forward secrecy [30]. When compared to the signature-then-encryption methodology, signcryption is way

more effective and secure. Furthermore, the considerable decrease in communicational overhead and computational cost is the most significant feature of signcryption against signature-then-encryption [34]. The signcryption and unsigncryption processes in the most known signcryption methods meet the security requirements of distinctive unsigncryptability, efficiency, and security [1].

In the signcryption technique, the recipient's public key is used to obtain a secret key for symmetric encryption by the sender. Later, the recipient's private key is applied to generate the corresponding secret key [10].

2.3 El-Gamal-Based Signcryption

Although signature-then-encryption, which is a cryptographical technique, had been widely utilised, a digital signcryption rose to prominence as it justifies both the purposes of encryption and digital signature with a lower cost than that of the former [35]. DLP based digital signature scheme was introduced by Taher ElGamal, which is called ElGamal scheme. It produces completely dissimilar ciphertext for the same message in every encryption [1]. It has been modified and broad assumptions have been made on a variety of systems thanks to its advantages [35]. All the parameters belong to the ElGamal based digital signature scheme are shown in Table 1.

It is vital to notice that similar to hash function, KH (Keyed Hash Function) yields hashed message by hashing it with *key*, which is collusive resistant. It indicates that the hash algorithm cannot produce message pairs with the same hash value [35].

Two relatively disparate shortened signatures are allowed by implementing the technique of shortening to DSS. Table 2 provides a brief explanation of the signature and verification phases of these two methods, which are signified by SDSS1

Table 1 Parameters of digital signature schemes, obtained and adapted from Zheng [35], p-294

p	A large prime number
q	A large prime factor of $p - 1$
g	An integer with $q \bmod p$ chosen randomly from $\{1, \dots, p - 1\}$
H	One-way Hash function (SHA-1)
KH	A keyed one-way hash function (SHA-1 keyed hash function)
E	Encryption algorithm of a private key cipher
D	Decryption algorithm of a private key cipher
Alice's keys:	
x_a	Alice's private key, chosen uniformly at random from $\{1, \dots, p - 1\}$
y_a	Alice's public key ($y_a = g^{x_a} \bmod p$)
Bob's keys:	
x_b	Bob's private key, chosen uniformly at random from $\{1, \dots, p - 1\}$
y_b	Bob's public key ($y_b = g^{x_b} \bmod p$)

Table 2 Shortened digital signature schemes, obtained and adapted from Zheng [35], p-293

Shortened schemes	Signature (r,s) on a message m	Verification of signature
SDSS1	$r = hash(g^x \text{ mod } p, m) \quad s = x/(r + x_a) \text{ mod } q$	$k = (y_a g^r)^s \text{ mod } p$ check whether $hash(k, m) = r$
SDSS2	$r = hash(g^x \text{ mod } p, m) \quad s = x/(1 + x_a r) \text{ mod } q$	$k = (g y_a^r)^s \text{ mod } p$ check whether $hash(k, m) = r$

(Shortened Digital Signature Standard 1) and SDSS2 (Shortened Digital Signature Standard 2) respectively. Due to the fact that these two shortened schemes offer incontestable security and a signature shorter, they are favored to DSS [35].

It can be seen from the methods shown above that k can be computed from parameters public and r, s values, even though the equation $(g^x \text{ mod } p)$ does not precisely include the (r, s) parameters. It was the motivation behind building a signcryption technique, which is a form of shortened signature [35].

As a side note,

$$r \equiv hash(g^x \text{ mod } p, m) \text{ mod } (p - 1) \tag{1}$$

The algorithms of signcryption schemes, which were proposed by Zheng [34], based on those two shortened signatures are given below in Tables 3 and 4.

Shortened ElGamal-based signcryption and unsigncryption methodologies can be executed effortlessly, as can be perceived from the Tables 3 and 4. Furthermore, $c, r,$ and s values orchestrate the signcrypted message m from which the original message can be unsigncrypted by the recipient [35].

Table 3 Implementation of signcryption based on SDSS1, obtained and adapted from Zheng [35], p-295

Signcryption	Parameters	Unsigncryption
$x \in R\{1, \dots, q - 1\}$ $(k_1, k_2) = hash(y_b^x \text{ mod } p)$		$(k_1, k_2) = hash((y_a g^r)^{sxb} \text{ mod } p)$ $m = D_{k_1}(c)$
$c = E_{k_1}(m)$	c, r, s	$m = D_{k_1}(c)$
$r = KH_{k_2}(m) \quad s = x/(1 + x_a r) \text{ mod } q$		Accept m only if $KH_{k_2}(m) = r$

Table 4 Implementation of signcryption based on SDSS2, obtained and adapted from Zheng [35], p-295

Signcryption	Parameters	Unsigncryption
$x \in R\{1, \dots, q - 1\}$ $(k_1, k_2) = hash(y_b^x \text{ mod } p)$		$(k_1, k_2) = hash((y_a g^r)^{sxb} \text{ mod } p)$
$c = E_{k_1}(m)$	c, r, s	$m = D_{k_1}(c)$
$r = KH_{k_2}(m) \quad s = x/(r + x_a) \text{ mod } q$		Accept m only if $KH_{k_2}(m) = r$

3 Proposed Scheme

Achieving to thwart any new attack is to need to propose new techniques as hackers or attackers are able to comprehend what is the weak side of the system. Therefore, it is crucial to develop a method that is capable of divesting the chance of an attacker from retrieving systems' any vulnerable points. As it is stated in the previous sections, signcryption can not only provide more effective performance but also make the system more secure. Hence, whether perpetual advancement has been vital to keeping the system secure, whether introducing new signcryption methods or improving previous signcryption techniques [1].

Meshram [18]'s encryption scheme has been proposed for ID-based cryptosystem. However, in this study, it has been modified to some extent to be appropriate for the signcryption scheme since Meshram's cryptosystem offers more security on the encryption side, and it is relatively uncomplicated to decrypt the message. Moreover, the intricacy of resolving IFP and DLP contributes to the success of his technique.

Integer Factorization Problem Definition of the IFP as seen in [18], Def. 2.1.1) "Let $N = pq$ and $\gcd(e, \phi(N)) = 1$, where p and q are randomly selected primes. Given $y \in Z_N^*$, it is computationally infeasible to obtain x such that

$$y = x^e \text{ mod } N$$

with the knowledge of e and N ." [18]

Discrete Logarithmic Problem Definition of the DLP as seen in (Meshram [18], Def. 2.1.2) (DLP over Z_N^*). Let $N = pq$ and g be a primitive root for both Z_p^* and Z_q^* , where p and q are randomly selected primes. Given;

$$y = g^x \text{ mod } N$$

it is computationally infeasible to derive x ." [18]

Rijndael Algorithm The Rijndael Algorithm, which Dr. Rijmen and Dr. Daemen have developed, is a symmetric encryption technique that encrypts 128, 192 or 256 bits blocks utilizing the symmetric key with different sizes, such as 128, 192 or 256 bits. This algorithm processes the plaintext by deriving it into blocks and involves r rounds, which the first one is called *AddRoundKey*. The number of standard r rounds is based on the block and key lengths, such as $r = 10, 12, 14$ [11].

It has been accepted as an AES (Advanced Encryption Standard) and encouraged to be employed by NIST (the US National Institute of Standards and Technology) as it features a variety of capabilities, for instance, low computational cost, the time that has been spent during the process, flexibility, and implementation ease. Furthermore, it performs steadily in software and hardware platforms [11].

3.1 Proposed Scheme Analysis

A new signcryption approach based on IFP and DLP has been proposed and discussed in this sub section. The signcryption scheme is a modified version of [18]’s encryption scheme, specified as follows: Key Generation, Signcryption and Unsigncryption. It is assumed that after acquiring the senders’ and the receivers’ public key, anyone can prove those are valid certificates through CA (Certification Authority). The technique has been detailed in the following phases.

Global Parameters

The global parameters, which have been agreed upon by Alice and Bob before the message is sent, is shown in Table 5.

Z_N^* means that all the values are less than and relatively prime to N . As it is stated that relatively prime implies that it satisfies $gcd(x,N) = 1$.

Key Generation

Alice(Sender)

- Alice selects her secret key x_a where it is randomly selected from Z_N^* .
- Alice computes her public key as $y_a = g^{x_a} \text{ mod } N$.

Bob(Receiver)

- Bob selects his secret key x_b where it is randomly selected from Z_N^* .
- Bob computes his public key as $y_b = g^{x_b} \text{ mod } N$.

Signcryption

In the signcryption process, the sender takes the parameters, which are selected and computed from the previous parts, and follows the steps shown below.

- The sender picks the random values k_1 Secret key and z in the order $\{1, \dots, N - 1\}$.

Table 5 Global parameters

p	A large prime number
q	A large prime factor of $p - 1$
N	A large number such that $N = p * q$ with $q (p - 1)$
g	An integer with $q \text{ mod } p$ chosen randomly from Z_N^* and satisfies $g^{\phi(N)} \equiv 1 \text{ mod } N$
H	One-way hash function (SHA-1)
KH	A keyed one-way hash function (SHA-1 keyed hash function)
Ekey	Encryption process with <i>key</i> (Rijndael Algorithm)
Dkey	Decryption process with <i>key</i> (Rijndael Algorithm)

- Alice calculates the β with using her private key:

$$\beta = z^{xa} \bmod N \quad (2)$$

- Alice computes

$$a_1 = g^\beta \bmod N \quad (3)$$

- Alice computes

$$a_2 = k_1(y_b)^\beta \bmod N \quad (4)$$

- k_2 is computed by hashing k_1 using one-way hash function

$$k_2 = H(k_1) \quad (5)$$

- The ciphertext is being computed by the Rijndael encryption algorithm of the message m with the k_2

$$C = E_{k_2}(m) \quad (6)$$

- The message and the public key of the receiver are hashed with the k_2 by the SHA-1 keyed hash algorithm.

$$r = KH_{k_2}(m, y_b) \quad (7)$$

After signcryption process is completed, Alice sends a_1 , a_2 , C and r to Bob the receiver.

Unsigncryption

In the receiver's side, Bob follows the steps shown below.

- Bob obtains the k_1 solving the Eq. 8 shown below

$$k_1 = \frac{a_2}{a_1^{x_b}} \bmod N \quad (8)$$

which can be written;

$$k_1 = a_2(a_1^{x_b})^{-1} \bmod N$$

- The k_2 is obtained as in the Eq. 5,
- The message is being recovered by the Rijndael decryption process with the k_2

$$m = D_{k_2}(C) \quad (9)$$

- The receiver computes r to verify whether or not the message is valid.

$$r = K H_{k_2}(m, y_b) \quad (10)$$

- After r is generated, the receiver makes a comparison between r and r as in the Eq. 11. If those values are not equal, Bob refuses the message. Otherwise, he accepts the message.

$$r' = r \quad (11)$$

Verification of Obtaining k_1 .

To obtain the k_1 , the receiver Bob can obtain k_1 using his private key as follows.

$$\begin{aligned} k_1 &= \frac{a_2}{a_1^{x_b}} \text{mod } N \\ &= \frac{k_1(y_b^\beta) \text{mod } N}{g^\beta \text{mod } N^{x_b}} \end{aligned}$$

It is clear that

$$y_b = g^{x_b} \text{mod } N$$

Using this equation:

$$= \frac{k_1(g^{x_b \beta})}{g^{\beta x_b}} \text{mod } N$$

As it is stated above, k_1 is obtained through the Eq. 8. Since $k_1 < N$

$$k_1 = k_1 \text{ mod } N$$

In a nutshell, the signcryption scheme proposed is to assemble a signcrypted text by generating k_2 for encrypting and signing the message. It is vital to comprehend that the message and the public key of the receiver are signed with k_2 to prevent collusive attacks. The a_1 , a_2 and r , which are sent to the receiver by the sender, make the scheme more secure in comparison to the SDSS1 and SDSS2.

4 Results Achieved

Information about the results obtained from the proposed work has been discussed further. The proposed signcryption, SDSS1 and SDSS2 has been implemented and executed at the same time to allow a comparison among those schemes for measuring the full execution time and their performance. During the implementation process, text files of different sizes (from 4.79 kb to 10 Mb) have been obtained from the website (<https://onlinefiletools.com/>), which generates text files with random characters for different sizes given as an input. For each text file, the implementation has been executed using each scheme in order to obtain the time spent for full execution.

4.1 Efficiency

In this section, benefits and drawbacks of the scheme proposed has been critically examined. The exponent size has the greatest impact on the processing cost of modular exponentiation, and cryptosystems based on diverse mathematical algorithms or problems often utilise exponents of varying sizes. Hence, it is necessary to evaluate signcryption schemes that are proposed in terms of the size of exponents that are involved in both unsigncryption and signcryption phases [34], which affects the efficiency of the algorithm. Alternatively, the execution time of each scheme consumed in the computer environment could be taken as a reference for analysing computational cost.

Computational Cost

It is clear that one among the most substantial ambitions of introducing the new signcryption approach is to provide less computing cost as well as security. Although the proposed signcryption could not satisfy the low computational cost compared to the original signcryption schemes, it takes more or less the same time. As seen in the Table 6, theoretically, there are no such big differences in computational cost between the proposed scheme and the original schemes. However, it could be believed that the work proposed would be more secure since it is based on both IFP and DLP.

Table 6 Computational cost of signcryption schemes

Schemes	ENC	DEC	H	MA	MM	MD	ME
SDSS1	1	1	4	1	3	1	3
SDSS2	1	1	4	0	2	1	3
Proposed	1	1	4	0	1	1	4

Where ENC: The number of Encryption process, DEC: The number of decryption process, H: The number of one way or keyed hash functions, MA: The number of modular addition, MM: The number of modular multiplication, MD: The number of modular division, ME: The number of modular exponentiation

Performance Analysis of the Proposed Scheme

The speed of the full execution process is required to evaluate the effectiveness of the signcryption algorithm. The full execution time that the proposed signcryption process takes is measured by applying it to a chosen data set that contains text files of various sizes. The SDSS1 and SDSS2 algorithms are also employed. More process time has been taken by the proposed work. According to the result that is obtained by the implementation, the scheme proposed is not much effective than the original signcryption schemes in terms of full execution speed, which is shown in Figs. 1 and 2. These results are acquired by running the code implemented 100 times for each file and each scheme in order to obtain average execution speed. It could be significant to consider that the execution speed could change from environment to environment, which has different features.

As seen in Fig. 1, the signcryption scheme proposed has almost the same speed as the original signcryption schemes SDSS1 and SDSS2, which supports the theoretical results in Table 6. Nevertheless, as seen in Fig. 2, the proposed work achieved better result as opposed to other schemes. It can be accomplished that the rate of reduction declines with larger file sizes. Since the proposed scheme is not affected by the message size, in this case file size, while computing the keys for each side. Additionally, the work introduced in this project is more secure as it has security attributes that original schemes do not have, which is detailed in Sect. 4.2, even though it does not satisfy the computational speed.

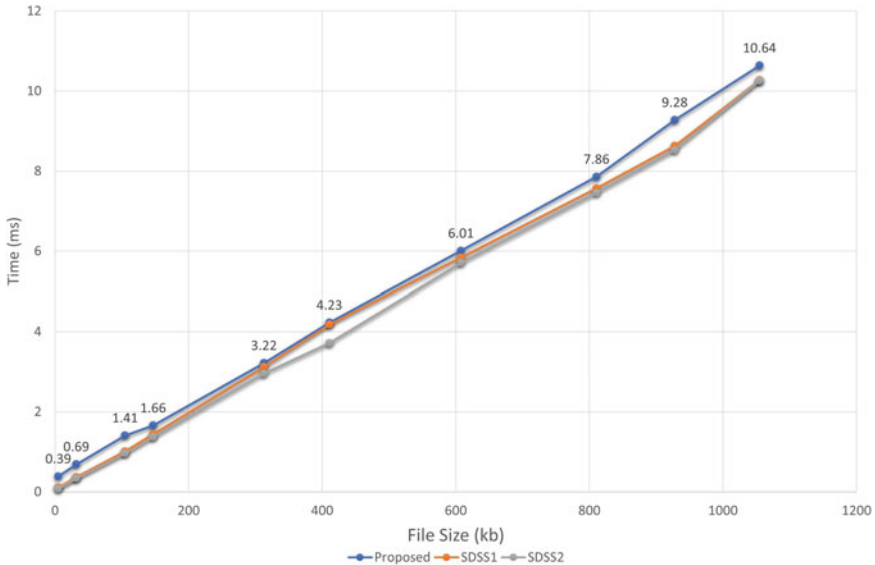
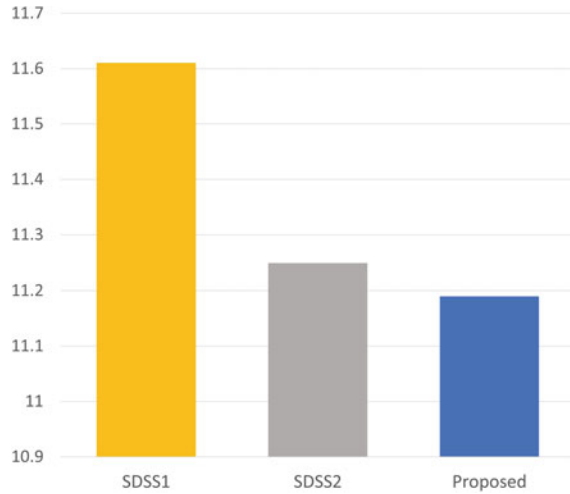


Fig. 1 Full execution time of signcryption schemes

Fig. 2 Full execution time of signcryption schemes for a single file with 10 Mb size



4.2 Security Attributes of Proposed Scheme

All security properties, which are mentioned below, have been met by the scheme proposed, which is based on the problems both IFP and DLP that are infeasible to solve. As shown in the Table 7, the proposed scheme has more security properties than the original schemes SDSS1 and SDSS2. Hence, the introduced signcryption scheme is more invulnerable to various attacks than that of the original schemes due to the fact that it distributes more security attributes.

Confidentiality

The main aim of confidentiality is to guarantee that the intended recipient of the communication receives the message. This is achieved by keeping unauthorised users from obtaining the content of the message. The attacker, Charlie, employs a number of techniques to attempt to retrieve the message content. However, these are thwarted from taking place by keeping the private key of the receiver secret. As a result, unless the private key is compromised, no one can decrypt the message and may not masquerade to be the recipient. The designated recipient performs

Table 7 Comparison of the scheme proposed with the original schemes, obtained and adapted from [9], p-55

Properties	Proposed	Sdss1	Sdss2
Confidentiality	✓	✓	✓
Integrity	✓	✓	✓
Non-repudiation	✓	✓	✓
Unforgeability	✓	✓	✓
Forward secrecy	✓	X	X
Authentication	✓	X	X

the unsigncryption operation and she is the only person who covers the message [1]. The scheme proposed provides confidentiality. That is, if Charlie intercepts the communication between sender and receiver and retrieves a_1 and a_2 , he must know the x_b (private key of Bob) so as to compute k_1 , yet it is only known and kept secret by the receiver. Additionally, the value of k_1 may be acquired by solving Eq. 8, which is computationally infeasible.

Authenticity

CA can verify the public key of pairs that are involved in the communication. CA may not be implemented in this project since it has not been deployed. However, as it is stated previously, if there is a trusted CA, verification of senders' and receivers' can take place, which means that the proposed scheme may offer authenticity by including certificate authority. In addition, it is also assumed that they agreed on global parameters to compute their public key before the message is sent to the receiver by the sender.

Integrity

The proposed scheme has been designed to ensure that message is not tampered with during the transmission even if the channel, which transits the message from the sender to the intended receiver, is not secure. Since Bob can make sure that the message m is the original message that Alice sent after receiving the signcrypted text. In the case that Charlie interferes with this communication, he may be able to modify all the values that Alice sent, such as the encrypted message C' to C , r' to r . However, Bob acknowledges this intervention since he could no longer confirm r through Eq. 10 as a consequence of these alterations. Hence, if there is any modification of the values that Alice sent, Bob would be aware that Charlie attempted to change the message during transit.

Non-repudiation

There is a trusted third party other than the pairs that participate in the communication, which accomplishes preserving the communication when disagreement occurs among the pairs. The presence of it guarantees that Bob, who is the designated receiver, has successfully received the message [1]. If there is a conflict between Alice and Bob, m is able to be validated by sending a_1, a_2, C, r to the trustworthy third party in order to agree on whether or not Alice is the sender of the message. It is supposed that the third party can determine the authenticity of m by computing r in Eq. 10 since β is generated using both the secret key of the sender and the randomly selected number z , which is only known by the sender. Therefore, Alice might be unable to refuse that message m is delivered by her.

Unforgeability

Suppose that Charlie the attacker intervened in the communication between Alice and Bob and forged a message. It must be signcrypted and sent to the receiver Bob. In this case, the receiver rejects the message as it is supposed to be delivered by the real sender. The message sender's validation achieves unforgeability, which verifies

the genuine sender and that the message was only written by this sender [1]. Intruder is unable to generate arbitrary proper values a_1, a_2, C, r starved of the Alice's private key as all the values that are created by the sender depend on the private number β , which is based on random number z and private key x_a in the signcryption process of Eq. 2. The message will not be validated because of the fact that valid values have not been produced by the attacker. Therefore, the introduced signcryption scheme in this project satisfies the security property of unforgeability.

Forward Secrecy

Providing that the attacker Charlie acquired the private key of the sender x_a , he could still need to recover the random number z , which implies that he could not get any message m from it, since the signcrypted text is computed by using random number z . Without a doubt, computing z with the knowledge of the secret key x_a , a large number N , and β is infeasible because of the fact that it involves solving IFP. Naturally, every attempt at signcrypting the message m necessitates selecting the random number z . Hence, even if the message m is the same, the signcryption scheme produces different values, which proves that the scheme proposed is of a forward secrecy.

5 Conclusion

Cyber-attacks have emerged as a myriad of difficulties, such as stealing users' personal information or bank details. Children are particularly vulnerable due to the lack of knowledge and understanding about those issues, and they are seen as an unchallenged target by cyber attackers. Despite children not always taking these risks seriously, it is crucial for security researchers to address and mitigate potential threats.

A novel signcryption scheme has been proposed in this paper in order to enhance secure communication between clients and servers, specifically focusing on the context of online game users and online game platforms. This scheme is based on a modified version of [18]'s encryption, which relies on the assumptions of DLP and IFP. As signcryption integrates both the role of encryption and digital signature into a logical one step, whilst it distributes additional security properties, such as forward secrecy. Although the signcryption scheme proposed does not have a lower computational cost than that of the original schemes SDSS1 and SDSS2, it is more secure because of the fact that it has additional security properties that those original schemes do not have. The scheme introduced in this project meets the security properties of confidentiality, integrity, authenticity, non-repudiation, forward secrecy, and unforgeability.

The importance of this study was to prompt the development of new cryptographic methods or modification of existing algorithms for targeting online game platforms extensively. Moreover, the security of the suggested scheme shows that it is resistant to a variety of security issues, which arise due to the lack of cryptographic security attributes. Additionally, the signcryption method introduced in this paper can be

extended to generalized signcryption, enabling separate deployment of signcryption, signature, and encryption schemes based on authentication requirements. This flexibility can potentially reduce computational costs when authentication is not required for all messages, files, or images.

References

1. Alzeinat AM (2017) A signcryption approach based on rabin digital signature schemes, PhD thesis, Middle East University
2. Amounas F, Kinani EH (2013) An efficient signcryption scheme based on the elliptic curve discrete logarithm problem. *Int J Inf Netw Secur* 2(3):253
3. Bao F, Deng RH (1998) A signcryption scheme with signature directly verifiable by public key. In: *International workshop on public key cryptography*. Springer, pp 55–59
4. Bellare M, Rogaway P (2005) *Introduction to modern cryptography*
5. Dent AW (2005) Hybrid signcryption schemes with outsider security. In: *International conference on information security*. Springer, pp 203–217
6. Ekambaranathan A, Zhao J, Van Kleek M (2021) Money makes the world go around: identifying barriers to better privacy in children's apps from developers' perspectives. In: *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp 1–15
7. Esteves J, Ramalho E, De Haro G (2017) To improve cybersecurity, think like a hacker. *MIT Sloan Manage Rev* 58(3):71
8. Goyal N, Goyal D (2017) Cyber crime in the society: security issues, preventions and challenges. *Res J Eng Technol* 8(2):73
9. Hasan S (2021) Generalized signcryption based on elliptic curve, PhD thesis, Capital University
10. Hwang R-J, Lai C-H, Su F-F (2005) An efficient signcryption scheme with forward secrecy based on elliptic curve. *Appl Math Comput* 167(2):870–881
11. Jamil T (2004) The Rijndael algorithm. *IEEE Potentials* 23(2):36–38
12. Kumar M, Gupta P (2019) An efficient and authentication signcryption scheme based on elliptic curves. *Matematika: Malaysian J Indus Appl Math* 1–11
13. Lapenta GH, Jørgensen RF (2015) Youth, privacy and online media: framing the right to privacy in public policy-making. *First Monday*
14. Li X, Chen K (2004) Identity based proxy-signcryption scheme from pairings. In: *IEEE international conference on services computing, 2004 (SCC 2004)*. *Proceedings. 2004'*. IEEE, pp 494–497
15. Libert B, Quisquater J-J (2004) Efficient signcryption with key privacy from gap diffie-hellman groups. In: *International workshop on public key cryptography*. Springer, pp 187–200
16. Lupton D, Williamson B (2017) The datafied child: the dataveillance of children and implications for their rights. *New Media Soc* 19(5):780–794
17. Malone-Lee J, Mao W (2003) Two birds one stone: signcryption using RSA. In: *'Cryptographers' track at the RSA conference*. Springer, pp 211–226
18. Meshram C (2015) An efficient id-based cryptographic encryption based on discrete logarithm problem and integer factorization problem. *Inf Process Lett* 115(2):351–358
19. Montgomery KC (2015) Youth and surveillance in the facebook era: policy interventions and social implications. *Telecommun Policy* 39(9):771–786
20. Ofcom U (2016) Children and parents: media use and attitudes report. Office of Communications London, London

21. Olivier F, Carlos G, Florent N (2015) New security architecture for IoT network. *Procedia Comput Sci* 52:1028–1033
22. Razaghpanah A, Nithyanand R, Vallina-Rodriguez N, Sundaresan S, Allman M, Kreibich C, Gill P, et al (2018) Apps, trackers, privacy, and regulators: a global study of the mobile tracking ecosystem. In: *The 25th annual network and distributed system security symposium (NDSS 2018)*
23. Saldana J, Fern´andez-Navajas J, Ruiz-Mas J, Viruete-Navarro E, Casadesus L (2014) Online fps games: effect of router buffer and multiplexing techniques on subjective quality estimators. *Multimedia Tools Appl* 71(3):1823–1856
24. Saleh ME, Aly AA, Omara FA (2016) Data security using cryptography and steganography techniques. *Int J Adv Comput Sci Appl* 7(6)
25. Savu L (2012) Signcryption scheme based on schnorr digital signature. arXiv preprint [arXiv:1202.1663](https://arxiv.org/abs/1202.1663)
26. Siddiqui Z, Zeeshan N (2020) A survey on cybersecurity challenges and awareness for children of all ages. In: *2020 International conference on computing, electronics & communications engineering (iCCECE)*, pp 131–136
27. Singh AK, Patro B (2017) Performance comparison of signcryption schemes—a step towards designing lightweight cryptographic mechanism. *Int J Eng Technol (IJET)* 9(2)
28. Stoilova M, Livingstone S, Nandagiri R (2020) Digital by default: children’s capacity to understand and manage online data and privacy. *Media Commun* 8(4):197–207
29. Taylor CR, Shue CA, Najd ME (2016) Whole home proxies: bringing enterprise-grade security to residential networks. In: *2016 IEEE international conference on communications (ICC)*. IEEE, pp 1–6
30. Toorani M, Beheshti A (2008) Cryptanalysis of an efficient signcryption scheme with forward secrecy based on elliptic curve. In: *2008 international conference on computer and electrical engineering*. IEEE, pp 428–432
31. van Summeren R (2011) ‘Security in online gaming’. RadboundUniversity Nijmegen: Bachelor Thesis Information Science
32. Vanover JK (2018) Exploring the cyber security improvements needed by internet game users to reduce cyber security threats, PhD thesis, Colorado Technical University
33. Yu G, Ma X, Shen Y, Han W (2010) Provable secure identity based generalized signcryption scheme. *Theoret Comput Sci* 411(4042):3614–3624
34. Zheng Y (1997) Digital signcryption or how to achieve cost (signature & encryption) cost (signature)+ cost (encryption). In: *Annual international cryptology conference*. Springer, pp 165–179
35. Zheng Y (1997) Signcryption and its applications in efficient public key solutions. In: *International workshop on information security*. Springer, pp 291–312
36. Zheng Y, Imai H (1998) How to construct efficient signcryption schemes on elliptic curves. *Inf Process Lett* 68(5):227–233

ChildSecurity: A Web-Based Game to Raise Awareness of Cybersecurity and Privacy in Children



Yang Zou and Tasmina Islam 

Abstract COVID-19 pandemic has significantly shaped the way children use the internet, leading to a notable increase in their engagement with online gaming. This raises concerns regarding their privacy and online security while playing online games. This paper aims to pinpoint prevalent security vulnerabilities children encounter through an analysis of survey data. Employing the Attention, Relevance, Confidence, and Satisfaction (ARCS) model as the study's conceptual framework, a web-based game named 'ChildSecurity' has been proposed and developed. The game addresses the critical issue of cybersecurity in the context of increased online gaming among children, by educating children about protecting themselves from potential online risks while participating in gaming activities, thus contributing to enhancing their digital safety awareness.

Keywords Cybersecurity · Privacy · Web-based game · Awareness

1 Introduction

Children now-a-days have become more reliant on the internet and personal devices and the onset of COVID-19 pandemic saw a surge in internet usage, driven by the transition to online classes. Concurrently, this has also increased the time they engage in online gaming. This increased exposure to online activities potentially opens avenues for cybersecurity and privacy vulnerabilities. These challenges have become an integral aspect of children's upbringing, given their immersion in a technology-centric world. Notably, the dynamics of player interactions within online games can present more substantial risks to players' cybersecurity and privacy compared to interactions in standalone games. Addressing these concerns, it becomes imperative for parents

Y. Zou · T. Islam (✉)

Department of Informatics, King's College London, London, UK

e-mail: tasmina.islam@kcl.ac.uk

Y. Zou

e-mail: yang.zou@kcl.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_9

187

and educators to be attuned to these risks and proactively impart knowledge to children about responsible and secure internet navigation. Strategies can include screen time boundaries, vigilant monitoring of online activities, and equipping them with skills to identify and evade potential cyber threats in this educational effort [19, 21].

Given the growing concern over threats to children's cybersecurity and privacy when playing online games, it is crucial to find solutions to this problem. Currently, there are several proposed solutions, but they do not provide a comprehensive solution to the issue. With the rise of "Big Data", online gaming has become a major online social network, and the most effective way to protect user data in these networks is through user education and empowerment [4]. This can be achieved by increasing children's awareness and knowledge of the importance of protecting their cybersecurity and privacy when using the internet. Studies have shown that training approaches that incorporate games can be effective in promoting behavioural change, as demonstrated by the research of Alotaibi et al. [2]. Therefore, it is important to focus on educating and empowering children to be able to protect themselves in the online gaming environment.

This paper proposes an educational application for children designed to incorporate gaming elements with the aim of increasing their awareness of cybersecurity and privacy risks associated with online gaming. The application is presented as a web-based platform, leveraging the inherent advantages of widespread accessibility and cross-platform compatibility offered by web browsers. Additionally, the web delivery model allows for efficient centralised maintenance and instant distribution of software updates at a cost-effective measure [10].

The remainder of the paper is organised as follows: Sect. 2 provides a brief overview of relevant studies in this area. Section 3 describes the ARCS model and the design of the application along with the perception survey questionnaire analysis. Sect. 4 describes the implementation of the 'ChildSecurity' application with evaluation. Finally, Sect. 5 concludes the paper with a summary of the study and a discussion of potential future work.

2 Literature Review

Researchers have been studying to get an understanding of the risks to which children are commonly exposed during their internet usage, as discussed in a previous review [19]. However, a predominant portion of these studies has focused on specialised themes such as anti-virus technology, spam, encryption, and other related subjects [18, 19]. But much less attention has been paid towards the potential dangers that children may encounter while participating in online video game activities.

Researchers have been working on increasing 'Cybersecurity Awareness'—a way of educating internet users to be alert to the many cyber risks and the susceptibility of computers and data to these attacks [20]. In other words, [10], this is the level of understanding of users regarding the significance of information security, as well as the responsibilities and actions they must take in order to maintain adequate levels

of information security control in order to safeguard an organization's data and networks. is referred to as their "cybersecurity awareness." For children, the most important aspect of teaching about cyber security awareness is to make them more aware of the potential risks and how they can protect themselves. This can include use of educational training movies to improve children's privacy literacy [7]. The age of the child is an important factor to consider when assessing the effectiveness of this type of training, as younger children have a more challenging learning curve. Efforts have also been made to create engaging educational materials, such as the e-book "CyberHeroes" [24]. However, studies have shown that the presence of multimedia elements in educational materials may divert children's attention and reduce their understanding [6]. Additionally, the book's content may be perceived as uninteresting and daunting for children.

Some studies have reported attempt to use the format of boot camps, such as "Cyber boot camp" [3], to organize learning activities for cybersecurity education. These types of boot camps have been shown to be successful in some cases, with participants reporting improvements in their engagement with cybersecurity and their sense of self-efficacy in the field. However, this type of intervention requires a significant investment of time and resources. Similarly, training camp activities such as "Gencyber" [9] offer children in grades K-12 the opportunity to gain practical experience in cybersecurity, but are only held in certain regions of the United States and the materials taught are not widely distributed. The cybersecurity GUIDE app [1] is an Android application that provides assistance from cybersecurity professionals. The app received positive feedback from users but is no longer available for download and primarily targets young children and adolescents in India.

Some experts have also suggested that gamification may play a role in the future of cybersecurity education. The strategy of learning through play offers a dynamic approach that is both entertaining and participatory, allowing players to gain new skills and engage in new mental processes. Video games such as CyberCIEGE [22], which simulates the management of resources within an organization, can provide users with the opportunity to make decisions that impact the safety of corporate assets and learn how to prevent potential cyber attacks. Additionally, Bioglio et al. [4] propose a gamified web application, Social4School, as a way to increase young people's understanding of privacy procedures in online communities.

Overall, the field of children's internet safety is an important and necessary area of research. While there have been studies conducted on specialized themes such as anti-virus technology and encryption, there remains a gap in the literature regarding the potential dangers that children may encounter while playing video games online. Further research in this area is crucial in order to fully understand the risks and develop effective strategies for educating children about internet safety. Additionally, the selection of appropriate educational materials and methods should be carefully considered and evaluated for their effectiveness in increasing children's understanding and awareness of internet safety.

3 Survey Analysis and Design Specification

A survey titled “Perception of Parents: Security and Privacy Risks in Online Gaming Platforms for Children” has been conducted to gather information on parents’ perceptions of security and privacy risks in online gaming platforms for children. The survey was made available through the WeChat app and targeted at parents who have children between the ages of 12 and 16. A total of 72 responses have been received over a month period. All necessary ethical approval has been obtained from the institution to conduct the survey. The questionnaire flow diagram is shown in Fig. 1. The questionnaire was divided into 4 main themes: malicious links, cyberbullying, sensitive information, and password security. These themes were selected based on the assumption that the majority of parents’ children have been exposed to risks in these areas at some point. The questionnaire consisted of a mix of multiple-choice and open-ended questions to gather both quantitative and qualitative data. The data collected from the questionnaire have been used to identify the areas of concern for parents and guide the development of the game’s design direction.

The analysis of the survey responses revealed that about 68% parents mentioned that their child has opened malicious links while playing online games (as shown in Fig. 2). In response to the details of this question, they mentioned about pop-up windows that appear in the lower right-hand corner of the gaming interface while the child is playing the game. There are certain cartoon designs that are designed to entice toddlers to click and modify them. There are also other possibilities, such as providing a link for you to use to sell your account. Someone takes on the identity of a buyer and purchases your account. When you attempt to withdraw cash, you are informed that there is an issue with your bank card and that the monies need to be unfrozen before you may withdraw them. Parents also stated that their children had accessed several malicious online interfaces, that they had trouble leaving the game after they had entered it, and that their children continued to export internet infections. Some parents are also quick to point out that a number of the links in their browser to download games are not genuine, but rather lead to games that are either fraudulent or strangely connected to other games. When youngsters download one application, a large number of other unnecessary programmes are installed on the computer. When a kid plays programmes on the website that he or she is unable to download, the child may also encounter fraudulent pop-ups such as gaming activity top-ups that show on the screen. A child was playing a game as they normally would when suddenly a website appeared on the screen. It said that the account was locked and needed to be unlocked, and that the child should click on some unknown links to continue. The child later discovered that his account had been taken by another player. When youngsters click on certain malicious links, they are led to believe that they would get virtual goods; nevertheless, their accounts are really taken over by cybercriminals.

In response to the issue of cyberbullying, the majority of respondents reported that they had experienced it to some degree at some point in their lives. Less than 10 percent (shown in Fig. 2) of respondents claimed to have never been bullied online.

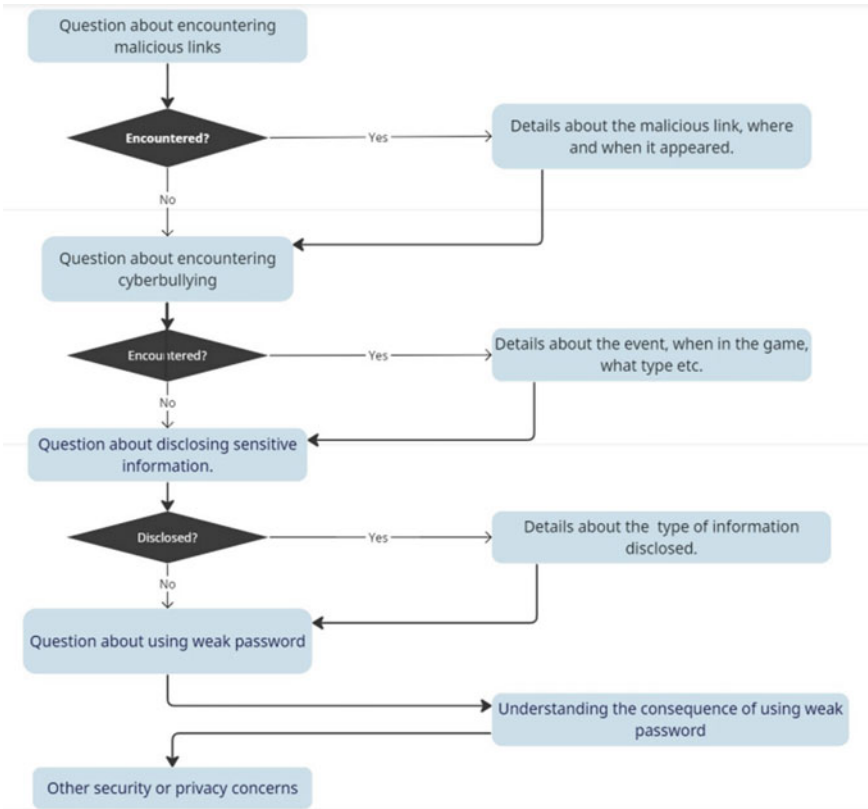


Fig.1 Questionnaire questions flow diagram

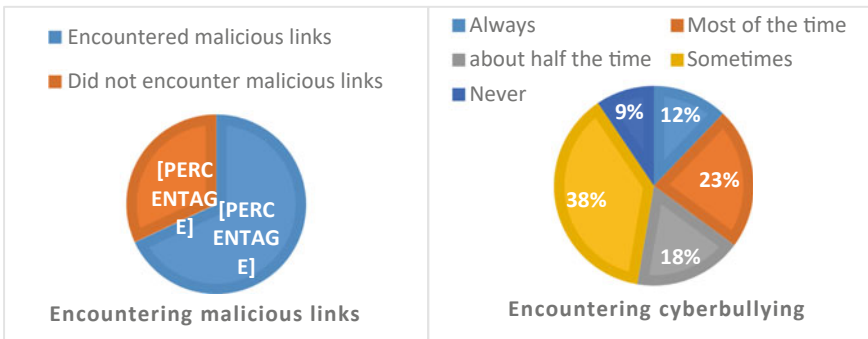


Fig. 2 Response to questions if they have encountered malicious links (left) and cyberbullying (right)

Therefore, the application plans to assist teenagers in understanding the concept of cyberbullying and the associated risks. Given that all concepts and explanations can be dry, the goal is to create a game where players spend in-game gold coins. As the user spends gold coins, they will have a deeper understanding of the topic and a final test will assess their level of understanding. It is hoped that through this approach, children will be better equipped to recognise instances of cyberbullying in their online gaming experiences, thereby reducing the negative impact of cyberbullying on children.

As the analysis of the survey responses revealed that the theme of malicious links was of particular concern for parents, the proposed application for children who play online games, will need to educate them on how to recognize malicious links. The initial step in the application would be to provide users with an explanation of what links are, followed by information on what constitutes a harmful link. This would then lead into a simulated version of a fake operating system, featuring a brief game. Users would be prompted to create an account for the game and log in to begin playing. A timer would be established and after a certain amount of time, a malicious warning message would appear on the screen, informing the user that their account has been locked and requesting that they click a button to unlock it. Upon clicking the button, the user's account password would be immediately generated, and the user would be informed of the consequences of clicking potentially dangerous links.

In response to the question about disclosing personal information about 62 percent of respondents (as shown in Fig. 3) confirmed that their child has at some point disclosed sensitive information, such as phone numbers, real names, ages, home addresses, and email addresses. These findings suggest that many young individuals may not fully understand the implications of sharing such personal information. As such, the proposed application aims to educate users on the concept of sensitive information and the risks associated with its disclosure.

As shown in Fig. 3 about 65 percent of respondents said that their children attempt to use insecure passwords for their online accounts. In addition, around three fourth

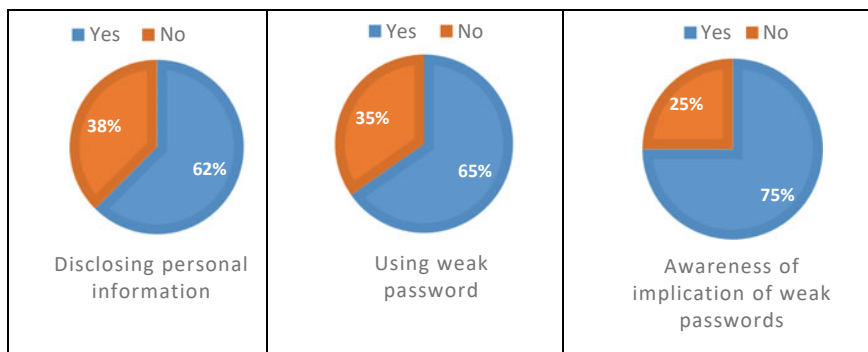


Fig. 3 Response analysis to questions if they have disclosed personal information (left), use weak password (middle), are aware about the implications of weak passwords (right)

of youngsters do not understand or aware of the implications using weak passwords for online accounts. Consequently, it is necessary for youngsters to have an understanding of what a weak password is and the implications of using it. Therefore, a simulated operating system is proposed to be constructed. This system will encompass a login interface mimicking that of some popular games, along with a password-cracking tool commonly employed by security researchers. This programme provides users with extensive instructions on how to employ the password-cracking tool for attempting the compromise of a weak password. This immersive experience aims to provide a clear understanding to young users about the inherent risks associated with employing easily breakable passwords that do not ensure an adequate level of security.

From the analysis of all the survey responses, it becomes evident that the cybersecurity concerns expressed by parents can be succinctly categorized into the four previously mentioned topics: malicious links, cyberbullying, safeguarding sensitive information, and password security. Consequently, educating young individuals on these four topics serves as a comprehensive strategy for instilling online safety awareness. Through the completion of this program, children will acquire the knowledge required to protect themselves to some extent from potential cybersecurity threats while engaged in online gaming.

4 Conceptual Framework

Motivation is a useful conceptual framework for interpreting the behaviour of learners. When it comes to changing or improving learning outcomes, one of the most important factors that is regarded to be at play is motivation, both internal and extrinsic [12, 13]. An Instructional Design Model (IDM) may be used to remedy this situation [11]. In the context of IDMs, there are a number of different instructional models available in the literature. Some examples of these models are Isman model [14], Flipped learning [17], ICCEE model [5], ARCS, and others. The ARCS model is a framework for the creation of motivating learning that incorporates the incorporation of several motivational ideas and theories [23]. Attention, relevance, confidence, and satisfaction are the four elements that may be used to classify the company's instructional approaches [15]. If instructors are able to successfully organise teaching materials and procedures in line with the ACRS model, then this model may help them develop courses or enhance teaching and make learning process more motivating.

Figure 4 illustrates the connections that are made between the structural parts of ChildSecurity and the components of the ARCS system. Attention, Relevance, Confidence, and Satisfaction are the four main components that make up the ARCS model, and they are responsible for fostering and maintaining motivation throughout the learning process. This can be seen clearly in the image.

The success of self-directed learning motivation is contingent on the presence of each of these components, each of which plays a significant part in the learning process. In the following paragraphs, we will go into more depth on the manner

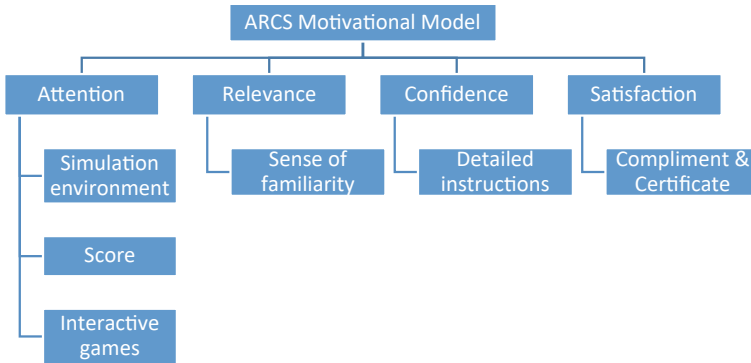


Fig. 4 Interconnection between ChildSecurity and ARCS model

in which the aforementioned characteristics are included into the design process of ChildSecurity.

4.1 Attention

The ARCS motivating model outlines the essential steps that need to be included in a learning approach in order to maintain a learner’s focus while they are in the process of acquiring new information. This trait has been shown to be one of the most important components in ARCS since it is necessary to maintain the learner’s level of interaction at a high level in order to hold the learner’s attention throughout the process of learning. Figure 4 demonstrates that the subcomponents known as “Simulation environment”, “Score”, and “Interactive games” are responsible for achieving this goal.

Users are made aware of the hazards posed by cybersecurity threats by way of a simulated fake operating system. Scores for eating gold coins are also presented in the interactive trivia game for eating gold coins, which displays scores for eating gold coins. All of these characteristics have the potential to greatly boost user attention and play an essential part in making the learning process more manageable.

4.2 Relevance

The ARCS approach also relies heavily on a concept known as relevance. According to the model, which is shown in Fig. 4, the student has to have a clear understanding of why it is important to complete this course and how it is related to issues and circumstances that occur in real life. The technique takes into account both of these prerequisites in some way. When you open the software and choose the “about” tab,

you will get a comprehensive explanation of the application as well as the primary investigation. Users are also able to comprehend their very own objectives and values in relation to the utilisation of the application. In addition, the topics covered in ChildSecurity training are very applicable and are the product of an examination of the viewpoints expressed by parents. The training components of the software consist of the user going through situations that they have encountered in the past, which demonstrates a high degree of relevance.

4.3 Confidence

Confidence may be gained from the program's tiny games, as shown in Fig. 4, thanks to the in-depth operating instructions, which can be found in the application. To be more specific, it is of the utmost importance for the player to have the feeling that he is capable of effectively completing a job. If a user has already played a minigame and wishes to play it again, all they have to do is click the "play again" button once they have finished playing the minigame. This also implies that the user is able to relearn the subject in order to answer the quiz after each topic, in the event that they do not comprehend the question being asked during the quiz after each topic. The user's chances of effectively learning the topic are increased when detailed explanations of the principles associated with each subject as well as instructions on how to play the game are provided. Keeping these considerations in mind, ChildSecurity has established distinct learning goals with the intention of acquainting users with the typical cybersecurity dangers that are present while online gaming. To be more specific, the learning goals linked with the software may be summed up as follows:

- In order to prevent their online gaming accounts from being misused, children should be able to determine whether or not strange links that they come across while playing online games are harmful.
- It is important for children to have an understanding of cyberbullying in order to safeguard their physical and mental health while using the internet.
- Children should be aware that their actual personal information cannot be freely shared in online games, and that this is a problem of personal privacy; this is a concept that should be communicated to them.
- It is important for children to be aware of the reasons why insecure passwords are weak passwords.

4.4 Satisfaction

The ARCS model is complete with this very final component. It is believed to be the most important factor in maintaining motivation, since if the user is happy with the outcomes of their learning, then it is very probable that they will want to play these games again in the near future. The end goal of the application is to provide the

learner with a “ChildSecurity certificate,” which verifies that she is knowledgeable about and skilled in matters pertaining to security.

5 ChildSecurity Implementation

As mentioned in the previous sections, internet use is increasingly prevalent among young people in today’s society. This is largely attributable, among other things, to the proliferation of popular multiplayer online games. As a result of this circumstance, education on cybersecurity for teenagers has become a significant and urgent problem. In this light, we have developed an application called ChildSecurity, which aims to provide children with an approach to learning about data security challenges and developing security awareness that is not only enjoyable but also more productive. In a nutshell, the objective of the game that is described in this article is to acquaint children with fundamental cybersecurity technologies. These technologies are necessary in order to protect children’s online security and privacy against legacy threats such as malicious links, cyberbullying, sensitive information and password security.

The learner is given the option of choose a different subject to study before beginning the accompanying minigame. Following the successful completion of each subject, the associated topic in the skill tree will be lighted. This will serve to give both inner and extrinsic incentive for the learner. The last step in the process is for the student to complete all of the tasks inside the game in order to get their “ChildSecurity certificate.” The ChildSecurity app provides the learner with instructions on how to play each fast session game and notifies them of the current learning objective before they begin playing a minigame. It should be emphasised that the programme was not developed with the intention of supplying the student with high-quality reading material for an extended period of time. It is common knowledge that students become disinterested and bored when they are required to read an entire passage and then respond to a series of questions [8]. As a result, our design stands in contrast to traditional methods, which demand of students that they complete an entire reading process and afterwards answer a set of questions.

The learner is provided with instructions on how to play every minigame and informed of the current learning objective before the app actually begins the minigame. For each and every security-related learning subject, there is a full explanation of the idea, followed by something like a mini-game that demonstrates the unique dangers, and by a quiz that tests whether or not the user has learned the issue. If the user gives an inaccurate response, the question may be asked again and again until the user gets it right. In a setting like this, the user is encouraged to not only comprehend the many different ideas that are being sent, but also recognise them while they are engaged in the activity of playing online games.

HTML, CSS, and JavaScript were the specific technologies that were used throughout the creation of ChildSecurity. ChildSecurity is capable of running on a range of major desktop platforms as a result of its use of a merged technological

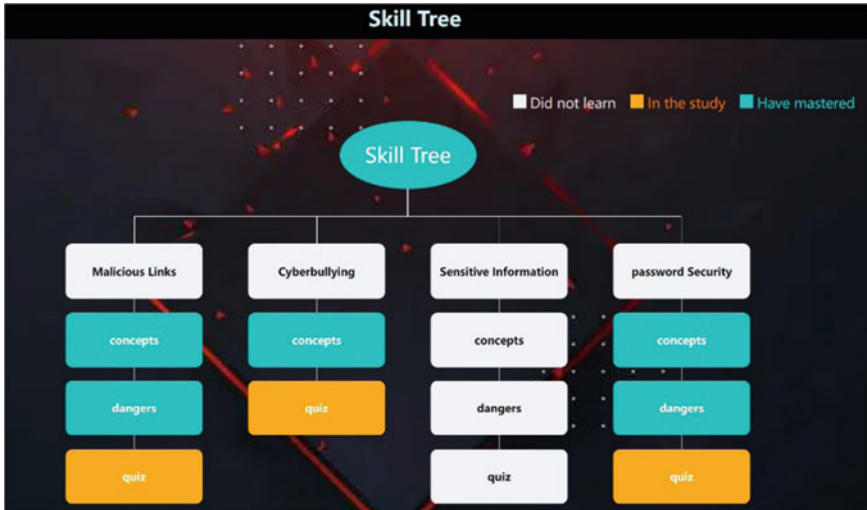


Fig. 5 Skill tree

stack. These platforms include Windows, Linux, and iOS. The code for ChildSecurity is available for download from github [25].

5.1 Skill Tree

Figure 5 depicts the page that may be found in the skill tree. White is shown in the colour block that corresponds to a subject when the user is not currently exploring that topic. The colour of the block changes to orange as soon as the user begins to browse the subject matter. The colour of the block will change to blue after the user has reached the end of the learning process, at which point they will have achieved mastery of the subject matter.

5.2 Malicious Links

Understanding the notion of a URL is the first step in the training process for this subject, which begins with youngsters. The next step after this is to get familiar with the types of URLs that are regarded as harmful. A gluttonous snake-themed quiz game was developed to simulate the experience of clicking on a malicious link. In order to participate in the game, users were forced to register for an account. As soon as the user begins participating in the trivia game, the computer software will automatically start a timer. After around 5 s, a fraudulent pop-up window will appear, indicating

that the currently active account has been locked. `unlock-account.com` is a malicious website that, when visited, causes a modification to be made to the user's previously registered password for their gaming account. In the realm of cybersecurity, this particular technique is referred to as a Cross Site Request Forgery vulnerability (CSRF). An adversary may assume the identity of a user and submit fraudulent requests using CSRF in order to steal the victim's identity [16]. Sending emails and messages, hijacking accounts, adding system administrators, and even buying items and transferring virtual money are all examples of activities that fall under this category.

5.3 Cyberbullying

Because the information is presented in the form of lengthy paragraphs of written material, it might be difficult to motivate youngsters to read while conducting training on the subject of cyberbullying. For the purpose of this study, a gold coin eating game was developed. The game has a pop-up text that discusses cyberbullying on each occasion a gold coin is consumed. The paragraphs focus on the nature of cyberbullying as well as possible responses to it.

5.4 Sensitive Information

It would indicate, on the basis of the analysis of the findings of the questionnaire, that it is important for children to have a clear understanding of what information is private or sensitive to them. The children need to be made aware of the risks associated with divulging such information to others they have never met in an online setting.

5.5 Password Security

When it comes to using weak passwords, the first thing that youngsters need to do is learn what kinds of passwords are considered weak. The next step has the children enter a mock operating system that is simulated so that they may experience how a password cracking programme can break a weak password. Children will get a better understanding of the consequences of using weak passwords as a result of this cracking experience and will be better able to avoid using weak passwords for future online games.

5.6 *Evaluation*

The most notable feature of this programme is that it can operate on any platform that has a browser. This means that you may begin your workout routine with this application as long as the device that you are using possesses a browser. In addition, this application has gaming components and simulated operating system, both of which entice children to engage in the actual hacking process and, as a result, increase their level of comprehension about the subject matter covered in the training. In conclusion, the application includes a section for quizzes, which serves not only as a test of the user's knowledge but also as a gauge of how well they have retained it.

The programme begins with a screen called "skill tree" that provides the user with a visual representation of the skills that may be obtained via training. Children have the option to train themselves in any of the skills that are shown to them by clicking on the one that most interest them. The application mimics a genuine computer's user interface so that the process of teaching children about cybersecurity risks may be presented in a way that is more engaging for them. This strategy instils in the user the desire to continue their education and maintains their concentration throughout the process of training new skills. This indicates an improvement in the effectiveness of the learning process. In a nutshell, this training application offers a more forward-thinking and instructive approach to the task of educating children on how to handle some of the potential threats to their cybersecurity that they may face while participating in online gaming.

6 **Conclusion**

This paper presented a web-based application aimed at educating young individuals about the significance of safeguarding their privacy and cybersecurity during online gaming activities. While existing research focuses on teaching children to be mindful of cybersecurity in general, there has not been much attention on specifically addressing cybersecurity issues encountered while playing online games. Using this application children will increase their awareness and protect themselves against the cyber risks that they may encounter when playing online games. In addition, the method by which this application simulates the operating system to provide the user with the opportunity to experience some cybersecurity risks or vulnerability procedures is a relatively recent development. This approach contributes to enhancing users' understanding of potential threats, thereby augmenting the training process.

References

1. Agarwal C, Singhal A (2017) Securing our digital natives: a study of commonly experience internet safety issues and a one-stop solution. In: Proceedings of the 10th international conference on theory and practice of electronic governance, pp 178–186
2. Alotaibi F, Furnell S, Stengel I, Papadaki M (2016) A review of using gaming technology for cyber-security awareness. *Int J Inf Secur Res (IJISR)* 6(2):660–666
3. Amo LC, Liao R, Frank E, Rao HR, Upadhyaya S (2018) Cybersecurity interventions for teens: two time-based approaches. *IEEE Trans Educ* 62(2):134–140
4. Bioglio L, Capecchi S, Peiretti F, Sayed D, Torasso A, Pensa RG (2018) A social network simulation game to raise awareness of privacy among school children. *IEEE Trans Learn Technol* 12(4):456–469
5. Chen LL (2016) A model for effective online instructional design. *Literacy Inf Comput Educ J* 6(2):2303–2308
6. Dalla Longa N, Mich O (2013) Do animations in enhanced ebooks for children favour the reading comprehension process? a pilot study. In: Proceedings of the 12th international conference on interaction design and children, pp 621–624
7. Desimpelaere L, Hudders L, Van de Sompel D (2020) Knowledge as a strategy for privacy protection: how a privacy literacy training affects children’s online disclosure behavior. *Comput Hum Behav* 110:106382
8. Dong C, Cao S, Li H (2020) Young children’s online learning during covid-19 pandemic: Chinese parents’ beliefs and attitudes. *Child Youth Serv Rev* 118:105440
9. Dryer A, Walia N, Chattopadhyay A (2018) A middle-school module for introducing data-mining, big-data, ethics and privacy using rapidminer and a hollywood theme. In: Proceedings of the 49th ACM technical symposium on computer science education, pp 753–758
10. Gellersen HW, Gaedke M (1999) Object-oriented web application development. *IEEE Internet Comput* 3(1):60–68
11. Gibbons AS, Boling E, Smith KM (2014) Instructional design models. In: Handbook of research on educational communications and technology. Springer, pp 607–615
12. Haque MF, Haque MA, Islam M, et al (2014) Motivational theories-a critical analysis. *ASA Univ Rev* 8(1)
13. Hodges CB (2004) Designing to motivate: motivational techniques to incorporate in elearning experiences. *The J Interact Online Learn* 2(3):1–7
14. Isman A (2011) Instructional design in education: new model. *Turkish Online J Educ Technol-TOJET* 10(1):136–142
15. Keller JM (1983) Motivational design of instruction. *Instructional design theories and models: an overview of their current status* 1(1983):383–434
16. Kombade RD, Meshram B (2012) Csrfl vulnerabilities and defensive techniques. *Int J Comput Netw Inf Secur* 4(1):31
17. Lee J, Lim C, Kim H (2017) Development of an instructional design model for flipped learning in higher education. *Educ Tech Res Dev* 65(2):427–453
18. Pinter AT, Wisniewski PJ, Xu H, Rosson MB, Carroll JM (2017) Adolescent online safety: moving beyond formative evaluations to designing solutions for the future. In: 16th international ACM conference on interaction design and children, IDC 2017, Association for Computing Machinery, Inc., pp 352–357
19. Quayyum F, Cruzes DS, Jaccheri L (2021) Cybersecurity awareness for children: a systematic literature review. *Int J Child-Comput Interact* 30:100343
20. Rahim NHA, Hamid S, Kiah MLM, Shamshirband S, Furnell S (2015) A systematic review of approaches to assessing cybersecurity awareness. *Kybernetes*
21. Read JC, Markopoulos P (2013) Child-computer interaction. *Int J Child-Comput Interact* 1(1):2–6
22. Thompson M, Irvine C (2011) Active learning with the {CyberCIEGE} video game. In: 4th workshop on cyber security experimentation and test (CSET 11)

23. Tsai CY, Shih WL, Hsieh FP, Chen YA, Lin CL, Wu HJ (2022) Using the arcs model to improve undergraduates' perceived information security protection motivation and behavior. *Comput Educ* 181:104449
24. Zhang-Kennedy L, Abdelaziz Y, Chiasson S (2017) Cyberheroes: The design and evaluation of an interactive ebook to educate children about online privacy. *Int J Child-Comput Interact* 13:10–18
25. Zou Y (2022) Childsecurity. <https://github.com/ezy0812/kclChildSecurity.git>

Subverting Biometric Security Using 3D Printed Biometrics Characteristics



Bobby L. Tait 

Abstract 3D printing technology is making major strides into various manufacturing domains, allowing for rapid prototyping and even manufacturing of novel and task specific hardware. Various 3D printing technologies are available for an exceedingly affordable price, allowing any user to purchase 3D printers and create any object that the imagination allows. This paper asks the question if the 3D printing technology can be utilized to create objects which can facilitate in subverting biometric authentication systems? The first aim of the research presented in this paper is to test in what way commonly available 3D printing technology can be used to create fake biometric characteristics. Secondly, tests are presented which evaluate if these fake 3D printed characteristics can successfully be used to subvert various biometric systems. It was found that it is indeed possible to use various 3D printing approaches to generate fake biometric characteristics that successfully subvert biometric authentication systems. Though commercially available 3D printing technology was used, the paper concludes with proposing future research investigating 3D Bioprinting, allowing for actual organs to be 3D printed.

Keywords 3D printing · Finger prints · Authentication · Biometrics · 3D Bioprinting · Security

1 Introduction

Biometric technology is often suggested to allow humans to effortlessly authenticate themselves in various environments, and biometric technology is by no means a novel technology, as mentioned in the 2001 IBM systems Journal [27]. Even though this paper by IBM, published in 2001, seems dated, biometric technology, used to authenticate humans, are much older and can be traced back to times when the Bible

B. L. Tait (✉)
University of South Africa, Pretoria, Gauteng, South Africa
e-mail: taitbl@Unisa.ac.za

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_10

203

was written [30]. Aspects related to biometrics pertinent to this paper will briefly be discussed in Sect. 2.1.

To secure a system, five security services are typically considered [3]. These five security services are defined as Identification, Authentication, Confidentiality, Integrity and Non-Repudiation. It must be noted that Authentication is in essence the gatekeeper to all the other security services. If the system fails to authenticate the object, it is not certain if a secret is confidential. If Authentication is subverted, the integrity of data is in question, as it is unsure if changes made to data is done by the person authorised to make the changes. The Authentication service is the service that ensures that only authorized objects are allowed to make updates to data, or that access to secret information is only allowed to objects that have rightful access to that data.

It is clear that if one desires to subvert a secure system, a way to subvert the authentication phase of a security system needs to be found. The most commonplace used authentication method is still where a secret number, a word or a phrase, is offered for authentication [33]. However, using a secret that only an object or an user knows, such as a password, has shown in various research tests to be subverted in various ways [7, 10, 16].

On the other hand, if a biometric authentication mechanism is utilized to secure an environment, various test have shown that biometric systems can also be subverted [21, 22].

3D printing technology, if compared to biometric technology, is a fairly novel technology. 3D printing makes it possible to generate a real world object using various methods, and is discussed in Sect. 2.2.

The research presented in this paper investigates in what way commonly available and affordable 3D printing technology as available to the man on the street, can be utilized to create physical objects to subvert a biometric authentication systems? To prove the concept, fingerprint biometrics will be used, as this technology is in commonplace in various devices such as Smartphones [29], Laptops [18] and physical access control devices [15].

The remainder of this paper is compiled in the following way: Sect. 2 introduces the background of the technologies used during this research, followed by Sect. 3 which describes the experimental setup to test the ability to subvert biometric fingerprint systems using 3D printed artefacts. Section 4 provide insight into the outcomes of the experiments, followed by the last section, which concludes the paper and discuss future research.

2 Background

This section introduces the various technologies that is applicable to the research conducted and presented in this paper. To this extent, biometrics is discussed briefly as a technology. However as the experiments in this paper were conducted using fingerprint biometric characteristics, Sect. 2.1.2 provides fundamental information

related to fingerprint biometrics and provides a brief discussion of the aspects related to fingerprint biometrics as an authentication technology.

3D printing as technology is briefly discussed in Sect. 2.2. Fused deposition modelling and Vat polymerization are the two 3D printing technologies used in this research. Fused deposit modeling is discussed under Sect. 2.2.1 and Vat polymerization is discussed under Sect. 2.2.2.

2.1 Biometrics

The term “biometrics” is derived from the Greek words “bio” (life) and “metrics” (to measure). The utilization of biometrics is by no means novel, Identification and authentication of a human by means of their physical traits can be traced back thousands of years. One of the oldest and most basic examples of a biometric characteristic that is applied for identification and authentication, is the human face. Since the very early times of civilization, humans have used faces to identify known and unknown individuals. This basic biometric recognition worked well in small communities and small villages, but as populations grew, and travel to distant places became common, recognition and authenticity of a individual became increasingly difficult. Also keep in mind that people are not just recognized by their face, but also by means of their voice or even the way they walk. All of these factors contributed to a unique biometric characteristics that makes up the uniqueness of a human being.

2.1.1 Biometric Technology and Considerations

In the biometric domain implemented in computing environments different names are often used to describe the state of a biometric identifier. For example, in many cases an author would refer to “biometric token” [12], however this term is ambiguous, as it is not clear to which biometric state such a name refers to. Is this the actual biometric characteristic (or trait) which is part of the human, or is this the biometric characteristic (or trait) as digitized by the biometric device? For this reason the following terminology is used throughout this paper, to clearly specify the biometric state being referred to [30].

Biometric characteristic. The word “biometric token” is often used to describe the physical biometric characteristic. However, the word token is associated with a man made article such as a RF-id card, magnetic card or special key. Concatenation of the word *biometric* with *token* is ambiguous. The part of the individual presently offered to the biometric scanner, should be referred to as a *biometric characteristic*.

Fake biometric characteristic. Fake biometric characteristic is a biometric characteristic fraudulently manufactured [21].

Latent biometric image. Latent biometric image is an image left behind by an individual as he/she interacts with the environment, i.e. the imprints left on a glass,

handled by an individual. This image can possibly be used to manufacture (create) a fake biometric characteristic.

Genuine biometric data. This term defines the legitimate biometric data as indeed being generated by the authentic user. *Genuine biometric data* is the result of generating a legit electronic representation of the biometric characteristic. Genuine biometric data is found if the biometric is digitized from the authentic user.

Illicit biometric data. This phrase refers to biometric data obtained by means of an illegitimate process and subsequently offered illicitly as biometric data of the authentic user. This term extends the understanding of the particular state of the biometric data. Illicit biometric data is all instances of biometric data that has not been acquired by legal means.

Reference biometric data. During the initial controlled enrolment process, the biometric system creates a special biometric template [32] which will be used by the matching algorithm to test any offered biometric data for a successful match. *Reference biometric data* refers to biometric data stored by the biometric registration system for testing offered biometric data.

In the paper entitled “The biometric landscape—towards a sustainable biometric terminology framework” [30], further clarification, as well as a diagrammatic explanation can be obtained related to the terminology used in this paper.

2.1.2 Fingerprint Biometrics

In this paper an investigation is made into the possibility to create a fake biometric characteristic by using 3D printing technology. The question is asked “Can a fake biometric characteristic be created by means of 3D printing technology?”.

Due to the common usage of finger print biometric technology, this research focused on the creation of fake fingerprint biometric characteristics using either “Fused deposit modelling” (FDM) or “Vat polymerization” (also known resin 3D printing) 3D printing technologies, as is discussed in Sect. 2.2.

Finger print biometrics utilizes the ridge details, or more specifically, the impressions of the papillary or friction ridges on the surface of the finger, which are used to uniquely identify a person [5]. In broad terms, biometric fingerprints are classified into 3 categories by the Henry system [8]. Using the Henry system, each individual fingerprint is classified as either a loop, an arch or a whorl (Fig. 1).

Loop: In a loop fingerprint pattern, the ridges enter from either side, re-curve and pass out or tend to pass out the same side they entered. In the loop pattern there are two focal points: the Core, or the centre of the loop, and the delta. The delta is the area of the pattern where there is a triangulation or a dividing of the ridges.

Arch: In an arch fingerprint pattern the ridges enter from one side, make a rise in the center and exit on the opposite side. The arch pattern has no delta or core.

Whorl: In a whorl pattern the ridges are usually circular, furthermore a whorl pattern will have two or more deltas. For a whorl pattern, all deltas and the areas between them must be recorded.



Fig. 1 Fingerprint classification

In order to distinguish between two fingerprints in the same class, for e.g. two fingerprints from the whorl class, a system which allows for identification of micro characteristics, as found in the ridge pattern of an individual print is used [17]. The National Institute of Standards and Technology (NIST) [24] created a standard for forensic identification and called these micro characteristics minutia (singular) or minutiae (plural).

2.1.3 Minutia

Minutiae are a number of small changes in the friction ridges of a fingerprint. These minutiae are recorded and allow for a fingerprint to be uniquely identified. During the classification of a fingerprint, various minutia points will be recorded according to the following types of minutiae that are commonly found in a fingerprint [24] (Fig. 2):

- **Endings**, the points at which a ridge stops.
- **Bifurcations**, the point at which one ridge divides into two ridges.
- **Dots**, very small ridges.

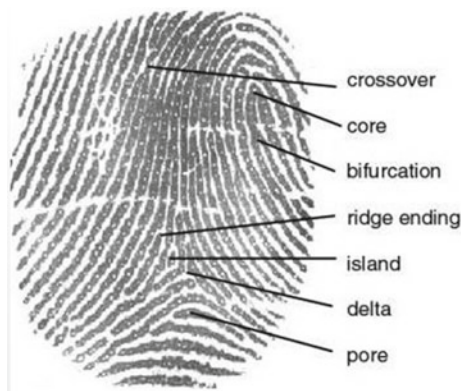


Fig. 2 Example of minutiae on a fingerprint

- **Islands**, ridges slightly longer than dots, occupying a middle space between two temporarily divergent ridges.
- **Ponds or lakes**, empty spaces between two temporarily divergent ridges.
- **Spurs**, a notch protruding from a ridge.
- **Bridges**, small ridges joining two longer adjacent ridges.
- **Crossovers**, two ridges which cross each other.
- The **core** is the inner point, normally in the middle of the print, around which swirls, loops, or arches centre. It is frequently characterized by a ridge ending and several acutely curved ridges.
- **Deltas** are the points, normally at the lower left and right hand of the fingerprint, around which a triangular series of ridges centre.

2.1.4 Finger Print Digitizer

A fingerprint digitizer creates a digital representation of the biometric characteristic that a person is offering for authentication. This digitizer is commonly referred to as a fingerprint scanner or fingerprint sensor. A typical fingerprint scanner [28] is presented in Fig. 3, and is used to capture ridge detail from a finger to create biometric data.

In order to subvert a biometric system by providing a fake biometric characteristic, a fake biometric characteristic must be manufactured. In the past, research by Tsutomu Matsumoto [21], demonstrated one approach to manufacture a fake biometric fingerprint by using gelatin and purpose made moulds to cast the gelatin. This paper investigates the possibility of creating a fake biometric characteristic by utilizing 3D printing technology. For this reason, the next section presents 3D printing technology, and discusses the aspects that must be considered when using 3D printing technology.

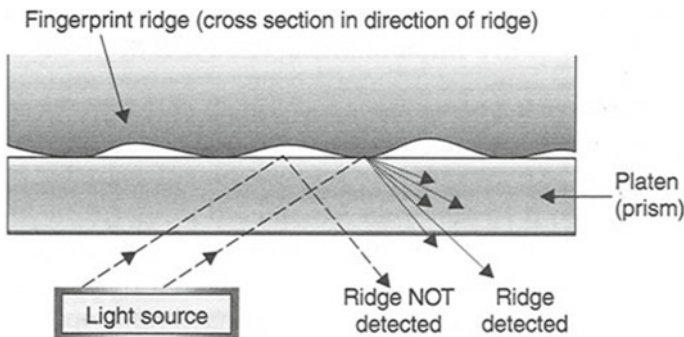


Fig. 3 Cross-section view of optical scanner

2.2 3D Printing

Various 3D printing approaches exist currently. There are seven main types and more than 20 subtypes of 3D printing technology used today [2]. Noteworthy 3D printing technologies are:

- Fused Deposition Modelling
- Vat polymerization—which uses 3 different technologies, known as Stereolithography, Digital Light Processing and Liquid Crystal Display (also known as masked stereolithography). The fundamental difference between these types of 3D printing technology is the light source and how it is used to cure the resin in the vat
- Powder Bed Fusion
- Material Jetting (which works similar to inkjet and bubble jet technologies commonly used in document printers.)
- Binder Jetting
- Directed Energy Deposition
- Sheet Lamination.

An in-depth discussion of all the mentioned 3D printing technologies is beyond the scope of this paper.

As Fused Deposit Modelling (FDM) and Vat polymerization is used in the research conducted in this research project, these two technologies are discussed in Sects. 2.2.1 and 2.2.2 respectively.

A 3D-printed artefact utilized to create a fake biometric characteristic requires a fine enough printing resolution to provide sufficient detail that correctly represent the fingerprint ridge detail as discussed in the minutiae section. With this in mind, the following technologies were tested to see if a fake biometric characteristic can be 3D printed that will subsequently be capable of subverting the biometric authentication system.

2.2.1 Fused Deposition Modelling

In 2000 a colleague, Dr. Emil Marais conducted research in a field called “Tele-manufacturing” [20]. This technology used FDM to create 3D models based on a detailed mathematical instruction set. Due to the rarity of FDM printers in 2000, the instructions to create the 3D model had to be sent to Germany, where the model was printed, and sent back to Dr. Marais via courier. In 2000 the technology only allowed for basic shapes and figures to be created and the models created had very poor resolution, and not suitable for attempting to print a fake biometric fingerprint.

Today, FDM printers are quite cheap, and definitely within reach of any house hold. Futhermore the modern 3D printers also print with a significantly higher resolution. FDM printers are available in various sizes, and can utilize various materials to 3D print an object. Noteworthy 3D printers available at an affordable price are the

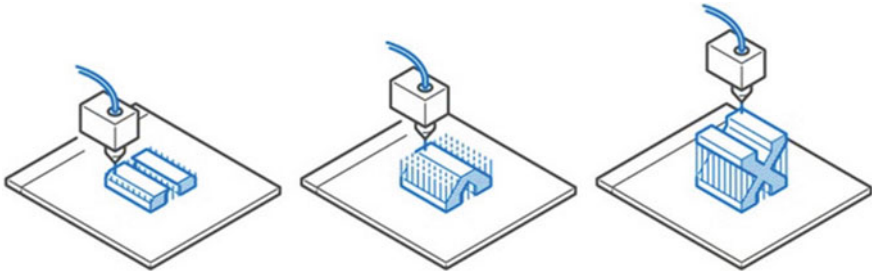


Fig. 4 Material extrusion 3D printing [2]

Creality Ender 3 V2 [9] (used in this research) and the **Prusa Prusa MK4 3D Printer [26]**

Figure 4 illustrates the basic working of a FDM 3D Printer [2]. In relation to FDM 3D printing the following must be noted:

- **Printing method:** The FDM printer feeds filament from a spool to a HotEnd, using small gears. The HotEnd melts the filament and then the melted filament is deposited onto the build surface by the HotEnd nozzle. X-axes, Y-axes and Z-axes movement of the print nozzle is controlled via stepper motors, allowing for very accurate deposit of the filament onto the build surface as illustrated in Fig. 4.
- **Printing material:** 6 main filament types can be used in the commercial 3D printing machines [19]. PLA (Polylactic Acid), PETG (Polyethylene Terephthalate Glycol), TPE/TPU (Thermoplastic Elastomer/Polyurethane), ABS (Acrylonitrile Butadiene Styrene), ASA (Acrylonitrile Styrene Acrylate), PA (Polyamide or Nylon). Each filament type has specific strengths and noteworthy weaknesses. The exact filament used in this research is discussed in Sect. 3.1.2.
- **Extruder technology** FDM technology uses one of two extruder technologies [13]. If a *Bowden extruder* is used, the extruders are designed to be mounted away from the moving mass of the printing head. A piece of Bowden tube, varying in length depending on the size of printer and the mounting position of the extruder, is then used to connect the extruder and the HotEnd. Alternatively *direct Drive* extruder setups have the extruder mounted immediately before the HotEnd on the printhead assembly. In general, this results in a heavier moving mass than Bowden, but offer better control over the filament due to the shorter distance between the filament feeding gears and HotEnd printing nozzle.
- **Printing resolution.** The printing resolution of an FDM printer is firstly determined by the size of the printing nozzle. The second aspect that influences the printing resolution, is the micro movements of the X, Y, and Z stepper motors, as well as the stepper motor feeding the extruder HotEnd. The typical HotEnd nozzle size is 0.4 mm, with the smallest nozzle size currently being 0.2 mm [2].

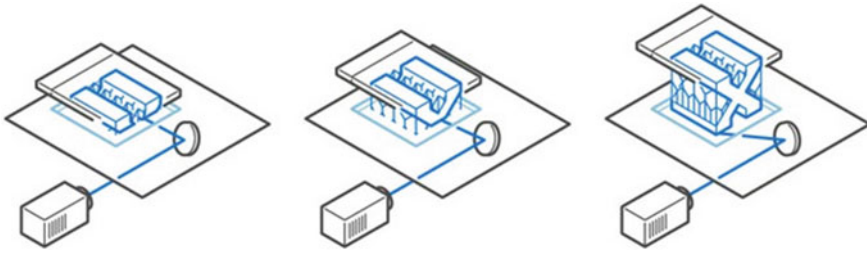


Fig. 5 Vat polymerization using a laser [2]

2.2.2 Vat Polymerization

Vat polymerization, or resin 3D printing, yields exceptionally detailed 3D models due to the fact that the technology uses a light source to selectively cure (or harden) photopolymer resin in a vat [2].

Figure 5 illustrates the basic working of a Vat polymerization 3D Printer [2].

- **Printing method:** As illustrated in Fig. 5, a light source (in this case a laser), is precisely directed to a specific point or area of the liquid using mirrors. This light source is known as galvanometers or galvos, with one positioned on the X-axis and another on the Y-axis. These galvos rapidly aim a laser beam across a vat of resin, selectively curing and solidifying a cross-section of the object inside this building area. This results in building the 3D model layer by layer. Once the first layer is cured, the build platform is moved up or down (depending on the printer design) by a small amount (typically between 0.01 and 0.05 mm), and the next layer is cured using the light source, joining the new layer to previous hardened one. This process is repeated layer by layer until the 3D object is complete. Once the printing process is complete, the remaining, unused resin, is either drained from the vat, or the 3D printed model is lifted out of the resin.
- **Printing material:** Photopolymer resins which can be cast-able, transparent, industrial, bio-compatible.
- **Polymerization technology** The 3 most commonly used polymerization approaches are stereolithography (SLA), digital light processing (DLP), and liquid crystal display (LCD)-also known as masked stereolithography [2]. SLA was used in this research.
- **Printing resolution.** The accuracy of printing a 3D model ranges for $\pm 0.15\text{mm}$ to $5\mu\text{m}$, depending on the specific technology used [25].

3 Experimental Setup

To test if it is possible to 3D print an artefact that can be used to subvert a biometric scanner, a number of aspects were considered to define the test domain. These test premises are listed here:

- Lifting of a latent biometric characteristic from the environment that the user interacts with (such as lifting a latent fingerprint from a glass handled by the user) falls outside the scope of this research. To that extent a fingerprint, already digitized and in digital format, was used to 3D print all artefacts. An example of how a latent fingerprint can be lifted from the user's environment is presented in the research by Tsutomu Matsumoto [21].
- The purpose of this research was to determine if the biometric scanner will accept a fake biometric characteristic to force a false acceptance of the presented fake biometric characteristic. To ensure that the ridge detail are easy to read, the image of the fingerprint was enhanced and processed using CAD software before the printing process started. This is similar to the processing done by biometric system software, which include image processing to segment, Thin and binarize and prune the digital image of the fingerprint. The software packages used is mentioned in Sect. 3.1.
- For a 3D model to be printed, the 3D model must be sliced using software that define each layer to be printed. In order to get a usable 3D printed artefact, various settings in the slicing software were manipulated. Unfortunately quite a few 3D artefacts were not usable and disregarded. The settings that were found to be successful are discussed in Sect. 3.1.

3.1 Experimental Setup

This section outlines all aspects as incorporated during the experimental phase of this research. Attention is given to the CAD and slicing software used to create the 3D biometric models, as well as the 3D printers and materials used to print the artefacts. The 3 biometric scanners tested, and the testing approaches followed to test them, are discussed.

3.1.1 Software

Two CAD software packages were used to create the 3D models to be printed namely *Autodesk 3ds Max* [4] and *Adobe Substance 3D Designer* [1]. A two dimensional JPG image of a fingerprint was used and enhanced using Autodesk 3ds Max. The JPG is then converted to Scalable Vector Graphics (SVG) as the SVG graphic format ensures that the dimensional ratios remains constant when scaling the image of the

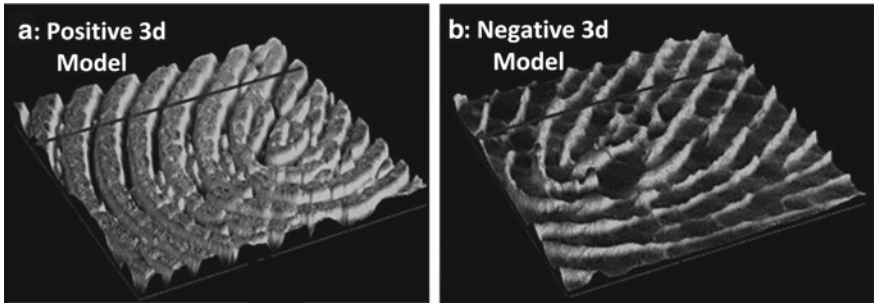


Fig. 6 Example of 3D image of fingerprint

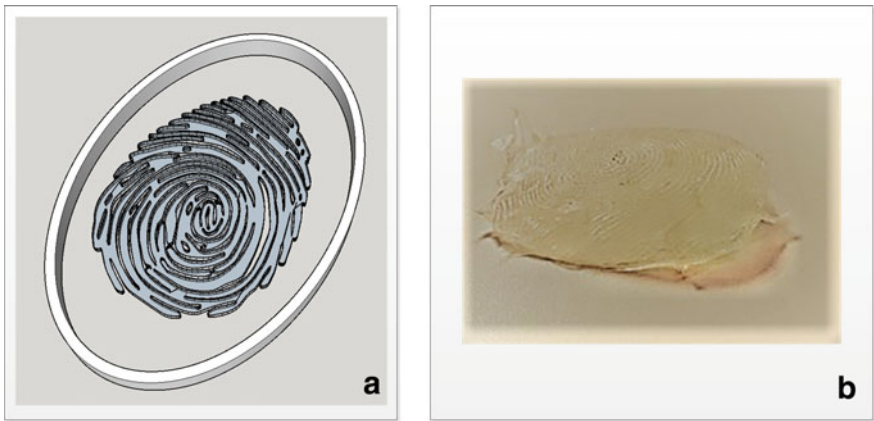


Fig. 7 CAD 3D model for 3D printing and fake latex fingerprint

fingerprint. The SVG is then converted to a 3D model of the fingerprint image, using Adobe Substance 3D designer. At this point the software 3D model is an enhanced digital representation of the biometric characteristic. Figure 6a shows the enhanced 3d image of such a fingerprint.

Considering that a Latex mould will be created of this 3D image, a “negative” (or inverse) is required, otherwise the eventual latex fingerprint will be a mirror image of the actual fingerprint. This negative version of the 3D image is illustrated in Fig. 6b.

The negative 3D model of the finger print is added to a rudimentary container using the 3D modeling software, to form a mould which can be used to cast latex into. The completed software designed mould is shown in Fig. 7a. This model destined to be 3D printed, is saved as a STereoLithography (STL) file.

The STL file is loaded into software that will slice the 3D model into layers to be printed by the relevant 3D printing technology. Cura version 4.12.1 [31], as developed by Ultimaker, was used to slice the STL file and generate a gcode file. The gcode file contains specific G- and M- commands, instructing the 3D printer to move the various stepper motors or galvos, toward construction of the 3D model. As

mentioned earlier, the settings in the slicer software were altered to yield a model to allow for an optimal mould.

3.1.2 3D Printing and Fingerprint Moulding

In this research, only two 3D printing approaches were used to create the moulds.

Firstly a **FDM 3D printer**, the Ender 3 V2 printer was used with a 0.2 mm nozzle with the layer height set in Cura (the slicer software) to 0.1mm. The Ender 3 printer uses a Bowden extruder, and to attempt to get a clean print between HotEnd extruder movements, the “enable retraction” setting was selected. As the mould created does not require durability, PLA filament was used during the printing process. Lastly a slow print speed of 5.0 mm/s was selected to ensure optimum possible accuracy during the mould printing process. It took roughly 68 min to print the mould.

Considering that generally human papillary have a height of between $20\mu\text{m}$ and $60\mu\text{m}$, it was decided to use a **resin polymer 3D printer** as the second option to print the mould. The Creality Halot-one Plus Resin 3D Printer was used to print the same 3D mould. After the printing process completed, the mould was cleaned and rinsed using Isopropyl alcohol and allowed to fully cured using a UV lamp.

To cast the fake biometric finger print, the 3D printed hollow moulds were filled with just a few drops of liquid latex, resulting in a paper-thin fake biometric characteristic, which can then be used as an overlay on the hacker’s finger. Figure 7b shows an example of the fake biometric finger print.

3.1.3 Fingerprint Scanner Testing

The biometric scanner of three devices were tested namely a Notebook, cell phone and a door access scanner.

The notebook was a Lenovo notebook which uses an inline scanner (this means that the biometric characteristic must be dragged over a small scanner). The cellphone used was the Samsung Note 9 cellphone, and a physical access control scanner that was tested, is mounted as an access point to a secure area (the manufacturer information must be excluded at this point).

Test 1—Vanilla test In all 3 cases the biometric scanner of each device was tested using the fake latex fingerprint from the FDM mould and the fake latex fingerprint from the resin polymer mould, without any alterations to the latex and recorded as a “vanilla test”.

Test 2—Gel test Due to observations during the first test a second test was conducted on all three devices using the same fake latex fingerprint, but during the second test, Dynarex electrically conductive gel [11] was applied to the fake latex fingerprint. This test was recorded as “Gel test”.

4 Findings and Results

From outset it was clear that the mould created using the FDM process did not yield a mould that had enough ridge detail to create a fake biometric characteristic that will fool a biometric scanner. Numerous print attempts were made, altering various settings in the slicer software, but it was clear that the FDM printer used in this research introduces too many anomalies in the mould, resulting in a fake biometric characteristic that causes the authentication attempt to fail. Even though a fake biometric cast was created from the best of the 3D printed moulds, neither of devices accepted the fake biometric characteristic for successful authentication.

As the resin polymer produces a mould with exceptional detail, it was clear that the resulting mould was a very good (inverse) representation of the biometric characteristic, and showed a lot of promise to yield a usable fake biometric characteristic after the casting process. However, not all tested biometric scanners detected the fake biometric characteristic due to the different technologies employed in the biometric scanner to digitize the biometric characteristic. The laptop and physical access control digitizer detected the fake biometric characteristic and authenticated the fake characteristic. The Note 9 did not detect the latex material at all, due to the fact that the biometric scanner used by the Samsung is a *capacitive CMOS sensor* [14]. This type of biometric scanner are reliant on reading the electrical potential differences between ridges of the fingerprint. As latex is not electrically conductive, the sensor does not detect the fingerprint.

In the second test, called the “Gel test”, Dynarex electrically conductive gel was applied to the human finger and also to the inside and outside of the fake latex fingerprint. At this point the cellphone did detect the fake latex fingerprint (overlayed on the human finger), but it took a number of attempts to get the phone to authenticate the fake biometric characteristic. The results are shown in Tables 1 and 2.

Table 1 Test 1—vanilla test

Mould type	Detected	Device	Authenticated	Attempts	Success rate (%)
FDM	Yes	Physical access	No	0 out 10	0
FDM	No	Samsung note 9	No	0 out 10	0
FDM	Yes	Lenovo laptop	No	0 out 10	0
Resin	Yes	Physical access	Yes	7 out 10	70
Resin	No	Samsung note 9	No	0 out 10	0
Resin	Yes	Lenovo laptop	Yes	6 out 10	60

Table 2 Test 2—gel test

Mould type	Detected	Device	Authenticated	Attempts	Success rate (%)
FDM	Yes	Physical access	No	0 out 10	0
FDM	Yes	Samsung note 9	No	0 out 10	0
FDM	Yes	Lenovo laptop	No	0 out 10	0
Resin	Yes	Physical access	Yes	10 out 10	0
Resin	Yes	Samsung note 9	No	3 out 10	30
Resin	Yes	Lenovo laptop	Yes	9 out 10	90

From the data presented in Tables 1 and 2, it is clear that the addition of the electrically conductive gel, facilitated the various scanners to successfully detect and authenticate the fake biometric characteristic.

5 Conclusion and Future Research

From the experiments conducted in this research, it is clear that commonly available commercial 3D printing technology makes it possible to generate an artefact that can be used to successfully subvert systems using biometric technology for security and access control. It was found at this point that a Vat polymer 3D printer can create a mould that provides ridge detail which is on par with the ridge detail found on the actual biometric characteristic of the human finger.

However, this research was done as a proof of concept, but it must be noted that 3D technology is not limited to only printing resins, or different filament types. **An alarming prospect to consider** is that 3D printing technology has progressed to the point where human skin and human organs can be 3D printed, this technology is called 3D bioprinting [6, 23]. 3D bioprinting is an additive manufacturing process where organic or biological materials, such as living cells and nutrients, are combined to create natural tissue-like three-dimensional structures. In other words, 3D bioprinting is a type of 3D printing that can potentially produce anything from bone tissue and blood vessels to living tissues.

At this point these 3D printers are very expensive, requires very specialized equipment and are not accessible to the general public. However this opens the door for future research to test if 3D bioprinting can be used to 3D print a fake biometric characteristic from real human skin, or even 3D bioprint an actual organic human iris...

References

1. Adobe: Adobe substance 3d designer software (2023). <https://www.adobe.com/za/products/substance3d-designer.html>. Accessed 25 July 2023
2. All3DP: the 7 main types of 3d printing technology (2021). <https://all3dp.com/1/types-of-3d-printers-3d-printing-technology>. Accessed 20 July 2023
3. Andrade C, von Solms SH (2008) Investigating and comparing multimodal biometric techniques. In: Policies and research in identity management: first IFIP WG11. 6 working conference on policies and research in identity management (IDMAN'07), RSM Erasmus University, Rotterdam, The Netherlands, October 11-12, 2007. Springer, pp 79–90
4. Autodesk: Autodesk 3ds max: create massive worlds and high-quality designs (2023). <https://www.autodesk.com/products/3ds-max/overview>. Accessed 25 July 2023
5. Bansal R, Sehgal P, Bedi P (2011) Minutiae extraction from fingerprint images-a review. arXiv preprint [arXiv:1201.1422](https://arxiv.org/abs/1201.1422)
6. Bliley JM, Shiwerski DJ, Feinberg AW (2022) 3d-bioprinted human tissue and the path toward clinical translation. *Sci Transl Med* 14(666): eabo7047
7. Chapple M, Seidl D (2021) Social engineering, physical, and password attacks. *CompTIA Security+ Study Guide: Exam SY0-601*, pp 65–81
8. Cole SA (2004) History of fingerprint pattern recognition. In: *Automatic fingerprint recognition systems*. Springer, pp 1–25
9. Creality: creality ender 3 v2 3d printer (2021). <https://www.creality.com/products/ender-3-v2-3d-printer-csco>. Accessed 20 July 2023
10. Di Campi AM, Focardi R, Luccio FL (2022) The revenge of password crackers: automated training of password cracking tools. In: *European symposium on research in computer security*. Springer, pp 317–336
11. Dynarex: Dynarex electrically conductive gel ecg (2023). <https://dynarex.com/base-electrically-conductive-gel-ecg.html>. Accessed 25 July 2023
12. El Fkharany HM (2022) Applying biometric technology for enhancing airports efficiency during covid-19 pandemic: A case study of egyptian destination. *The Int J Tourism Hospitality Stud* 3(2):104–117
13. Hardy W (2022) Bowden vs direct drive extruder. <https://e3d-online.com/blogs/news/bowden-vs-direct-drive>. Accessed 20 July 2023
14. Hassan H, Kim HW (2018) Cmos capacitive fingerprint sensor based on differential sensing circuit with noise cancellation. *Sensors* 18(7):2200
15. Johnsoncontrols: biometric physical access control (2023). <https://www.johnsoncontrols.com/security/access-control/biometrics-access-control>. Accessed 19 July 2023
16. Kanta A, Coisel I, Scanlon M (2022) A novel dictionary generation methodology for contextual-based password cracking. *IEEE Access* 10:59178–59188
17. Krish RP, Fierrez J, Ramos D, Alonso-Fernandez F, Bigun J (2019) Improving automated latent fingerprint identification using extended minutia types. *Inf Fusion* 50:9–19
18. Lenovo: setting up fingerprint reader on your lenovo laptop (2023). <https://pcsupport.lenovo.com/us/en/products/laptops-and-netbooks/thinkbook-series/thinkbook-14-impl/videos/nvid500012-setting-up-fingerprint-reader-on-your-lenovo-pc>. Accessed 19 July 2023
19. MakeUseOf: the 6 different fdm 3d printer filament types (2022). <https://www.makeuseof.com/what-are-the-3d-printer-filament-types>. Accessed 20 July 2023
20. Marais E, Ehlers E, Dutta D (2000) Adaptive slicing by telemanufacturing. In: *Proceedings of the third international symposium on tools and methods of competitive engineering-TMCE 2000*, pp 18–21
21. Matsumoto T, Matsumoto H, Yamada K, Hoshino S (2002) Impact of artificial "gummy" fingers on fingerprint systems. In: *Optical security and counterfeit deterrence techniques IV*. vol 4677. SPIE, pp 275–289
22. McGrath J, Bowyer KW, Czajka A (2018) Open source presentation attack detection baseline for iris recognition. arXiv preprint [arXiv:1809.10172](https://arxiv.org/abs/1809.10172)

23. Mudigonda J (2024) 3d bioprinting of tissues and organs: a comprehensive review of the techniques, recent advances, and their applications in organ engineering and regenerative medicine. *Adv 3D Bioprinting* 1–54
24. NIST: fingerprint classification (2010). <https://www.nist.gov/programs-projects/fingerprint>. Accessed 20 July 2023
25. Pagac M, Hajnys J, Ma QP, Jancar L, Jansa J, Stefek P, Mesicek J (2021) A review of vat photopolymerization technology: materials, applications, challenges, and future trends of 3d printing. *Polymers* 13(4):598
26. Prusa: Prusa mk4 3d printer (2022). <https://www.prusa3d.com/product/original-prusa-mk4-2>. Accessed 20 July 2023
27. Ratha NK, Connell JH, Bolle RM (2001) Enhancing security and privacy in biometrics-based authentication systems. *IBM Syst J* 40(3):614–634
28. Samir N, Michael T, Raj N (2002) Biometrics: Identity verification in a networked world
29. Samsung: how to set up and use fingerprint recognition sensor on galaxy s23 (2023). <https://www.samsung.com/uk/support/mobile-devices/how-to-set-up-and-use-fingerprint-recognition-sensor-on-galaxy-s23-5g>. Accessed 19 July 2023
30. Tait BL (2014) The biometric landscape-towards a sustainable biometric terminology framework. *Int J Electron Secur Digital Forensics* 9 6(2): 147–156
31. Ultimaker: Ultimaker cura free, easy-to-use 3d slicing software (2023). <https://ultimaker.com/software/ultimaker-cura/>. Accessed 25 July 2023
32. Zhang X, Cheng D, Jia P, Dai Y, Xu X (2020) An efficient android-based multimodal biometric authentication system with face and voice. *IEEE Access* 8:102757–102772
33. Zimmermann V, Gerber N (2020) The password is dead, long live the password-a laboratory study on user perceptions of authentication schemes. *Int J Human-Comput Stud* 133:26–44

Image Band-Distributive PCA Based Face Recognition Technique



Henry Ndu , Akbar Sheikh-Akbari , Iosif Mporas , and Jiamei Deng

Abstract This paper presents an Image Band-Distributive PCA (IBD-PCA) based technique for face recognition. The proposed method consists of four steps. In the first step, the reference image is pre-processed by converting its pixel values and performing histogram equalization to increase its contrast. In the second step, the equal-size boundary calculation method is used to calculate the boundary splitting values to divide the input image into multiple images with respect to band intensities of pixels. In the third step, Principal Component Analysis (PCA) is used to extract features from the images which will then be used as the input for the fourth step. In the last step, matching is performed by calculating the Euclidean distance between principal components. The proposed technique has been tested on the ORL Face Database and Yale Face Database. The experimental results demonstrate that the proposed technique outperforms other techniques on the same database.

Keywords Face recognition · Principal component analysis · Image band-distributive PCA

H. Ndu · A. Sheikh-Akbari (✉) · J. Deng
School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds, UK
e-mail: a.sheikh-akbari@leedsbeckett.ac.uk

H. Ndu
e-mail: h.ndu6387@student.leedsbeckett.ac.uk

J. Deng
e-mail: j.deng@leedsbeckett.ac.uk

I. Mporas
School of Engineering, Engineering and Computer Science, University of Hertfordshire, Hatfield, UK
e-mail: i.mporas@herts.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_11

1 Introduction

Face recognition is a highly researched field within the realm of biometrics, focusing on the identification of individuals based on their facial images [1]. It offers numerous advantages, particularly in areas such as security and surveillance, banking, database retrieval, law enforcement, criminal justice systems, and virtual reality [1]. Despite the considerable amount of research dedicated to face recognition, there is still a significant need for improved techniques that can accurately and effectively recognize individuals.

Principal Component Analysis (PCA) is one of the techniques that has been extensively utilized for face recognition [2]. PCA is commonly employed to reduce the dimensionality of input image data while preserving its essential characteristics. By revealing vital aspects of the data, PCA enables the identification of patterns and the exploration of similarities and differences between features in the input data, facilitating feasible investigations.

However, it is important to note that face recognition is a complex and evolving field, and researchers continue to explore and develop new methods to overcome the challenges associated with accurate identification. While PCA has proven to be effective in many cases, it is not without limitations. One limitation is its sensitivity to variations in lighting conditions, facial expressions, and pose. In real-world scenarios, these factors can significantly affect the performance of PCA-based face recognition systems.

In recent years, researchers have proposed various face recognition methods [3–5]. These methods can generally be categorized into two groups: holistic template matching and geometrical local feature-based techniques [6]. Among these approaches, PCA-based methods employ the holistic template matching technique and have shown promising results for face recognition [7].

1.1 PCA for Face Recognition

In face recognition, PCA is utilized as a technique for feature extraction and dimensionality reduction. In PCA-based face recognition, the face images are transformed into a lower-dimensional subspace, known as the eigenspace, using PCA. This subspace captures the most significant variations in the face images, and they serve as a basis for representing and recognizing faces. The eigenvectors, called eigenfaces, are the principal components extracted during the PCA process. These eigenfaces form a set of basis vectors that span the face image space, allowing for efficient representation and classification of face images.

1.2 Eigen Faces

One of the pioneering works in PCA-based face recognition was introduced by Turk and Pentland [8]. They developed an approach for using PCA for face recognition known as “eigenfaces” that involves applying PCA to a database of face images to derive eigenfaces, which are then used to represent and classify new face images. The eigenfaces technique represents each face in the database as a vector of weights, obtained by projecting the image onto eigenface components. When a test image is provided for identification, it is also represented by its vector of weights, and identification is performed by finding the image in the database with weights closest (in terms of Euclidean distance) to the test image’s weights [6].

On the other hand, geometrical local feature-based methods focus on extracting specific facial features, such as the nose, eyes, and mouth, and utilize their locations for recognition [6]. Jafri and Arabnia [9] employed a local feature-based approach for face recognition and reported encouraging results.

In this paper, we address the existing gap in the literature by introducing a novel approach for face recognition that combines image band-distributive generation with PCA. While researchers have made significant advancements in enhancing PCA-based face recognition methods, the incorporation of image band-distributive generation techniques has not been explored. The proposed method, called Image Band-Distributive PCA (IBD-PCA), fills this research gap and offers a robust solution for improving the performance of PCA-based face recognition systems. By generating multiple images from the input grayscale image based on pixel intensities, enables the extraction of facial features from different intensity ranges, leading to enhanced matching accuracy.

The rest of the paper is organized as follows: Sect. 2 describes the proposed technique. Section 3 provides details about the experiments performed and the results obtained in each case, and finally Sect. 4 draws the conclusion.

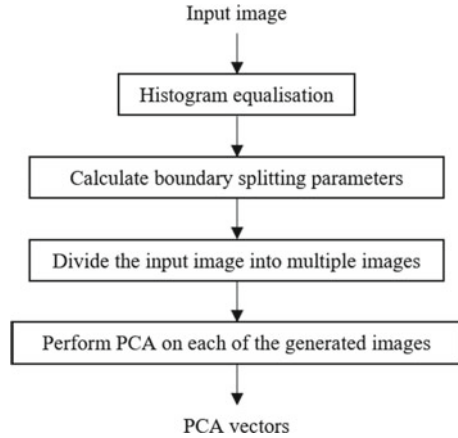
2 Proposed Technique

In this section, the proposed Image Band-Distributive PCA (IBD-PCA) for face recognition technique is presented.

The IBD-PCA technique introduces a unique and effective strategy to enhance the performance of traditional methods by leveraging the power of feature extraction. Unlike conventional approaches, the proposed method adopts a distinctive perspective by dividing the input grey image into multiple images, each capturing specific facial characteristics. This distribution of image data across multiple images enables the incorporation of both global and local facial details, resulting in a more comprehensive representation of the face features.

The core idea behind IBD-PCA lies in applying Principal Component Analysis (PCA) independently on each of the divided images. By extracting the principal

Fig. 1 Flowchart of the proposed technique



components from these images, discriminative features encapsulating the essential facial information are obtained. This image band-distributive approach enables the exploitation of diverse facial appearance variations across different regions of the face, leading to improved recognition accuracy. A flowchart of the proposed technique is shown in Fig. 1.

The proposed technique consists of four stages namely: Pre-processing, Multiple-Image Generation, conventional PCA, and Matching, which are detailed in the following sections.

2.1 Pre-Processing

Let I represent a dataset of face images, with each image in I of size $x \times y$. The input image $i \in I$ is assumed to be an 8-bit grayscale image. The input image pixels are first normalised to be mapped to the range of $[0, 1]$. The contrast of the resulting image is then increased using histogram equalization. This is done by first calculating the Probability Mass Function (PMF) PMF_X of the input image as follows:

$$PMF_X(x_j) = PMF(X = x_j) \text{ for } j = 0, 1, \dots, 255 \tag{1}$$

where $X = x_0, x_1, \dots, x_{255}$ presents the pixels values and $PMF_X(x_j)$ represents the probability of coefficients in bin j . The resulting PMF is then used to derive the Cumulative Distribution Function (CDF) CDF_x by considering the following expression:

$$CDF_x(j) = PMF(X \leq x_j) \text{ for } j = 0, 1, \dots, 255 \tag{2}$$

where $CDF_x(j)$ is the cumulative probability of $X \leq x_j$. The pixels within the input image are mapped using the resulting CDF. The resulting histogram equalized image is then used for multiple-image generation.

2.2 Multi-image Generation

To generate multiple-images from the resulting histogram equalized input image, the proposed IBD-PCA method splits the input image into predefined number of images according to intensity of its pixels using an *equal size* boundary calculation method. If i is the input histogram equalized image and N is the predefined number of images desired from the input image i , the proposed algorithm calculates $N-1$ boundaries to split the pixels of the input image into N target images. The pixel value boundaries $B = \{b_1, b_2, \dots, b_{N-1}\}$ are then calculated using Eq. 3:

$$b_n = n/N \text{ for } n = 1, 2, \dots, (N - 1) \quad (3)$$

If $B = \{b_1, b_2, \dots, b_{N-1}\}$ represents the resulting boundary values, the input image i is then divided into N target images by:

1. Generating N images that have the same size as the input i image with pixel values of zero. Let's assume the generated images be: $R = [r_1, r_2, \dots, r_N]$.
2. Assigning all valued pixels p of i in the range $[0, b_1], [b_1, b_2], \dots, [b_{(N-1)}, 1]$ to image r_1, r_2, \dots, r_N , respectively.

Each resulting image in the image set of R has its own band intensity. The coefficients in each resulting image are then mapped to $[0, 1]$. This enables the algorithm to extract face features from different pixel intensity of the image, increasing the matching accuracy of the proposed IBD-PCA based face recognition technique.

2.3 Principal Component Analysis

Assume R is the set of the resulting multiple-images, which their coefficients have been mapped to $[0, 1]$, where $R = \{r_1, r_2, \dots, r_N\}$. Each image $r_j \in R$ is mean adjusted, generating r'_j as follows:

$$r'_j = r_j - \bar{r}_j \quad (4)$$

where \bar{r}_j represents the mean value of the pixels in image r_j .

Each resulting mean adjusted image r'_j is then converted to a column wise vector. Hence, R can be represented as a two-dimensional matrix S .

Principal Component Analysis (PCA) is then performed using Singular Value Decomposition (SVD) on the resulting matrix S , generating the following decomposition:

$$S = U \Sigma V^T \quad (5)$$

where U is a unity matrix, and the columns of V are the orthonormal eigenvectors of the covariance matrix of S and Σ is a diagonal matrix of their respective eigenvalues. The eigenvectors form a basis for an eigenspace for each set of images R . The resulting principal components in V are assumed to be the extracted features of the input image, which are used for finding the best match.

2.4 Matching

The matching process of faces is based on Euclidean distance measurement. Euclidean distance calculations can be used to determine the distance between any two points in a two-dimensional space, and to derive the absolute distance between points in N -dimensional space. In face recognition, smaller Euclidean distance values indicates more similarity between faces.

The face image to be detected is processed to obtain a N -dimensional face feature vector, which enables a condition for Euclidean distance calculations.

Let $\mathbf{A} = (x_1, x_2, \dots, x_n)$ be the principal components of an input face image \mathbf{A} to be detected, and $\mathbf{B} = (y_1, y_2, \dots, y_n)$ is the principal components of an image \mathbf{B} in a dataset of images, the Euclidean distance can be calculated as follows:

$$AB = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} \quad (6)$$

The resulting value from calculating AB represents the Euclidean distance between principal components of both face images.

After calculations of Euclidean distances between principal components have been done, their average is determined in the form of an average distance metric, as in the expression below, and used for finding the best match:

$$Av\mathbf{AB} = \sum \mathbf{AB} / (N - 1) \quad (7)$$

3 Experiments and Results

This section presents the experiments performed to assess the validity and performance of the developed IBD-PCA technique in the domain of face recognition. The primary objective of these experiments is to establish the effectiveness of the proposed technique by benchmarking its performance against existing methods and thoroughly analyzing the obtained results.

3.1 Database

The proposed IBD-PCA method was used for face recognition and tested using two benchmark face image databases called: the ORL Face Database [10] and Yale Face Database [11]. The ORL database was used to assess the performance of IBD-PCA under varying conditions of pose and sample size. The Yale database was used to assess the performance of the proposed method with varying facial expressions and illumination.

3.1.1 ORL Face Database

The ORL Face Database [10] contains ten face images of each of 40 distinct subjects, some of which were captured under different times, varying light, facial expressions (open/closed eyes, smiling/not smiling), and facial details (glasses/no glasses). The images are 8-bit grayscale of size 92×112 pixels. Figure 2 shows ten sample images of three individuals from the ORL Face Database.

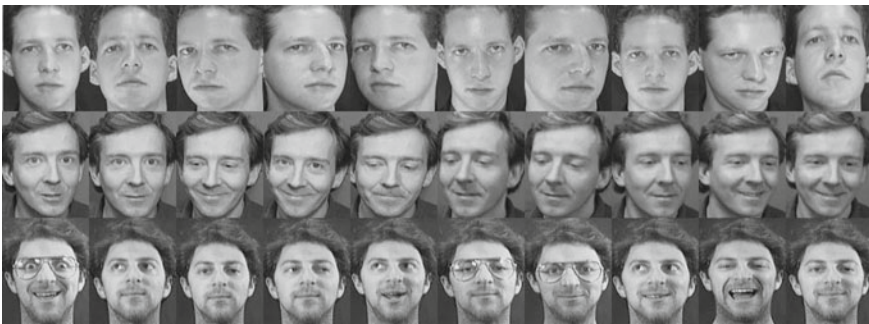


Fig. 2 Sample images from ORL Face Database [10]



Fig. 3 Sample images from Yale Face Database [11]

3.1.2 Yale Face Database

The Yale Face Database [11] contains 11 different face images of each of 15 distinct subjects. The images were also captured in different conditions like the ORL Face Database, and the images were in 8-bit grayscale and each of size 320×243 pixels. Figure 3 shows eleven sample images from of three individuals from the Yale Face Database.

To assess the performance of the proposed method, experiments were conducted on both databases for two main purposes. The first was to generate a baseline for comparison and the second was to determine the optimum matching accuracy.

3.2 Methodology

To establish a meaningful baseline for comparing the performance of the proposed IBD-PCA technique with other existing methods, a comprehensive evaluation using three algorithms: IBD-PCA, conventional PCA, and eigenfaces was conducted. These algorithms were applied to both the ORL and Yale datasets, allowing for the assessment of their respective performances and gaining insights into their unique strengths.

In the experimental setup, Rank-1 and Rank-5 evaluation criteria were employed to gauge the effectiveness of each algorithm. For the ORL dataset, ten captured face images per subject were utilized, with nine images serving as the database and one randomly selected image as the query. Similarly, the Yale dataset consisted of eleven face images per subject, with ten database images and one query image randomly chosen for each subject.

To determine the accuracy of the algorithms, the query image was compared against the database images. If the query image matched correctly with an image from the same subject in the database, it was labelled as a Rank-1 match. Similarly, if the subject was found among the top five closest images to the query image, it was labelled as a Rank-5 match. The percentage of Rank-1 and Rank-5 matches was then calculated based on the total number of query images, yielding the Rank-1 and Rank-5 accuracies.

To ensure the reliability of the findings, this process was repeated for multiple permutations, conducting three trials in total. The average accuracies of Rank-1 and Rank-5 across. These trials are then calculated, providing a robust assessment of each algorithm's performance.

By following this rigorous experimental procedure, a solid foundation for comparing the proposed IBD-PCA technique with conventional PCA and eigenfaces was established. The Rank-1 and Rank-5 accuracies obtained from these evaluations offer valuable insights into the relative strengths and weaknesses of each method, enabling a comprehensive understanding of their respective capabilities in face recognition tasks.

3.3 Experimental Results and Observations

The experimental results, as presented in Table 1 and 2, provide a comprehensive evaluation of the performance of the proposed IBD-PCA technique when assessed using the ORL Face Database and Yale Face Database, conventional PCA, and eigenfaces algorithms using the Rank-1 and Rank-5 evaluation criteria.

In this study, the performance of three face recognition methods: PCA, Eigenfaces, and the proposed IBD-PCA method, using the ORL Face Database and Yale Face Database was evaluated. The Rank 1 and Rank 5 accuracy results were used as performance metrics to assess the effectiveness of these methods.

For the ORL Face Database, the PCA method achieved a Rank 1 accuracy of 49% and a Rank 5 accuracy of 70.75%. This indicates that while PCA captures some discriminative facial features, it struggles to provide precise identification, resulting in relatively low accuracy. In comparison, Eigenfaces achieved a significantly higher Rank 1 accuracy of 91.94% but a slightly lower Rank 5 accuracy of 81.82%. This

Table 1 Rank-1 matching accuracies (%) for PCA, eigenfaces, and proposed IBD-PCA method

Method/dataset	ORL Face Database (%)	Yale Face Database (%)
PCA	49	75.15
Eigenfaces	91.94	75
IBD-PCA	99.25	75.33

Table 2 Rank-5 matching accuracies (%) for PCA, eigenfaces, and proposed IBD-PCA method

Method/dataset	ORL Face Database (%)	Yale Face Database (%)
PCA	70.75	87.27
Eigenfaces	97.78	81.82
IBD-PCA	100	85.33

suggests that Eigenfaces excel in accurately recognizing the most similar face images but exhibit challenges in distinguishing among more dissimilar faces.

Remarkably, the proposed IBD-PCA method outperformed both PCA and Eigenfaces, achieving a remarkable Rank 1 accuracy of 99.25% and a Rank 5 accuracy of 100% on the ORL Face Database. These results demonstrate the effectiveness of the IBD-PCA approach in capturing fine-grained facial variations and enabling highly accurate identification. By utilizing image band-distributive generation based on pixel intensity and incorporating PCA, the proposed method effectively leverages the complementary information contained in the images, resulting in superior performance.

Moving to the Yale Face Database, the comparative analysis yields interesting insights. The PCA method exhibited an improved Rank 1 accuracy of 75.15% and a Rank 5 accuracy of 87.27%. This indicates that the performance of PCA-based methods can vary depending on the complexity of the dataset. Eigenfaces achieved a Rank 1 accuracy of 75% and a Rank 5 accuracy of 81.82%, demonstrating competitive performance. Notably, the proposed IBD-PCA method showcased a promising Rank 1 accuracy of 75.33% and a Rank 5 accuracy of 85.33%. These results underscore the robustness and efficacy of the IBD-PCA approach in handling diverse facial variations, despite the challenges posed by the dataset.

Overall, the findings highlight the limitations of traditional PCA-based methods and the potential of the proposed IBD-PCA approach. By incorporating image band-distributive generation and leveraging the power of PCA, IBD-PCA method exhibits superior performance in face recognition tasks, particularly in achieving highly accurate identification even in complex datasets. These results contribute to the

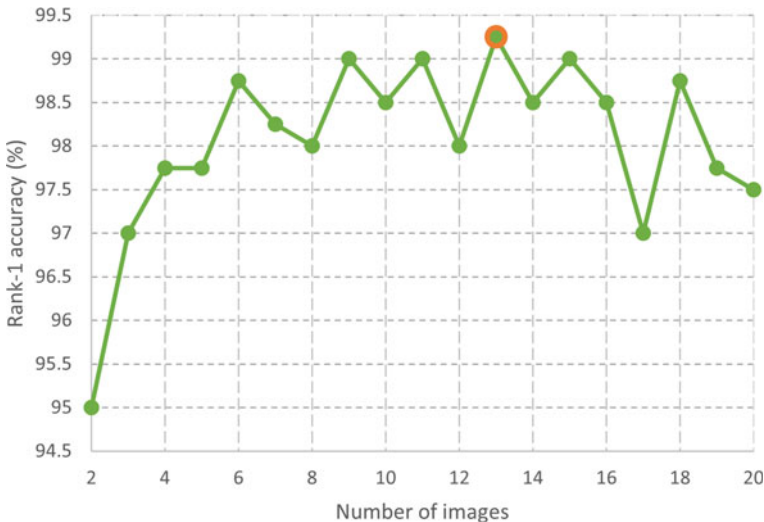


Fig. 4 Rank-1 matching accuracies (%) for PCA, Eigenfaces, and Proposed IBD-PCA method on the ORL Face Database

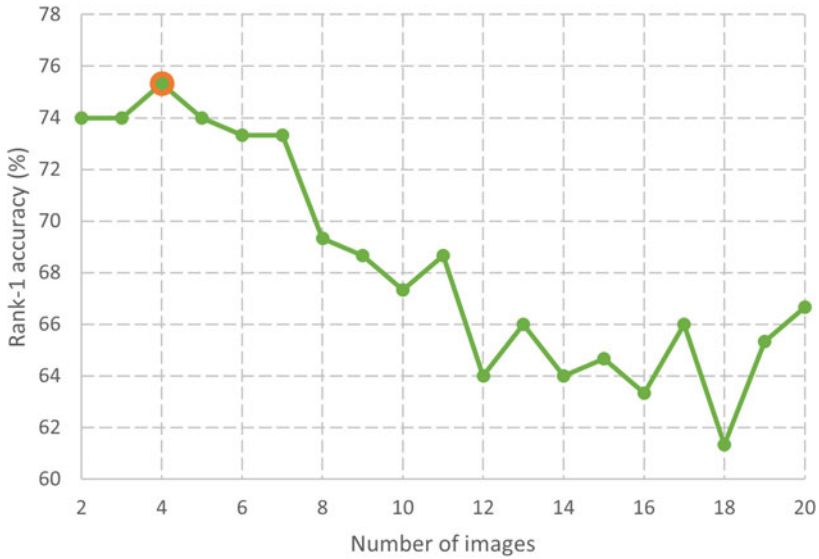


Fig. 5 Rank-1 matching accuracies (%) for PCA, Eigenfaces, and Proposed IBD-PCA method on the Yale Face Database

advancement of PCA-based methods and provide valuable insights for researchers and practitioners in the field of face recognition (Figs. 4 and 5).

4 Conclusion

In this paper, an Image Band-Distributive PCA (IBD-PCA) based method for face recognition was presented. With a face image as input, the proposed algorithm first performs histogram equalization on the image, it then generates multiple images from the resulting histogram-equalized image using the equal-size boundary selection method. The eigenvectors of the resulting images are then extracted by applying standard PCA, and the resulting eigenvectors were then used for feature matching.


The experimental results obtained from comparing the proposed IBD-PCA method with PCA and eigenfaces methods using ORL Face database and Yale Face Database show that the proposed method is superior and outperforms other techniques, giving improved matching accuracy.

References

1. Archana T, Venugopal T (2015) Face recognition: a template based approach. In: International conference on green computing and internet of things (ICGCIoT), Greater Noida, India, pp 966–969. <https://doi.org/10.1109/ICGCIoT.2015.7380602>
2. Puyati W, Walairacht A (2008) Efficiency improvement for unconstrained face recognition by weighting probability values of modular PCA and wavelet PCA. In: 10th International conference on advanced communication technology, Gangwon, Korea (South), pp 1449–1453. <https://doi.org/10.1109/ICACT.2008.4494037>
3. Abbas EI, Safi ME, Rijab KS (2017) Face recognition rate using different classifier methods based on PCA. In: International conference on current research in computer science and information technology (ICCSIT), Sulaymaniyah, Iraq, pp 37–40. <https://doi.org/10.1109/CRCSSIT.2017.7965559>
4. Zhao W, Chellappa R, Krishnaswamy A (1998) Discriminant analysis of principal components for face recognition. In: Proceedings third IEEE international conference on automatic face and gesture recognition, Nara, Japan, pp 336–341. <https://doi.org/10.1109/AFGR.1998.670971>
5. Poon B, Amin MA, Yan H (2009) PCA based face recognition and testing criteria. In: International conference on machine learning and cybernetics, Baoding, China, pp 2945–2949. <https://doi.org/10.1109/ICMLC.2009.5212591>
6. Chellappa R, Wilson CL, Sirohey S (1995) Human and machine recognition of faces: a survey. *Proc IEEE* 83(5):705–741. <https://doi.org/10.1109/5.381842>
7. Winarno E, Al Amin IH, Februariyanti H, Adi PW, Hadikurniawati W, Anwar MT (2019) Attendance system based on face recognition system using CNN-PCA method and real-time camera. In: International seminar on research of information technology and intelligent systems (ISRITI), Yogyakarta, Indonesia, pp 301–304. <https://doi.org/10.1109/ISRITI48646.2019.9034596>
8. Turk MA, Pentland AP (1991) Face recognition using eigenfaces. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, Maui, HI, USA, pp 586–591. <https://doi.org/10.1109/CVPR.1991.139758>
9. Jafri R, Arabia H (2009) A survey of face recognition techniques. *JIPS* 5:41–68. <https://doi.org/10.3745/JIPS.2009.5.2.041>
10. ORL face databases (n.d.) http://www.uk.research.att.com/pub/data/orl_faces.zip. Accessed 28 Mar 2022
11. Yale face databases (n.d.) <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>. Accessed 28 Mar 2022

RAFA Model. Rethinking Cyber Risk Management in Organizations



Jeimy J. Cano M 

Abstract In a highly volatile scenario such as the current one, current cyber risk management practices, based on standards and best practices, begin to lose ground in their effectiveness, given that their scope and proposals are restricted to known and certain scenarios, while the concrete reality of the company is configured in a new abnormality, in uncertain and unknown conditions, which increases tensions in its supply chain, generates instability in geopolitical conditions, warns of disruption with the incorporation of new technologies and implies new conditions to ensure regulatory compliance required by regulators. In this sense, this paper introduces a conceptual and practical model of cyber risk management called RAFA (Resilience, Antifragility, Flexibility and Anticipation) that allows developing a vigilant and active position of organizations as a way to propose alternatives to mobilize the efforts of the organization before the inevitability of failure, making it more resistant to attacks, creating incomplete maps of the reality and challenges of adversaries, and ways to move forward even before the materialization of adverse events.

Keywords Cyberrisk · Resilience · Antifragility · Flexibility · Anticipation

1 Introduction

In a highly volatile scenario such as the current one, organizations need to remain attentive to the instabilities of the context to mobilize their efforts and channel their strategies in order to maintain a privileged strategic position and thus navigate in the midst of uncertainty. In this sense, cyber risk is configured as a systemic and emerging reality that challenges the operating and business conditions of today's companies, since due to the increase in digital density, every key initiative ends up with greater connectivity and flows of data and personal and corporate information,

J. J. Cano M (✉)

Universidad de los Andes, Law Faculty. Cra 1E No.18A-10, Bogotá, Colombia

e-mail: jcano@uniandes.edu.co

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_12

231

which increases the attack surface and therefore, more opportunities for the adversary [1].

In view of the above, current risk management practices, based on standards and best practices, are beginning to lose ground in their effectiveness, given that their scope and proposals are restricted to known and certain scenarios, while the concrete reality of the company is configured in a new abnormality that increases tensions in its supply chain, generates instability in geopolitical conditions, warns of disruption with the incorporation of new technologies and implies new conditions to ensure the regulatory compliance required by regulators [2].

Thus, organizations will have to leave the comfort zone of standards, to become uncomfortable with uncertainties and embark on navigating in the midst of the environment's.

own fragilities. That is, to transform their risk management practice based on continuity, assurance, control panels and knowledge (which seeks to achieve certainties), to one that privileges resilience, antifragility, flexibility and anticipation (which seeks to trace new paths and options) as the natural response to meet the new challenges involved in achieving a position and strategic advantage in the midst of a complex digital world, crossed by systemic risks that require a more holistic and interdisciplinary view [3].

Consequently, this article presents a conceptual and practical proposal for rethinking cyber risk management in organizations, which, understanding the scenario of interaction and coupling of the different elements and components of a digital initiative, offers more than an invulnerability or mitigation response, a way to propose alternatives to mobilize the organization's efforts in the face of the inevitability of failure, making it more resistant to attacks, creating incomplete maps of the reality and challenges of adversaries, and ways to move forward even in the midst of the materialization of adverse events.

To this end, this document is initially developed by situating the new BANI world (Brittle, Anxious, Non-linear and Incompressible) [4] where the abnormal natural conditions that organizations must assume are established, then cyber risks are presented as systemic risks that cross this new reality. Next, these ideas are connected with the false sense of security that is inherent to the use of standards and good practices, and then the proposed model of cyber risk management called RAFA (Resilience, Antifragility, Flexibility and Anticipation) that allows the development of a vigilant and active posture of the organizations. Finally, some conclusions and challenges implied by this model and its implementation in organizations that want to maintain strategic positions in a dynamic and asymmetric digital context are detailed.

2 The BANI World-Brittle, Anxious, Non-linear and Incomprehensible

The global instabilities and the context of polycrisis [5] that the current world is going through establishes more clearly the scenario of a brittle, anxious, Non-Linear and incomprehensible world [4], characteristics that clearly surpass any risk management strategy available to date, since trying to shape a reality that can be considered risky will necessarily have to be situated in a systemic context and embedded in a socio-technical system that changes and mobilizes all the time.

In this sense, cyber crises arising from a successful cyber attack create unprecedented conditions that both organizations and States must begin to understand and address, given the interrelationships and couplings that exist between the different actors involved, which have different characteristics from the effects of a traditional cyber attack. The particular characteristics are typical of systemic risks (which will be reviewed in the next section) and detail cross-border crises typical of a cyberspace and cyberdomain where borders are not delimited. These characteristics are [6] :

- exceed national and even regional capacities to address the crisis,
- defies the traditional time sequence and its effects are not necessarily identified in the same period of time as the explosion of the crisis,
- do not necessarily result in loss of life or pose a direct threat to human lives. However, they can cause significant or even greater damage by undermining the legitimacy of the state and public institutions,
- company's operations are based on deeply intertwined and complex systems that expose societies to serious threats, since a single vulnerability is enough to open the door to a crisis with great potential for escalation,
- exposed to authority vacuums where questions are raised as to who owns the crisis and who has the authority to deal with it.

As it is, organizations face the challenge of the inevitability of a cyber attack where three questions arise on the part of executives:

- How exposed are we to cyber attacks?
- What is the probability that we will have a successful cyber attack?
- How quickly can we recover from a successful cyber attack?

These three concerns pose specific challenges for the security/cybersecurity executive in order to develop metrics that can guide the expected responses and from there make the necessary decisions on the issue. Therefore, when placing cyber risks in a BANI context, it is necessary to overcome the current practices of cyber risk management based on standards and good practices to take up the challenge of understanding and agreeing with the uncertain, in order to recognize from there an extended visual of the business reality where:

- The exposure level is visualized beyond the maturity level of assurance of known practices, i.e. understanding a complex socio-technical system.

- The level of vulnerability is understood beyond technological failures, i.e. recognizing the presence of digital ecosystems.
- Resilience is concrete beyond business continuity practices, i.e., building organizational resilience.

This is a concrete difference to traditional risk-control models such as ISO 31000, or NIST risk management frameworks that try to achieve certainties, overcome vulnerabilities and generally focus on known IT risks. Recognising a BANI environment means moving beyond the day-to-day risk management exercises that take the organisation into a comfort zone, into a more digitally dense and interconnected environment [1], where there is a larger attack surface for adversaries, less visibility of latent risks to risk areas, and a greater impact on successful events that materialise because of the blind spots that remain in the organisations' security and control model.

3 Cyber Risk. A Systemic Risk

Systemic risks are those found in the interaction and coupling between natural phenomena (partially altered and amplified by human action), economic, social and technological evolution and policy-driven actions, both nationally and internationally. In this sense, they are risks that go beyond the usual agent-consequence analysis and focus on the interdependencies and domino effects that may arise from the interaction of their different participants [7].

In particular, systemic risks have the following properties are[8]:

- intersectoral in terms of the scope of their consequences,
- highly interconnected and intertwined, giving rise to complex causal structures and dynamic evolutions,
- non-linear cause-effect relationships that show often unknown inflection points or zones,
- and stochastic in their effect structure.

In this context, cyber risk is systemic and dynamic in nature. It is a liquid risk that changes shape and flows in unexpected ways in the context of a digital ecosystem with high interaction and coupling. In this sense, its management cannot respond to a practice of standards based on known risks and traditional regulatory cycles such as plan, do, check and act. It is a risk, which necessarily leads organizations to leave their comfort zone to rethink their own strategies and practices and thus create new opportunities and challenges to visualize and realize new experiences for customers [9].

This dynamic of risk is reflected in the interaction between the attacker and the analyst, where the movement of one is the trigger for the response of the other and vice versa. It is a cycle of interaction that demands vigilance and anticipation on the part of both actors: While the attacker improves his attack techniques, develops

knowledge and tests the effectiveness of his new developments (of course, without restrictions and limitations), the analyst observes the new attack patterns, improves his defense capabilities and implements his new knowledge as a response and reading of the adversary's behavior (always based on the regulatory and normative framework where he/she operates) [10].

The complexity of cyber attacks shows the inability of organizations to recognize possible patterns of anomalous behavior, which leads to the creation of blind spots in their operation that are used by adversaries to stay off the control radars. Likewise, the aggressors, knowing the need to anticipate their counterparts, are opting to generate "false positives" as a strategy of distraction and generation of instability and chaos, in order to distract attention and achieve their planned agenda. In this way, the maxim that says: "the attacker does not lie, because he has no rules" is confirmed [11].

The dilemma that boards of directors or boards of directors have in the face of cyber risk beyond recognizing and understanding the technological challenges that enable them, involves taking on a leadership challenge in the face of a systemic risk that demands a holistic view of the business in a digital context, which involves governing and managing a new distributed risk environment that affects most organizations to preserve business value and leverage the full potential for value creation in a digital and technologically modified economy [12].

As connectivity and the characterization of new smart objects or devices increase, a new web of connections is defined within the framework of known and unknown relationships that transform the way of doing things and end up changing the reality of customers. In this sense, although spaces are enabled to achieve different experiences, there are also possibilities that may lead to unauthorized uses, since it is personal information that is now mobilized in these new interfaces [13].

In the systemic perspective of social construction, we can indicate, as Luhmann [14] notes that the subject ceases to be the center of the analysis, to focus on the systems and their relationships with the environment. That is, a person makes new indications and distinctions [15] of the interconnected reality to notice new emergent properties of the system he/she analyzes and that, in the scenario of digital density, is enabled by the connectivity and expectations of the different participants of society.

In reading an ecosystem view, it is possible to understand digital density as the enabler of a network of interconnected and interdependent phenomena, where the actions and activities of the participants are relevant to make sense of the dynamics of the interrelated system. In this context, it is understood that greater connectivity emphasizes the values and principles of cooperation, creativity, synthesis, association and life experience, with higher purposes to act in this same way [16], without prejudice to the channeling of these same principles and values, to realize actions openly contrary to the digital ethics of data and in favor of particular interests, affecting the general interest on which the ecosystem evolves.

4 The False Sense of Security. The Comfort Zone of Standards and Best Practices

With the accelerated advance of the implementation of emerging and disruptive technologies, organizations create scenarios of new experiences for their customers, and at the same time motivate new security and control gray zones that can end up being exploited by modified known vulnerabilities and other emerging ones, as a result of the current technological convergence. This reality of exponential connectivity reveals more clearly the coupling and interaction between physical objects and logical implementations, issues that end up defining the dynamics of operation and its reliability [17].

In this scenario, organizations must understand that the inevitability of failure becomes the new normal that they must address and face, since a greater connection between physical objects and logical implementations will necessarily have opaque points of control, which may be capitalized by adversaries now or in the future. Thus, the current mechanics of security and control, which are based on known risks and specific metrics, must be updated to account for this new augmented reality of digitally connected “things” [18].

This creates a zone of strategic instability that has the inherent risk of creating a “false sense of security”, which ends up favoring the need for certainty and productivity, for which investments in technological tools end up being the answer to the challenges of the inevitability of failure. Therefore, when this happens, the organization is weakened in its resilient posture in the face of unexpected events, reacting to these situations in uncoordinated ways, which does not allow greater room for maneuver by attackers seeking to destabilize the company and its business [19].

Thus, the security/cybersecurity executive is at the crossroads of knowing how much to trust the technological tools he has deployed and the practices and standards currently used in the development of security/cybersecurity management and governance. This feeling keeps this manager under constant tension, as he knows that at any moment a security breach could occur and result in instability for the organization, which necessarily means that all eyes will be on him.

Finding the common ground that enables a resilient posture for the organization in the face of unexpected events or emerging adversaries involves recognizing how much one believes in the tools (knowing that they all have known and hidden vulnerabilities) and how much one believes in the current practices and standards (knowing that they all have limitations) [20]. To achieve this, it is necessary for the security/cybersecurity manager to enable a learning culture founded on defend and anticipate, while maintaining and validating known risks through protect and assure.

5 RAFA Model. Rethinking Cyber Risk Management

Considering the current dynamics of the world and the systemic essence of cyber risk, it is clear that current frameworks fall short to understand and address the challenges of this risk. In this sense, it is necessary to move from a traditional risk management focused on the search for certainties to one that understands and takes advantage of uncertainties, that not only complies with the requirements of insurance and regulatory compliance, but also adjusts to the corporate risk appetite and prepares the organization to keep operating despite being successfully impacted by an adverse cyber event [21].

The model proposed below arises from the review of multiple corporate cybersecurity risk management exercises carried out in Colombia, as well as from the review of the generally accepted frameworks used for enterprise risk management, which demand certainty and concrete elements to adjust the decision-making of executive teams in the face of a cyber threat that challenges existing controls all the time, compliance regulations, people behaviours, practices and interconnections with trusted third parties and accountability of board members to stakeholders in the deployment of innovative digital initiatives that meet customer expectations.

In this sense, the RAFA model is proposed, which is based on resilience, anti-fragility, flexibility and anticipation, allowing organizations to focus their efforts on navigating, discovering and taking advantage of the instabilities and uncertainties inherent to business, while taking risks in an intelligent way to account for the company’s promise of value through digital initiatives that fit the risk appetite defined by the company.

The RAFA model is described in Table 1.

The four elements of the model reinforce each other in a way that allows the organization to maintain a vigilant posture towards its environment, so that by recognizing the context in which it operates, the challenges of its business scenario and the capabilities of its adversaries, it is able to learn quickly from those moments when things do not go as planned, to respond to the expectations of its customers in times of

Table 1 RAFA model

Elements	Concept	Key actions
Resilience	Cushioning shocks and operating in the midst of instability, uncertainty and chaos	<ul style="list-style-type: none"> ● Bounce ● Restore ● Renew
Antifragility	Understanding error as a learning window, taking advantage of adverse events	<ul style="list-style-type: none"> ● Vulnerability ● Uncertainty ● Failure
Flexibility	Reconciling strategy and vision in the face of the inevitability of failure	<ul style="list-style-type: none"> ● Audacity ● Adaptation
Anticipation	Imagine and act in the face of different possible and probable futures	<ul style="list-style-type: none"> ● Advance ● Latent risks ● Emerging risks

crisis without exceeding the risk capacity thresholds established and defined by the company’s executive team.

Consequently, organizations move from a traditional posture based on standards and best practices, which responds to a scenario of known risks, to one based on capacity building and identification of patterns that allow them to update their reading of the environment based on scenarios and simulations. See Table 2.

Thus, organizations that assume the RAFA model must address at least three key elements to manage cyber risk, where the challenge will not necessarily be to prevent or avoid, but to assume a specific level of risk to mobilize key digital initiatives for the organization. First define the *risk appetite*, then the *risk tolerance* and finally the *risk capacity*.

Risk appetite is defined as “the amount of risk that an entity is willing to accept or retains in order to achieve its strategic objectives” [22] which generally should contain at least the following components (interrelated and connected) in order to provide the best guidance to the different decision makers in the executive team:

- The stakeholder(s) affected by the materialization of the risk.
- The risk or set of risks that may be involved in the initiative.
- Mission critical systems that may be affected or compromised.
- Vulnerabilities that can be exploited by adversaries.
- Strategic advantage it expects to achieve with the implementation of the initiative.

Based on these components, the organization should answer the following questions to know that its risk appetite statement fits the company’s challenge and size its required *risk tolerance*: [23]

- Do senior decision-makers understand the extent to which they are authorized (individually) to expose the organization to the consequences of an adverse event or situation?
- Do managers understand the aggregate and interrelated level of risk of the components described in order to determine whether it is acceptable or not?
- Do the board of directors and executive management understand the level of aggregate and interrelated risk of the above components to the organization as a whole?
- Is it clear to both managers and executives that risk appetite is not constant and that it changes as the environment and business conditions evolve?

Table 2 Comparison traditional risk model and RAFA Model

Traditional risk model	RAFA model
Certainties	Uncertain
Standards and best practices	Scenarios and simulations
Known risks	Latent and emerging risks
Management and measurement (Feedback)	Defense and anticipation (Feedforward)
Probability × Impact	Uncertainty × Complexity

- Are risk decisions made with full consideration of the strategic advantages to be achieved?

Once this exercise has been carried out, which results in an executive statement with the aforementioned components, the next step is to detail the risk tolerance. Risk tolerance is defined as the threshold of action and maneuver that the organization has to assume key business risks and maintain them even if adverse events occur. *Risk tolerance* is expressed at least in the following terms [24]:

- expectations to mitigate, accept and pursue specific types of risk.
- limits and thresholds of acceptable risk assumption.
- measures to be taken or the consequences of acting beyond the approved tolerances.

The above translates into cybersecurity/security metrics that must be expressed in permitted thresholds of operation or performance of protection and control measures that attempt to quantify the uncertainties inherent to the executives raised around the defined risk appetite. The base structure for the definition of *risk tolerance* considers at least the following topics:

- concerns of executives about the materialization of the risks detailed in the risk appetite.
- The ratios or indices of operational performance (actual) and potential (improvement) of the security and control measures of interest versus risk appetite.
- analysis of behavior and performance of the defined metrics.

With this minimum data, executives will be able to validate whether what was initially defined in the risk appetite is within the conditions and expectations defined by the organization in order to achieve the competitive advantage to be achieved with the digital initiative to be implemented.

Finally, *risk capacity* is defined as the level of preparedness and response developed by an organization in the face of uncertain and adverse events that affect its business model in the medium and long term. This level of preparedness translates into the actual amount of risk that the company could bear based on its financial and operational capabilities [25].

This theme is located in the level of resilient capacity that the organization has developed to overcome any eventuality that takes it out of its comfort zone and calculated risk. In this sense, it is required that the organization has a framework of buffering and concrete flexibility that can maneuver in the midst of the materialization of an uncertain event and continue operating, while it recovers and secures its environment again, to give peace of mind and confidence to its customers, responding in a coordinated and properly communicated manner to all its stakeholders.

This resilient condition can only be achieved if the organization as a whole is articulated from the inevitability of failure and prepared from the understanding of adverse scenarios, which end in simulations where all levels of the company participate to account for the uncertainty, instability and chaos generated by the adverse event and thus show their adversaries the level of preparation that the organization

has achieved, creating a deterrent effect on them, diverting attention for the moment on the initially defined objective.

Thus, the RAFA model places the organisation in its new natural environment: a new “abnormal”, where a greater capacity to understand and anticipate the threats in its environment is required in order to leverage the company from its capacity and risk tolerance to achieve business objectives. Consequently, cyber threats are conceptualised beyond technological challenges, to place them in the business environment, which implies developing four key strategic practices such as: [26]

- Maintain focus—Focus on those features, capabilities and service deployment of technology and security services that are critical to operations.
- Develop testing—Recognise the inevitability of failure as a foundation of security and control practice to sharpen enterprise risk appetite.
- Monitor patterns—Incorporate automated tools to identify people behaviours, adaptive or asymmetric attacks and event correlation.
- Balancing exceptions—Connecting the defined risk threshold and tolerance to requests that require deactivation or downgrading of previously designed controls and their possible compensating controls.

6 Conclusions

The continuing effects of the pandemic, international supply chain tensions, the war in Ukraine and the looming global recession are issues that are and will be present in the executive teams of organizations. In this sense, the exercise of corporate governance will be motivated by strategies and perspectives that allow organizations to move in a shifting terrain and uncertain situations, which will necessarily lead them to understand the current scenario from at least five perspectives: [27]

- Operational and supply chain challenges.
- Technological challenges and cyber threats.
- Organizational challenges: hybrid work and talent retention.
- Environmental, social and corporate governance (ESG) challenges.
- Geopolitical challenges and their global impacts.

This new reality crossed in all its components by the digital environment, creates a zone of permanent instability for organizations that demands a systemic perspective in its treatment in order to explore and understand the new technological ecosystem where organizations operate today. Consequently, cyber risk is a business risk that connects and activates the different strategic initiatives of the company, for which companies must declare their risk appetite based on their strategic objectives and challenges of their business sector.

If the above is correct, the inevitability of failure will be present with possible cyberincidents that may: [28]

- jeopardize the cybersecurity of an information system or the information that the system processes, stores or transmits;
- violate security policies, security procedures or acceptable use policies, whether as a result of malicious activity or not;
- create political, economic or social instability based on fake news or videos, dis-information campaigns on social networks that compromise a country's governance or corporate reputation,

elements that will challenge and test the company's risk tolerance, and from there validate the level of business resilience to buffer, restore and maintain operations despite a successful cyber event. In this way, executive teams must maintain a visual monitoring and oversight of these new realities, not as an additional requirement to be achieved for corporate compliance purposes, but as a natural part of the strategic process that must now be followed in a digital context.

In this sense, the RAFA model recognizes the new "abnormal" of today's organizations based on a BANI world where cyber risks are configured as concrete and real realities that change the dynamics of business. This enables a business transformation process that, placed in the particular context of companies, allows them to: [29]

- Understand the past: recognize known patterns, cycles and risks, and thus learn and capitalize on lessons learned. That is, overcoming checklists.
- Facing the present: observing and acting in the face of new changes and destabilizing forces that allow it to adapt and respond to these dynamics in real time. That is, overcoming the illusion of control.
- Explore the future: navigate in different, possible and probable scenarios that allow you to prepare the company to make decisions and anticipate hitherto non-existent capabilities. That is, to overcome the surplus of the future, which generates anxiety and instability for what is still unknown.

This being the case, the challenge of cyber risk management will be to decrease the attractiveness to the attacker by reducing vulnerabilities, learning from their methods and resolving security incidents within the limits of the organization's capacity [10], an exercise that involves resilience, antifragility, flexibility and anticipation that will lead organizations to overcome traditional risk management methods, not to explain the past, but to improve the future.

References

1. Sieber S, Zamora J (2018) The cybersecurity challenge in a high digital density world. Eur Bus Rev. <https://www.europeanbusinessreview.com/the-cybersecurity-challenge-in-a-high-digital-density-world/>
2. Sheffi Y (2020) The new (ab)normal. Reshaping business and supply chain strategy beyond Covid-19. MIT CTL Media, Cambridge, MA, USA

3. Spitz R (2022) The definitive guide to thriving on disruption. Essential frameworks for disruption and uncertainty, vol 2. Disruptive Future Institute LLC, USA
4. Cascio J (2020) Facing the age of chaos. Medium. <https://medium.com/@cascio/facing-the-age-of-chaos-b00687b1f51d>
5. Colomina et al (2022) The world in 2023: ten issues that will set the international agenda. CIDOB Notes Internationals. No. 238. <https://bit.ly/3YHt7uK>
6. Prevezianou M (2021) Beyond ones and zeros: conceptualizing cyber crises. *Risks Hazards Crisis Public Policy* 12(1):51–72. <https://doi.org/10.1002/rhc3.12204>
7. Renn O (2020) New challenges for risk analysis: systemic risks. *J Risk Res* 24(1):127–133. <https://doi.org/10.1080/13669877.2020.1779787>
8. Renn O, Lucas K, Haas A, Jaeger C (2019) Things are different today: the challenge of global systemic risks. *J Risk Res* 22(4):401–415. <https://doi.org/10.1080/13669877.2017.1409252>
9. Ray A (2022) Liquid risks. The new challenges to global security. Kindle Direct Publishing, Tampa, Florida, USA
10. Zeijlemaker S (2022) Managing the dynamic nature of cyber security. A future-proof strategy, this is how it works. Netherlands. Disem Institute
11. Saydjari O (2018) Engineering trustworthy systems: get cybersecurity design right the first time. McGraw Hill, New York, USA
12. Zukis B, Ferrillo P, Veltos C (2022) The great reboot. Succeeding in a complex digital World under attack from systemic risk, 2nd edn. DDN Press, USA
13. Cano J (2023) The illusion of control and the challenge of an adaptable digital enterprise. *ISACA J 3*. <https://www.isaca.org/resources/isaca-journal/issues/2023/volume-3/the-illusion-of-control-and-the-challenge-of-an-adaptable-digital-enterprise>
14. Luhmann N (1998) Complexity and modernity. From unity to difference. Trotta, Madrid, Spain
15. Brown S (1979) Laws of form. E.P. Dutton, New York, USA
16. Capra F (2003) The hidden connections: social, environmental, economic and biological implications of a new vision of the world. Anagrama, Barcelona, Spain
17. Perrow C (1999) Normal accidents. Living with High-Risk Technologies. Princeton University Press, Princeton, NJ. USA
18. Cano J (2021) Business cybersecurity. Reflections and challenges for 21st century executives. Bogotá, Colombia: Lemoine Editores
19. McKinsey-IIF (2020) Cyber resilience survey. Cybersecurity posture of the financial services industry. https://www.iif.com/Portals/0/Files/content/cyber_resilience_survey_3.20.2020_print.pdf
20. Abraham C, Sims R, Gregorio T (2020) Develop your cyber resilience plan. *Sloan Manage Rev*. <https://sloanreview.mit.edu/article/develop-your-cyber-resilience-plan/>
21. Duane M, Brandenburg R, Gruber M (2018) When the going gets tough, the tough get going. Overcoming the cyber risk appetite challenge. Oliver Wyman. <https://www.oliverwyman.com/our-expertise/insights/2018/apr/overcoming-the-cyber-risk-appetite-challenge.html>
22. Institute of Risk Management IRM (n.d.) Risk appetite and tolerance. <https://www.theirm.org/what-we-say/thought-leadership/risk-appetite-and-tolerance/>
23. Institute of Risk Management-IRM (2011) Risk appetite and tolerance. Guidance paper. https://www.theirm.org/media/7239/64355_riskapp_a4_web.pdf
24. Australian Government-AG (2016) Defining risk appetite and tolerance. Department of Finance. <https://www.finance.gov.au/sites/default/files/2019-11/comcover-information-sheet-defining-risk-appetite-and-tolerance.pdf>
25. Lum R (2016) 4 Steps to the future: a quick and clean guide to create foresight. Vision Foresight Strategy, USA, Honolulu, HI
26. Lam J (2017) Implementing enterprise risk management: from methods to applications. John Wiley & Sons, Hoboken, NJ, USA
27. Lipton M, Rosenblum S, Cain K, Clark H (2022) Thoughts for boards: key issues in corporate governance for 2023. Harvard Law School Forum on Corporate Governance. <https://corpgov.law.harvard.edu/2022/12/01/thoughts-for-boards-key-issues-in-corporate-governance-for-2023/>

28. European Systemic Risk Board ESRB (2020) Systemic cyber risk. https://www.esrb.europa.eu/pub/pdf/reports/esrb.report200219_systemiccyberrisk~101a09685e.en.pdf
29. Boehm J, Kaplan J, Sportman N (2020) Cybersecurity's dual mission during the corona virus crisis. McKinsey Insights. <https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/cybersecurity/cybersecuritys-dual-mission-during-the-coronavirus-crisis>

AI and Ethics: Embedding Good Aspects of AI



Gordon Bowen, Deidre Bowen, and Lisa Bamford

Abstract The buzz around AI systems emphasises the dangers or negatives of AI generative systems. The AI debate needs to move towards the centre ground so that balanced discussion can be achieved. AI systems can make, and are making, many successful contributions to society and individuals, and this should be celebrated but one needs to be wary of the dangers. Governments, society, technology firms and consulting firms are developing and deploying guidelines that can help keep AI systems safe for society. However, there is an urgency to move forward fast because AI is in its infancy and there is an explosion of innovation of new AI systems and applications. Proactiveness is the way to harness the benefits of AI systems so that they are based on ethical AI. This article makes the case for AI as a force for good, but also realises pitfalls lie ahead that need addressing.

Keywords Box–Jenkins model · AI and society · Ethics and microethics · Trust and AI · OpenAI

1 Introduction

The discussion on generative AI tends to overemphasise the negatives and this, to a degree, diminishes the benefits. However, with most technologies there are positives and negatives to their application, and one needs to take a balanced perspective. Understanding the negative impacts of AI will make us more cautious and wary of what could go wrong and help us develop counterbalances to protect the benefits of the technology. Nuclear energy has many benefits but there are drawbacks, however, this has not stopped some governments from harnessing its benefits and being mindful of

G. Bowen (✉) · D. Bowen · L. Bamford
Anglia Ruskin University, Cambridge, UK
e-mail: gordon.bowen@aru.ac.uk

Mental Health, London, UK

Nottinghamshire Health Care Trust, Nottingham, UK

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_13

245

the dangers. AI has the potential to extinguish humankind, just like nuclear energy, but humans have demonstrated they can use it responsibly. Why the over-scaremongering on AI? There will always be actors that use AI and technology applications in a wrongful way, but these can be “controlled” through systems such as cybersecurity. Elon Musk is very vocal about the dangers of AI and calls for the development of the technology to be stopped and for global AI regulation. At the same time Elon Musk is a major investor in OpenAI [21]. Thus, his position is contradictory.

If AI is so dangerous why continue to invest in the technology? Safety is not just about regulations, but about applying ethics and developing an ethics framework and ecosystem so the benefits and dangers of AI can be identified readily and dealt with appropriately.

The aim of this paper is to discuss how the dangers of AI require balancing with the benefits to counter the “stop the world I want to get off” mentality. Thus, AI is a technology that could be embedded in systems, so they behave ethically and are protective of humankind. This paper discusses ethics and AI, developing trustworthy AI, AI is good for society and the Box–Jenkins model as well implication and closes with a conclusion.

2 Ethics and AI

There are no guiding mechanisms to enforce the application of the underlying principles of ethics in the use of AI. However, the impact of a failure of ethics could be wide-ranging: from reputational damage to the person or firm to restriction of the firm’s activities or the individual’s. A failure of ethics has no immediate impact on the firm or individual and, in general, this could be considered a weakness [27]. Ethics guidelines in the AI industry and science suggest that internal self-governance is sufficient, and no specific laws or external governance is necessary to eliminate or minimise the technological risks [7]. Google [25] called for governments and civil society groups to contribute to the AI governance discussion, but little progress has been made. The Partnership of AI (PAI) is an example of companies taking the implications of AI seriously; PAI [38] is a coalition of companies, including Amazon, Apple, Baidu, Facebook, Google, IBM and Intel, researching best practices for the responsible use of AI. PAI [38] was founded on six pillars in relation to the greatest risks and opportunities of AI:

Safety-critical AI. AI tools are used to supplement or replace human decision-making, we must be sure that they are safe, trustworthy, and aligned with the ethics and preferences of people who are influenced by their actions.

Fair, transparent, and accountable AI. Researchers, officials, and the public should be sensitive to these possibilities [hidden assumptions and biases in data], and we should seek to develop methods that detect and correct those errors and biases, not replicate them. We also need to work to develop systems that can explain the rationale for inferences.

We will pursue opportunities to develop best practices around the development and fielding of fair, explainable, and accountable AI systems).

AI, Labor, and the Economy. Discussions are rising on the best approaches to minimizing potential disruptions, making sure that the fruits of AI advances are widely shared and competition and innovation are encouraged and not stifled. We seek to study and understand best paths forward and play a role in this discussion.

Collaborations between People and AI systems. Opportunities for R&D and for the development of best practices on AI-human collaboration include methods that provide people with clarity about the understandings and confidence that AI systems have about situations, means for coordinating human and AI contributions to problem solving, and enabling AI systems to work with people to resolve uncertainties about human goals.

Social and Societal Influences of AI. We seek to promote thoughtful collaboration and open dialogue about the potential subtle and salient influences of AI on people and society.

AI and Social good. AI offers great potential for promoting the public good, for example in the realms of education, housing, public health, and sustainability. We see great value in collaborating with public and private organizations, including academia, scientific societies, NGOs, social entrepreneurs, and interested private citizens to promote discussions and catalyze efforts to address society's most pressing challenges.

However, Hagedorff [27] suggested that membership of such associations could be used as camouflage to provide cover when governments or professional bodies call for legal regulation of businesses' activities. Businesses' vested interests may not always align with societal or governmental requirements. Garzcarek and Steuer [22], Goldsmith and Burton [24] and Johnson [28] suggested teaching ethics to data scientists. However, very little has been written about the implementation of ethical goals and values [27].

The threat of AI to democratic control, governance and political interference in AI are explicitly addressed by the "AI Now 2019 Report" Crawford et al. [10] and Montréal Declaration for Responsible Development of Artificial Intelligence [35]. These documents seem to be the only ones with guidelines that explicitly rule out imposing certain lifestyles or ideas of "good living" on people under AI systems [27]. The use of AI to reduce social cohesion is also criticised, such as isolating people in echo chambers [16]. Hagedorff [27] suggested that there are hardly any guidelines that discuss the political abuse of AI systems in the context of automated propaganda. Guidelines also fail to discuss the lack of diversity in the AI community, which impacts the workplace culture (male dominated and white) that shapes the development and research of AI applications. There are no guidelines on algorithmic decision making, which is superior or inferior to human decision-making processes. Lastly, there are the hidden social and ecological costs of a lack of AI guidelines and its impact on private–public partnerships and industry-funded research.

The USA (California and New York) has the most AI start-ups [41] and China claims it will be the "world leader in AI" applications by 2030 [1]. China has at

its disposal an inordinate amount of data to train AI systems and many companies that could take over supervised machine learning [43]. Research and development in AI systems is driven by the military; militaries are nationalist and are influenced by governmental requirements. Constant comparisons between the AI capabilities of the USA, China and Europe render a fear factor so that none of the actors want to be viewed as inferior [27]. The danger of this mindset is that competitors are seen as enemies and ethics become a bystander in the pursuit of winning. After all, ethics are about rules and are seen as an artificial differentiator [13]. A foundational framework for good AI research and development cooperation and dialogue between firms and research groups is needed [18] and governments are required to be part of the dialogue and cooperation to stop in-groups and out-groups. Hagendorff [27] argued that only by abandoning such thinking (in-groups and out-groups) can the AI race be reframed into global cooperation for beneficial and safe AI.

Will the implementation of ethics guidelines bring about meaningful change? In recent research software engineers were taught ethics guidelines to incorporate in their decision making, but no effect was observed when compared with a control group; this finding indicates that ethics codes or guidelines do not change the behaviour of technical professionals [31]. The monetarisation of AI via many AI applications pushes ethics and ethical decision making into the background. Thus, ethics guidelines and codes will have little impact because of the competitive environment between firms to monetise AI successfully. To overcome these macrosociological barriers requires changes in organisational structures to empower engineers and raise ethics standards. The economic benefits of AI also need to be set alongside principle-based ethics. Organisational structures that enable engineers to raise ethical concerns will help to embed ethical thinking and decision making in AI research and development [27]. The “AI Now 2017 Report” by Campolo and team [8] states that AI ethics and self-governance “faces real challenges” (Campolo et al. [8]:5).

Training in ethics for software engineers requires movement from the generic term of AI to applying ethics to specific datasets. Questions that could help to unpack the specifics of AI training include “What does it mean to implement justice or transparency in AI systems?”, “What does a human-centred AI look like?” and “How can human oversight be obtained?” (Hagendorff [27]:111). Developing a workable training framework requires ethicists to develop technical knowledge. This will require ethicists to reflect on: the process of data collection and the datasets ranging from data generation, recording, processing, dissemination, sharing and use [11]; on the ways of designing algorithms and code [29]; and/or the selection of the training datasets [23]. Ethics that focus on a specific area or application (AI or any other technology) gives rise to “microethics”; so discussion of AI ethics in a broad context would focus on selected areas of ethics and become technology ethics [27]. Gebru et al. [23] have developed microethical training that is easily implementable and works in practice. Their microethical approach enables practitioners to make more informed decisions on the selection of training datasets and avoids the issue of algorithmic discrimination [5]. The ethics theory that will help with the implementation of microethics training is deontology and virtue [27]. The deontological approach to ethics is based on rules that are not open to interpretation and could be applied to AI

ethics [32], whereas virtue is based on the character disposition, personality traits and behavioural disposition of the technologist [30]. Consequently, deontological ethics needs to be augmented with virtue ethics. Ethics training is no longer considered a tick box exercise when augmented with virtue ethics, but a process to change attitudes and behavioural responsibility so that people have the courage and moral compass to refrain from certain actions that are deemed unethical [27]. Furthermore, Hagendorff (2020: 112–113) stated:

I want to resign this negative notion of ethics. It should not be the objective of ethics to stifle activity, but to do the exact opposite, i.e. broadening the scope of action, uncovering blind spots, promoting autonomy and freedom, and fostering self-responsibility.

The application of virtue ethics raises the level of responsibility of every technologist involved in data science, data engineering and data economics relating to AI applications; they all must take some responsibility for their actions [18]. Back-propagation was applied by Floridi [18] to distribute responsibilities: all actions in a group or individual situation require evaluation of the consequences that are intended or not intended and that are morally wrong. This was aptly described by Hagendorff ([27]: 113):

This means that practitioners from AI communities always need to discern the overarching, short- and long-term consequences of the technical artefacts they are building or maintaining, as well as to explore alternative ways of developing software or using data, including the option of completely refraining from carrying out particular tasks, which are considered unethical.

The acceptance of virtue ethics will have consequences for technology communities and will require the adoption of a legal framework, a mechanism for independent auditing of technologies and the establishment of institutions for complaints that can compensate for AI systems that cause harm. Universities' curricula will need to adapt and evolve, especially in the application of ethics of technology, media and information [9, 20].

3 Developing Trustworthy AI

The European Commission developed a set of guidelines on trustworthy artificial intelligence, which was prepared by the High-level Expert Group on Artificial Intelligence [2]. The group identified the following requirements:

Human agency and oversight—including fundamental human rights, human agency (users should be able to make autonomous decisions on AI systems and have the knowledge to interact with, and tools to comprehend, outcomes of AI systems) and human oversight (this precludes AI from undermining human autonomy or causing adverse effects).

Technical robustness and safety—including resilience to attack and security (AI systems are protected against vulnerabilities), fallback plan and general safety (fallback plan in case of problems requiring human intervention), accuracy (AI systems making correct judgements or making correct predictions), reliability and reproducibility (AI system's results are reproducible to prevent unintentional consequences and harm).

Privacy and data governance—including privacy and data protection (guarantee privacy for the lifetime of the AI system, this also includes user data at the beginning and at the end of the lifecycle), and access to data (data handling processes need to be in place for users or non-users of the AI systems).

Transparency—including traceability (documentation of AI decisions, data gathering, data labelling activities and algorithms used), explainability (the ability to explain the technical processes in the context of human decisions, and technical decisions made by AI systems are traceable and understood by humans) and communication (AI systems should not represent themselves as human and humans have the right to be informed when interacting with AI systems).

Diversity, non-discrimination and fairness—including avoidance of unfair bias (AI datasets for training and operations should not include historical data that could be biased and incomplete, and bad governance models. The continuation of these biases could lead to unintended prejudice and discrimination against certain groups), accessibility and universal design (AI systems to be user-centric and access for all people regardless of their age, gender, abilities and characteristics) and stakeholder participation (consult stakeholders to build trustworthiness of AI systems throughout the lifecycle).

Societal and environmental wellbeing—including sustainable and environmentally friendly AI (AI systems should tackle societal pressing concerns in an environmentally friendly way), social impact (ubiquitous exposure to AI social systems in all areas of life and impact on social relationships), society and democracy (the use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including not only political decision making but also electoral contexts).

Accountability—including auditability (the enablement of the assessment algorithms, data, and design processes), minimisation and reporting of negative impacts (both the ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome, must be ensured. Identifying, assessing, reporting and minimising the potential negative impacts of AI systems is especially crucial for those indirectly or directly affected), trade-offs (implementing minimisation and reporting of negative impacts requirements, tensions may arise between them, which may lead to inevitable trade-offs) and redress (knowing that redress is possible when things go wrong is key to ensuring trust).

Political institutions, commercial institutions and academic institutions have all responded to the challenge of AI ethical alignment by creating ethics guidelines for trustworthiness of AI systems [19]. Ethics-based auditing is one suggestion to bridge the gap between ethics theory and practice in AI ethics. In April 2020 leading researchers from Oxford, Cambridge, Stanford, Google, Intel, and others suggested the way forward is to use external auditors for AI systems. The third-party auditors would be tasked with assessing whether safety, security, privacy and fairness-related claims made by AI developers are accurate [4]. Professional services companies, such as PwC and Deloitte, are entering the AI ethics audit by designing and implementing frameworks for trustworthy AI [12, 39].

What is ethics-based auditing? Ethics-based auditing is a governance mechanism that organisations can use to design and deploy AI system. Ethics-based auditing is a structural process by which AI system behaviour is codified based on relevant norms and/or principles. The purpose of the process is to help identify, visualise and communicate whichever normative values are embedded in a system. Different approaches to ethics-based auditing exist: one approach is functionality audits that concentrate on the rationale behind the decision; code audits entail reviewing the source code; and impact audits investigate the effects of an algorithm's outputs [34].

Ethics-based auditing promotes good governance [17]. Mökander and Floridi ([34]: 325) suggested that ethics-based auditing can offer the following contributions:

1. Provide decision-making support by visualising and monitoring outcomes.
2. Inform individuals why a decision was reached and how to contest it.
3. Allow for a sector-specific approach to AI governance.
4. Relieve human suffering by anticipating and mitigating harms.
5. Allocate accountability by tapping into existing governance structures.
6. Balance conflicts of interest, e.g., by containing access to sensitive information to an authorised third-party.

Mökander and Floridi ([34]: 325) made the following points about ethics-based auditing. There is no requirement for ethics-based auditing to be challenging. Ethics-based auditing is a process so that continuous monitoring and evaluation is possible. AI systems do not operate in isolation but are part of a socio-technical ecosystem. Consequently, a holistic approach to ethics-based auditing is preferable because it takes account of the wider knowledge in the socio-technical system. Ethics-based auditing is for the auditors to ask the right questions. An ethics-based auditing process should be harmonised with organisational strategy, policies and staff incentives to ensure an alignment between the organisational policies and the ethics-based auditing framework. The last point is that ethics-based auditing is about design. Thus, “interpretability and robustness should be built into systems from the start. Ethics-based auditing support this aim by providing active feedback to the continuous (re-)design process” (Mökander and Floridi [34]: 325).

Constraints on the ethics-based auditing mechanism include the following [34]:

Conceptual—high level ethics principles lack a consensus, trade-offs between normative values, quantifying externalities of complex AI systems, and reductionist explanations mean information is lost.

Technical—AI systems can appear opaque and hard to interpret, data integrity and privacy are exposed during audits, compliance mechanism cannot be linear with agile software development, and tests may not be representative of real-world AI systems behavioural situations.

Economic and social—audits may disproportionately burden or disadvantage specific sectors or groups, ensuring ethical alignment will require organisational incentives, ethics-based auditing can be impacted by adversarial behaviour, the transformation impact of AI systems could define triggers for audits, and ethics-based auditing may reinforce and reflect existing power structures.

Organisational and institutional—there is a lack of organisational clarity on who audits whom, auditors may lack access to information required for the AI system audit, and the global nature of AI systems will be a challenge for national jurisdiction.

4 AI is Good for Society!

AI is beneficial to society, but it must be handled carefully so that the benefits of AI are also good for society. However, concerns about AI applications have surfaced from self-driving car failures (e.g., a pedestrian was killed in Arizona) [33]. There is concern about engineering pitfalls and the impact of AI systems on government, firms, individuals, and researchers in an unregulated environment in the process of developing strategies to make AI application safe for economic deployment. There is an urgent need for research on safe AI and a consensus on the implementation of new intelligent systems, especially those with a high impact on society and communities [36]. The complexity of machine language, especially deep learning techniques, makes it challenging or impossible to interpret and simplify AI systems into black boxes. Also, there is scepticism about the decision-making process because the lack of transparency of the models hides unfairness in the decision making of the models [40]. The Box–Jenkins model is a good practices model that can be used to overcome the pitfalls of AI and to refine outcomes in the statistical community; it is tool to identify the pitfalls of AI and make AI safer for society [36].

5 Box–Jenkins Model—Safe AI

The European Union [14] developed Ethics Guidelines for Trustworthy AI; a National AI R&D strategic plan was presented by the Executive Office of the President of the United States [15]; the Government of Canada [26] presented a report on guiding

principles of AI; and the China Ministry of Science and Technology presented a report on guidelines for ethical AI [37].

The Box–Jenkins model is an iterative statistical model that goes through all the stages of statistical modelling including regression and time series. The stages of the Box–Jenkins are Identification (data preparation and model selection), Estimation and Validation, and Application. The first stage is to identify the model or algorithm parameters, and these are estimated using appropriate criteria and the assumptions are statistically validated. Within the identification stage is the data preparation and it requires the data to meet required “criteria such as driverless cars must detect pedestrians regardless of their skin color, or a medical diagnostic device for a specific disease must work for both people with high and low socioeconomic status” (Morales-Forero et al., [36]: 687). Model selection is heavily influenced by sample selection, which decides the level of inference that could be drawn from the sample. Training and identifying appropriate samples are time consuming and expensive. Learning from a few datasets that have few anomalies can be used to improve anomaly detection. The estimation process is based on “training” the algorithm and requires optimisation of the training datasets and is an objective process (Morales-Forero et al. [36]: 691). Yao et al. [42] stated that generalisation is reached when the test set error message is small, and the model is accurate based on the test data. The test data and the training data are drawn from the same distribution. However, in practice the distributions are related and not identical.

If the results show the selected algorithm is suitable then the model can be used in the validation phase; if the validation is not successful then stage 1 is repeated. The validation stage is a purely statistical process to validate the statistical assumptions [3]. For the safety of AI systems to be realised requires internal and external validation using the Cabitza and Zeitoun [6] framework within the Box–Jenkins framework.

There are several types of validation [6, 36]:

Internal validation requires validation of the algorithm’s performance in control scenarios. External validation is validation in real-world situations.

Statistical validation refers to the statistical analysis of metrics and their accuracy sensitivity, specificity, robustness and consistency.

Relational validation refers to the interaction of the algorithmic model directly with system users, such as physicians, who directly interact; this helps to understand the system’s useability and human–machine interaction.

Pragmatic validation assesses the functionality of the algorithm in real-world conditions.

Ecological validation is a longitudinal assessment using the pragmatic validation set, which will include all the users, such as from physicians to patients.

6 Implications

AI has many benefits for society, from solving problems quickly to bringing new insights into technological situations. Nevertheless, AI has a dark side which means it can be abused by actors in firms and society. A concern is that guidelines on AI application, from implementation to applications that should be barred, are in an embryonic state. Although many governments are starting to develop guidelines and policies, existing weak guidelines are not being strengthened and refined because of vested interest in self-regulation. Allowing firms to police themselves on AI is a risky approach because AI applications are “easy” to monetise; this is because the technology is in its infancy and first mover advantage enables monetisation to lead to market dominance. AI is seen as the next big thing and thus firms are willing to bet heavily on AI technologies without considering the ethical implications and there are no critical guidelines to hold them to account.

Accountability is pressing due to the many opportunities to which AI technology can be applied. One significant driver of AI systems is the military. What is an acceptable purpose of AI in a military context needs addressing. Will it clearly recognise the enemy? There is the issue of discrimination where an AI drone could have a propensity to attack specific actors on the battlefield (Morales-Forero et al. [36]: 687). Does the military require specific guidelines because of the context AI systems operate in? Should the development of AI guidelines be left solely to the military or will a societal panel of various experts play an active part in developing the military’s AI guidelines? It is expected that the military would like the guidelines to be an exclusively military endeavour, but the threat to society is too serious to contemplate the exclusion of wider representation.

The practice of development, implementation and use of AI applications has very often little to do with the values and principles postulated by ethics. However, the situation is changing with guidelines on AI systems being developed by the EU, Canada, USA and China.

However, the guidelines need alignment so that nations are on the same page. This would bring global and regional regulation a step near, but there is still much to be done. The challenge is to get the major nations (good and bad actors) to work together constructively and with a single purpose. From the authors’ perspective the aim is to make AI systems safe for society and the world whilst enjoying the benefits but minimising the pitfalls and risks associated with AI systems. The regulatory bodies will need to “outlaw” groupthink and develop innovative and workable regulations that harness and do not hinder the many realised and potential benefits of AI. Nations need to accept responsibility for the outcomes produced by AI systems in their country and work across borders to enhance the ethical behaviour of the algorithms. After all, AI is trained and in theory should not depart from the training, but if it needs to evolve so it can make some decisions, then the issue is: is AI willing to act ethically?

AI-based auditing is another response to enhance the reputation of AI systems; it includes guidelines from governmental bodies and professional services such as

Deloitte and PwC. Nevertheless, oversight of the guidelines and deployment of AI systems developed by professional firms will require regulation from professional bodies recognised by governments and independent of the AI-based auditors. A similar principle applies to the self-regulation proposed by technology firms that are members of PAI. Commercial interest could outweigh the societal and ethical responsibilities of these organisations and self-regulation has not been an overwhelming success in the technology industry. However, giving credit where it is due, it is a step in the right direction and a growing recognition of the power and influence of AI systems that needs regulating. It is a much more appealing approach than the “stop the world I want to get off” approach, which would hinder the many benefits and opportunities that AI systems could bring to humankind. Would the military allow AI-based auditing of their AI systems? Pragmatically, there would be pushback and procrastination by the military and government bodies. Is this a sustainable long-term position? After all, we have the international regulatory body the International Atomic Energy Agency (IAEA) that monitors the peaceful use of atomic energy across borders for good and bad actors. Could the IAEA model be a solution for military applications of AI systems? The truth is that AI technology is in its infancy and society and governments do not know the scope of the guidelines required and thus guidelines for unknown areas of AI behaviour will evolve by experience and possibly by AI systems interacting with each other (AI systems training other AI systems). Could AI systems develop guidelines to eliminate bad actors? How do AI systems recognise good and bad guidelines? How can one check the decision-making process of an AI system to ensure decisions are reached ethically based on the guidelines?

The application of a training framework would enable standardisation of training and less opportunities to make ethical errors. This would be dependent on wide adoption of the framework, such as the Box–Jenkins model, and it becoming mandatory training for software engineering students and professionals. However, training of software and technology engineers will require an ethical backdrop, which will require classes in ethics theory, especially relating to virtue ethics and microethics (specific areas of the algorithm such as decision making). Ethical change is behavioural and takes time and hinges on the character of the actors on the course. How does one identify and weed out bad actors (due to their ethical behaviour) on a technological course? Some people have a propensity to do bad things; how does one identify them in the context of technology, particularly with AI systems? Then there is the issue of certification of the engineers in ethical behaviour and how long will the certification last. How do they renew the certification? Do we have courses in technology in which participants are certified to operate ethically in an AI system’s ecosystem and those who are not? How is the training algorithm for courses certified? Can anyone develop the training algorithms? Training will also need to be linked to decision making to understand whether the AI system is behaving ethically. How can this be achieved for large language models? This will require another AI system to analyse and evaluate the decision-making process. Training will be required on the use of AI systems’ interaction.

7 Conclusion

AI systems can be harmful, but the discussion about AI systems has become overemotive and balanced discussion is required. Humankind has the capability to sustain AI and ensure it does not harm or eliminate humanity. Technology can be used by good and bad actors, but this does not preclude the use of the technology. Governments have developed and are continuously updating guidelines on AI applications, and global and regional regulation is very much the focus of discussion between the many stakeholders, such as governments, business and academic institutions. Developing trust and trustworthy guidelines, such as those developed by the European Commission, will go a long way in sustaining societal confidence and trust in AI systems, but this will require time and it should be facilitated by supportive governments and action taken against bad actors. AI systems will help to change future wars. However, society will need to ensure AI systems in the military are regulated within an acceptable framework based on a modified Box–Jenkins model and linked to an ethical AI algorithm; thus ensuring AI is safe for society in any context.

Machines' lack of goodwill or commonsense can lead to unfortunate situations. So, human- in-the-loop or human-assisted systems might be safer than fully autonomous systems.

Although being aligned with human values is a subjective concept, general ethical considerations can be incorporated into the system and constantly evaluated.

References

1. Abacus (2018) China internet report 2018. <https://www.abacusnews.com/china-internet-report/china-internet-2018.pdf>. Accessed 6 July 2023
2. AI HLEG, European Commission (2019) Ethics guidelines for trustworthy artificial intelligence. AI HLEG, European Commission. <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html>. Accessed 20 July 2023
3. Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control. Wiley, New York
4. Brundage M et al (2020) Toward trustworthy AI development: mechanisms for supporting verifiable claims. <https://arxiv.org/abs/2004.07213>. Accessed 21 July 2023
5. Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of machine learning research conference on fairness, accountability, and transparency, vol 81, pp 1–15
6. Cabitza F, Zeitoun JD (2019) The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. *Ann Transl Med* 7(8):161. <https://doi.org/10.21037/atm.20>
7. Calo R (2017) Artificial intelligence policy: a primer and roadmap. Available at SSRN: <https://ssrn.com/abstract=3015350>
8. Campolo A, Sanfilippo M, Whittaker M, Crawford K (2017) AI now 2017 report. AI Now Institute, New York. <https://experts.illinois.edu/en/publications/ai-now-2017-report>. Accessed 6 July 2023
9. Cows J, Floridi L (2018) Prolegomena to a white paper on an ethical framework for a Good AI society, pp 1–14. <https://ssrn.com/abstract=3198732>

10. Crawford K, Dobbe R, Fried G, Kazianus E, Kak A, Mathur V, Richardson R, Schultz J, Schwartz O, West SM, Whittaker M (2019) AI now 2019 report. AI Now Institute, New York
11. de Bruin B, Floridi L (2017) The ethics of cloud computing. *Sci Eng Ethics* 23:21–39. <https://doi.org/10.1007/s11948-016-9759-0>
12. Deloitte (2020) Deloitte introduces trustworthy AI framework to guide organisations in ethical application of technology. Press Release, New York. <https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/deloitte-introduces-trustworthy-ai-framework.html>
13. Derrida J (1997) *Of grammatology*. Johns Hopkins University. Press, Baltimore
14. European Union (2019) EU Guidelines on ethics in artificial intelligence: Context and implementation (europa.eu). Accessed 8 July 2023
15. Executive Office of the President of the United States (2019) The national artificial intelligence R&D strategic plan. Retrieved from <https://trumpwhitehouse.archives.gov/wp-content/uploads/2019/06/National-AI-Research-and-Development-Strategic-Plan-2019-Update-June-2019.pdf>
16. Flaxman S, Goel S, Rao JM (2016) Filter bubbles, echo chambers, and online news consumption. *PUBOPQ* 80(S1):298–320
17. Floridi L (2014) *The 4th revolution: how the infosphere is reshaping human reality*. Oxford University Press
18. Floridi L (2016) Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions. *Philos Trans Ser A Math Phys Eng Sci* 374(2083):1–13. <https://doi.org/10.1098/rsta.2016.0112>
19. Floridi L, Cows J (2019) Unified framework of five principles for AI in society. *Harvard Data Sci Rev*. <https://doi.org/10.1162/99608f92.8cd550d1>
20. Floridi L, Cows J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach* 28(4):689–707. <https://ssrn.com/abstract=3284141>
21. Forbes (2023) Elon Musk has issued a stark warning over AI. This isn't his first time. *Forbes*. <https://www.forbes.com/sites/qai/2023/02/16/elon-musk-has-issued-a-stark-warning-over-ai-this-isnt-his-first-time/>. Accessed 6 July 2023
22. Garzcarek U, Steuer D (2019) Approaching ethical guidelines for data scientists. *Comput Sci*:1–18. arXiv: 1901.04824v1
23. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumeé III H, Crawford K (2018) Datasheets for datasets, pp 1–17. <https://arxiv.org/abs/1803.09010>
24. Goldsmith J, Burton E (2017) Why teaching ethics to AI practitioners is important. *ACM SIGCAS Comput Soc* 110–114. <https://doi.org/10.1609/aaai.v31i1.11139>
25. Google (2019) Perspectives on issues in AI governance. Google. <https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf>. Accessed 7 July 2023
26. Government of Canada (2021) Responsible use of artificial intelligence (AI). <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html>. Accessed 8 July 2023
27. Hagedorff T (2020) The ethics of AI ethics: an evaluation of guidelines. *Mind Mach* 30:99–120. <https://doi.org/10.1007/s11023-020-09517-8>
28. Johnson DG (2017) Can engineering ethics be taught? *The Bridge* 47(1):59–64
29. Kitchin R (2017) Thinking critically about and researching algorithms. *Inf Commun Soc* 20(1):14–29
30. Leonelli S (2016) Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. *Philos Trans Ser A Math Phys Eng Sci* 374(2083):1–12
31. McNamara A, Smith J, Murphy-Hill E (2018) Does ACM's code of ethics change ethical decision making in software development? In: Leavens GT, Garcia A, Păsăreanu CS (eds) *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering—ESEC/FSE 2018*. ACM Press, New York, pp 1–7

32. Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. *Nat Mach Intell* 1(11):501–507
33. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG (2015) Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533
34. Mökander J, Floridi L (2021) Ethics-based auditing to develop trustworthy AI. *Mind Mach* 31:323–327. <https://doi.org/10.1007/s11023-021-09557-8>
35. Montreal Declaration for a Responsible Development of Artificial Intelligence (2018). Monoskop. https://monoskop.org/images/d/d2/Montreal_Declaration_for_a_Responsible_Development_of_Artificial_Intelligence_2018.pdf
36. Morales-Forero A, Bassetto S, Coatanea E (2023) Toward safe AI. *AI Soc* 38(2):685–696. <https://doi.org/10.1007/s00146-022-01591-z>
37. OECD AI policy observation (2021) AI in China ethical norms for new generation ai policy—OECD. Accessed 8 July 2023
38. PAI (n.d.) About us. Advancing positive outcomes for people and society. Partnership on AI. <https://www.partnershiponai.org/about/>. Accessed 6 July 2023
39. PwC (2019) A practical guide to responsible Artificial Intelligence (AI). <https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/responsible-ai-practical-guide.pdf>
40. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1:206–215. <https://doi.org/10.1038/s4256-019-0048-x>
41. TRUiC Team (2023) 30 Top AI startups to watch in 2023. TRUiC Startup Savant. <https://startupsavant.com/startups-to-watch/ai>. Accessed 6 July 2023
42. Yao L, Chu Z, Li S, Li Y, Gao J, Zhang A (2020) A survey on causal inference. <https://arxiv.org/abs/2002.02770>. Accessed 10 July 2023
43. Yuan L (2018) How cheap labor drives China’s A.I. ambitions. www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html. Accessed 30 Nov 2018

Cyber Security-Related Awareness in Bangladesh: Relevant Challenges and Strategies



Kudrat-E-Khuda

Abstract Coupled with numerous blessings, the technology has also brought huge security concerns for users in the current digital era thanks to its evolving nature which has compelled us to create a cyber-environment. Like other parts of the globe, there was a significant boost of technology in Bangladesh in recent decades. Concurrently, it also endangered the country with the threat of cybercrimes and challenges to keep its cyberspace safe. Bangladesh government and other concerned stakeholders have already taken numerous efforts, including awareness-raising programs, so far to face the evolving challenges and threats. Amid the rising trends and frequency of cyberattacks, people in the country are more frightened of cyber threats than nuclear weapons or environmental destruction. This study finds that the prevention efforts to face the attacks and threats are not adequate compared to the severity and the frequency of the cyberattacks. Cybersecurity laws and procedures are unknown to the majority of the people in the country. The government and concerned organizations are not even worried enough about cybercrime-related issues. They should come up with pragmatic steps and bring a massive change to their systems to strengthen cybersecurity-related policies and procedures. The authorities concerned also should enhance vigilance and formulate new policies, if necessary, to address the cyber menace. This study suggests a thorough investigation into the degree of cyber security awareness among Bangladeshi nationals. To this end, this paper also investigates briefly the understanding of cyber security awareness as well as the route background strategies in the country. It suggests ideas and strategies that should be used to raise awareness among people to provide a stable, safer, and more effective cyberspace at all levels.

Keywords Bangladesh · Cyber security · Cybercrimes · ICT · Policy · Technology

Kudrat-E-Khuda (✉)

Department Law, Daffodil International University, Dhaka, Bangladesh

e-mail: kekbabu@gmail.com; kekbabu.law@diu.edu.bd

1 Introduction

The number of new users of smartphones and the internet considerably has increased in developing countries like Bangladesh due to the rapid growth of information and communication technology (ICT) as well as the region's economic development. In such a situation, it is very crucial to develop cyberspace, one of the most important factors for the development of a country, for internet users. This is why, a strong cyber security system is crucial to ensure the security of the cyber stakeholders as well as the progress and development in the technical, social, and economic sectors. So, the nations should introduce CCB into their development projects to assist in promoting their growth, and development and safeguarding them. Bangladesh, a nation of 160 million population with only about 55,000 square miles of land, has achieved remarkable economic growth in the sectors of science and versatile information and communication technology despite having not many natural resources [1]. The developed and industrialized countries of the world have made great strides in science and technology and by utilizing them they have already increased national and economic growth and ensured social security. Meanwhile, developing countries have largely lagged in these areas due to their insufficient knowledge and research about science and technology and their potential. However, the ICT sector is now hovering new hope and creating new possibilities for developing countries as these countries are moving towards growth and progress by exploiting their potential intelligently. For example, the information and communication technology ministry and the ICT division of the Bangladesh government have already given the highest priority to the applications of ICT to reduce poverty by improving education among poor people living in rural areas [2]. The country has planned to increase productivity and quality along with education by taking various timely initiatives and creating a conducive environment, which will result in the qualitative development of the common people and contribute to nation-building. As a result, rural people will have access to information needed to produce quality products and market them. They should be properly trained to perform their duties effectively. In recent years, the Bangladeshi Government has taken several steps to advance the ICT sector and enhance scientific and technological research. To this end, Bangladesh should render services as strictly as possible in light of international codes and guidelines. Bangladesh's government has already expressed its commitment to extend its growth through enhancement of the scientific innovation and technological development by establishing a powerful National Task Force on Biotechnology, Information, and Communications Technology, and the National Council for Science and Technology. As a secured cyber system is essential for the development of the ICT sector as a country's development, having adequate knowledge and expertise on information and communication technology and cyber system are also crucial for ensuring cyber security. Bangladesh's government is campaigning for IT promotion and cyber security awareness in collaboration with universities, non-profit organizations (NPOs), non-governmental organizations (NGOs), and private businesses, and other stakeholders. However, it is very important to increase awareness among the users and stakeholders to check cybercrime as well

as to promote the endangered groups for taking privacy measures and adopting the best practices to ensure cyber security. Workers of different organizations, political activists, industrial and private company employees; students, teachers, parents, and other professionals, etc. are among the risky and vulnerable groups [4]. The most important thing is to spread knowledge about cybercrimes and security measures among users. Bangladesh has experienced recent rapid growth, particularly in the areas of information and communication technology apart from economic gain. Data from the State Statistical Office for 2022 showed that the use of ICT is expanding rapidly. As per its data, the Internet is accessible to 69.4% of households and 93.5% of business entities having 10 or more employees. The growth of internet access through broadband connections is another significant figure. The country needs to update its information technology and cyber security environment and enhance skills in some manner, as evidenced by the growing use of ICT. Even though the government has allocated funds for infrastructure security, it is very crucial to have all-out preparation for ensuring the security of data and information. The end users are the most vulnerable to cyber security risks thanks to having less knowledge of cyber security in the current period's global, organizational, and social dimensions. Many end users even do not know how it may significantly affect them personally and their entities and even the country [3]. To this end, the constructive action plan should be adopted aiming to achieve this goal of creating a culture of cyber security and raising consciousness by attempting to maximize awareness of the weakest security link. In light of the above discussion, this paper will first highlight the government initiatives and current policies to create awareness among stakeholders. Then it provides a summary of the projects and initiatives undertaken by NGOs, NPOs, and educational institutions. The study will come up with some suggestions on how to raise and alter awareness of cyber and IT security to safeguard Bangladesh's weak points. Finally, this article will have a sound conclusion on the awareness initiatives and strategies of Bangladesh.

2 Problem Statement

This article shed light on cyber security awareness and strategies and challenges in Bangladesh, a developing nation that is presently experiencing an ICT transformation. Bangladesh is now embracing information and communication technology and significantly increasing its use of the Internet as well as ICT. Although the country has great expectations for the successful outcome of its investment, Bangladesh has not yet seen any return on its investment. This study demonstrated that the nation's policymakers require more specific direction regarding how to secure cyberspace. It has come to the conclusion the present situation does not suit the present guidelines. There should have a deeper understanding and proper knowledge to address the problems with cyber security in Bangladesh. The issues at hand must be better defined, and the situation must be examined to determine how to overcome the difficulties and where more study is required.

3 Literature Review

A. Bangladesh's current level of cyber security awareness

As a part of the number of policies to safeguard its cyber system, the Bangladeshi government established the National Council for Science and Technology (NCST) to expand and implement science and technology development activities to raise the quality of living for the general populace. The country also formed NCST Executive Committee which has been given the task to carry out the council's policy recommendations. Thanks to the most recent National Information and Communication Technology Policy (2002), the nation has been focusing on the enormous ICT growth potential to establish effective governance, integrate ICT-related policies, allocate resources for software development projects independently, and produce top-tier ICT specialists in addition to earning multiple billions of dollars from the software export market. By 2024, the country has the plan to build an ICT-driven country with a knowledge-based society. To accomplish this, a national ICT infrastructure will be built to ensure that every citizen has access to knowledge in order to promote empowerment [7]. This goal will be accomplished through the use of strategic human resource infrastructure, good governance, e-commerce, banking, public utilities, and all types of online ICT-enabled services to ensure sustainable economic growth and enhance democratic values. The development of the ICT sector, infrastructure development for ICT, support for ICT research and innovation, and economic and social growth are all prioritized in the national ICT strategy of Bangladesh. To capitalize on the growth opportunities in the rural and agricultural sectors, the country will also use ICT in agriculture as well as cover the significance of the hardware industries, e-commerce, e-governance, ICT-related legal problems, and utilize ICT in the healthcare sector. The country also emphasizes how ICT is being integrated into other fields like social assistance, transportation, and the legal system. The Model Law on Electronic Commerce was approved by the United Nations Commission on International Trade Law (UNCITRAL) in 1996 [21]. This model is known as the UNCITRAL Model Rule for e-commerce. Bangladesh approved an ICT legislation in February 2005 that was drafted in light of the UNCITRAL Model legislation in order to promote electronic commerce and advance information technology. The country's ICT law establishes the rules and specifications for the authentication and acceptance of contracts, forms submitted electronically, default rules for contract establishment, and management of the performance of e-contracts. In the law, the attributes of a genuine electronic document are specified. Additionally, computerized mechanisms were added by amending the Copy Right Law 2000. The government still needs to make a concerted effort to lessen inequality, improve livelihood security, end hunger and malnutrition, and generate employment. To achieve this, it will be necessary to develop and evaluate all pertinent technologies, to quickly disseminate them through networking, and to enlist the support of sizable, unorganized economic activity sectors.

In recognition of the importance of ICT and its enormous impact on people's lives, the government also renamed the Ministry of Science and Technology the Ministry

of Science and Information & Communication Technology. The Ministry of Sciences and ICT has been assigned to ensure the orderly growth of this sector in Bangladesh. The Bangladesh Computer Council (BCC), the primary organization in charge of supporting all ICT-related activities in the nation, is also in charge of regulating the Ministry of Science and ICT. The development of the telecommunications sector is necessary for the development of research and ICT. This industry is still in its infancy due to a dearth of deregulation and unfettered competition. In such a situation, the government established Bangladesh Telecommunication Regulatory Commission (BTRC) in 2002 which functions as an independent telecommunications regulatory authority.

B. Cyber security perception stability in Bangladesh

There are restrictions on information available in many developing nations, including Bangladesh. Besides, getting access is not economically viable due to the inadequate state of the current infrastructure and the absence of the necessary schooling. The lack of a comprehensive infrastructure for information safety and sufficient cyber security training and study are among the challenges [5]. To ensure a guarantee for addressing safety concerns, cooperation, collaboration, and security investment are also necessary. These factors also produce a culture of security standards. In business and other operations, confidence is essential, and it can be built when professionals are confident that the terms of the agreement are being upheld. As a result, from a business standpoint, protection needs to be seen as a strategic partner. However, one of our problems is the absence of a suitable framework for information security, and one of the most significant problems is the lack of understanding of cyber security. Bangladesh is getting ready and developing security standards as important next steps for increasing awareness and the accessibility of vital intelligence. But one of the major challenges of the country is a lack of a proper framework to ensure information security [16]. Apart from that, awareness regarding cyber security is one of the crucial issues. Bangladesh is getting ready to set and develop security standards, take significant steps of increasing awareness, and enhance the knowledge on the issues accessibility of required study and knowledge [22]. There is a need for additional implementation procedures for information system security standards. Global collaboration is essential to achieving these goals. We understand that research and development are more crucial to ensure the security of the information security system, but we also need to take the lead in a successful program that fosters understanding and cooperation, promotes risk management techniques and best practices, and pursues standardization initiatives through a general understanding of evolving technology around the world. Bangladesh as a nation has an understanding that secured information security of a country is also a key parameter for ensuring ease of doing business and encouraging foreign investment international and industry-wide cooperation as well as making efficient public-private partnerships. Thus, urgent and effective legal frameworks and regulations are needed to insure cyber security. There is also a need for necessary national training and awareness-raising efforts for the enforcement of cybercrime policy programs. In order to encourage customer trust, security controls, trust points, and other self-regulatory procedures should be created for the creation

of goods and the delivery of services. These procedures should also be implemented along with the necessary measures.

C. Recent Cyber Attacks in Bangladesh

Numerous commercial and service-providing entities across the country come under cyberattacks frequently. Those attacks are increasing at a time when the nation is registering a substantial advancement in digital technology and the socioeconomic sector. As such incidents are multiplying, the topic of cyber security comes as a matter of great concern for both public and private organizations. The number of cyberattacks and related incidents increased by more than 31% year-on-year in 2018 [20]. About 870 incidents took place in 2018 while the number of such incidents was 683 in 2017, the Ministry of Posts, Telecommunications, and Information Technology's government-run Bangladesh e-Government Computer Incident Response Team (BGD e-Gov. CIRT) provided data. In 2016, a total of 379 incidents occurred 2016. According to the government data, a vulnerability in the cyber security system accounts for 63.2% of the total number of attacks while the incidents of attack or hacking account for 5.7%, malicious code 22.5%, offensive materials 4.5%, and the rest of the incidents happened due to fraud, attempted invasion, request for service, identity security, and others [19].

However, the real number of such attacks would be much higher but it is not possible to ascertain due to a lack of proper statistics as the state-owned organization does not disclose such incidents that occurred in commercial or service-providing sectors, according to the business insiders. After the much-talked-about incident of reserve heist from the swift account of the Bangladesh Bank at New York Federal Bank in the US, the Bangladeshi government established "BGD e-Gov. CIRT" under the auspices of the Bangladesh Computer Council (BCC). It was created to bolster defenses against any such fatal intrusions [18]. Later, all banks in the country have been asked to take an additional level of security measures after the Bangladesh Bank alerted the lenders of being hit by a new cyberattack by North Korean-based hackers. According to a source in the Bangladesh Bank, the warning followed a cautionary alarm from the US Federal Reserve Bank that was subsequently disclosed by the central bank. Multiple bankers claim that as a result of the warning, some banks with sizable digital banking networks have scaled back their online activities. Mutual Trust Bank's managing director, Syed Mahbubur Rahman, remarked they received a warning signal from the central bank over a week back regarding the possible hacking attempt by a group of North Korea-based hackers. A group of hackers were trying to penetrate the US banking system recently. Sensing the matter, the authorities concerned with the US banking system sent letters to all central banks so that they can take preventive measures, he said. He also added that they had taken all-out preparation to stop the possible cyber-attacks on their IT system in the wake of the US alert. Sources at Dutch-Bangladesh Bank, the largest ATM booth network in the country, said as a part of their precautionary measures they suspended transactions at all booths after 11 pm for some period.

The country's banking system started to face the threats of cyber-attack at a time when this sector was rapidly extending thanks to the rapidly expanding digital

banking network since the middle of the COVID-19 pandemic. In February 2016, \$81 million has been a theft from Bangladesh Bank's New York Fed foreign currency reserve fund, which was the largest cybercrime ever [15]. After an investigation, FED revealed that a North Korean-based hacker group was involved in that bank heist. After the heist, Bangladesh Bank has already taken many steps to bring back the stolen money but still could not yield a fruitful result. Meanwhile, online financial services have been expanding in Bangladesh during the pandemic when people had to stay at home during the shutdowns and largely depended on online services. Data from the central bank showed a significant amount of money was transacted through the internet. It shows in June 2020, payments made through Internet banking surged by 12.6%. Online banking transaction was Tk6,588 core in March, the month the government discovered the first coronavirus case in the country [13]. By June of this year, monthly transactions reached about Tk 7,421 core. This rise in online banking, which comes at a time when the disease is on the rise, is the outcome of a major shift away from traditional banking toward digital banking. However, many Bangladeshi banks have significantly increased their digital networks, raising security issues, according to the bankers, because of a shortage of competent labor and lesser spending on IT security.

4 Methodology

Research methodology is defined as a number of procedures that are required for carrying out research or any systematic investigation. The investigative research approach was used for this work, and it is based on a thorough analysis of the literature and the calculation of secondary sources of information or data. It is essential to gather the necessary data from the identified records and sources regarding the subject-related issues in order to carry out this research. However, this research investigates the cyber security literature and resources by interacting with facilitators and systems. As a result, this article will pursue qualitative research to analyze and comprehend the knowledge and policy of Bangladesh's cyber security. In conclusion, a study analysis yields suggestions for various industries and fundamental laws. These recommendations include the requirement to amend some legal frameworks and regulations. To guarantee the privacy of individuals when their private data is collected, stored, processed, and transmitted and to provide end-users, some regulations need to be reviewed or new legislation needs to be properly implemented. The government needs to keep up its current awareness campaigns and carry out more training sessions on data protection. The nation will need capable and skilled labor to develop a strong infrastructure and technology to accomplish the necessary level of cyberspace security to carry out these activities.

5 Cyber Security in the Global Village

People are now extremely linked from one part to another part of the globe due to advancements in information and communication technology and can have contact with the people from other edges of the world like their neighbors. Not only that, multiple people can instantly join together from various locations around the world at a time. As a result, a new term known as “netizens” has developed among the world’s citizens, and now the world turned into like a village where netizens act like villagers or citizens. This is how the term “global village” originated. Meanwhile, along with bringing revolutionary changes across the globe, the internet and communication and information technology have also brought some security concerns for users as well. Now people from all walks of users across the globe are prone to cyber threats [9]. However, this concern is not only for the users, their countries are also at a security risk similarly. But these security concerns or cyber threats have not got priority in the national security policies of the countries with the same importance. Yet the matter of security concerns has not been an issue of global concern as well. Digital transactions account for the majority of global financial transactions of a day. This is why, hackers are constantly looking for ways to sneak into financial systems in order to steal money by exploiting the financial networks and systems. B. Williams claims that there are four different kinds of cybercrimes. Cybercriminals are primarily concerned with making money. For instance, in April 2013, a hacked Twitter news stream that circulated a false rumor about an explosion at the White House caused the US stock market to lose \$130 billion in a matter of minutes [17]. Meanwhile, rival groups of cyber criminals prey on one another in their competition for sensitive information or intellectual property, creating security concerns. The greatest known collection of stolen Internet data was amassed by a Russian criminal organization, and it includes 1.2 billion username and password combinations for more than 500 million email addresses. De facto insiders also appeared as threats in many cases. The incident involving the hacking of the Iranian nuclear program’s IT system and the disclosure of American diplomatic cables highlighted how vulnerable the current system is and emphasized the importance of ensuring cyber-security. Although the harm and losses caused by cybercrimes are not proportional on a worldwide scale, the crimes have gotten worse because of the flaws in cyberspace security. Cybercrime frequently portends the tensions that will inevitably result from the intimate contact between various cultural activities via the internet. Worldwide networks of people have been formed as a result of the information and communication technology revolution, linking people at different organizational levels. The process of communication between organizations and inside the government used to be extremely sluggish and expensive, but the existence of the internet has transformed the entire process to an unmatched level in terms of medium, speed, and cost [14]. There is a number of software that is used to send any information to a location without an information-transmission network. Users also used to pay for the software as it is really useful for them. Meanwhile, due to the increasing frequency of cybercrimes, many nations and companies are now exercising caution, viewing cyber security concerns as a threat,

and striving to address them as well. Data breaches are a sort of cybercrime that is very widespread that involves stealing data from different entities and businesses. Many organizations are targeted by outsiders or hackers who attempt to steal classified data. Most frequently, this kind of hacking happens when data is moved from one organization to another. Not only did this kind of criminal action occur domestically, but it also occurred internationally. The cross-border nature of data breaches may make it difficult to resist taking an overwhelming approach to both the investigation and the division of breach management alternatives [11]. Asia–Pacific nations are seen formulating different rules and regulations to ensure digital security in recent years. Many countries are setting up different organizations, regulatory bodies, or authorities to strengthen vigilance in their cyberspaces and handle cyber security risks as well as publishing frequent guidelines and circulars. East Asian nations, including Singapore and Indonesia, established cyber agencies in 2015, while Japan ratified the Cyber Security Basic Act. Asia Pacific countries have suddenly begun to draft legislation and/or regulations in an effort to protect their cyberspaces. Meanwhile, developed countries especially Western countries have already taken many effective steps to ensure the security of their cyberspace. It is already learned that the Australian Securities and Investments Commission has issued cyber resilience. In the meantime, the Justice Department of the United States published a directive in April 2015 titled “Best Practices for Victim Response and Reporting of Cyber Incidents.” Many other nations are likewise expanding their current frameworks to include cyber security rules. Although many Asia–Pacific countries have taken different steps to ensure cyber security as their own, still there is a lack of intensive measures in formulating related legal frameworks or cordial efforts of taking legal action among the nations in the region against the responsible people or groups of stealing data. Besides, the issues of data breaching can be included in a number of already-existing laws and regulations in addition to cyber security laws. It might be integrated with labor regulations and data protection legislation. It can also be incorporated into laws and requirements including corporate governance, equity laws, and fiduciary responsibilities. In some circumstances, when any data goes outside of the capacity of any authority, the laws or legal frameworks of a state take effect [14]. The duties of governments and how authorities or courts handle data breaches must also be understood on a local level. Additionally, it’s critical to understand the best legal defenses against cyber threats. Once equipped with the knowledge, the victim of a cyberattack or someone who could become a target can verify his system for data leaks. After then, it might be able to divide the responsibilities and come up with a strategy to prevent more data leaks. They can manage data breaches, understand their effects, and develop practical and legitimate ways to retrieve lost data or mitigate damage from breaches. Many state-run websites in many cases take support from foreign servers and suppliers. This type of support may put the states in danger of hacking by system insiders. Similarly, to address different types of cyber-related issues many governments take help from experts from abroad, putting their cyberspace in a vulnerable position. The security of a nation could be most at risk from the fourth category. This is a state-sponsored cyberattack. Such types of attacks are carried out with the aim to compromise the state’s security system.

For example, this type of attack is carried out at vital infrastructure or important national economic components aiming to gain a strategic advantage. Here we can take China as an example. China is considered a threat to cyber security by many powerful countries like the United States, the United Kingdom, France, Germany, and India. These countries have been accusing China of espionage to obtain strategic advantages. However, there are some reasons to consider China as a threat to the above-mentioned nations. In 2007, China carried out a series of cyberattacks against those countries [6]. However, these nations have more ambitious plans to increase their military's capability to engage in information or cyberwarfare if it is necessary in the future.

6 Challenges to Bangladesh and Bangladesh Government

Bangladesh is the most vulnerable country in the globe in terms of cybercrimes. One of the reasons behind its vulnerability is that the majority of the software used here is pirated. Apart from that the poor infrastructure of the country's cyber system is another reason. To this end, the South Asian country's prime challenge is to protect cyberspace. Here more than 90% of software is pirated. Typically, when pirated software is utilized and used, it creates loopholes or security flaws by which criminals can infiltrate the cyber system easily. However, there are no concerns among the users and stakeholders about the use of pirated software in Bangladesh although this is the main cause of the nation's cyber security system's vulnerability. It is impossible to overlook the implications and impact of this issue because many users pay the price when they have no other option than to become victims of hackers. Another big challenge for Bangladesh is that the criminals who conduct cybercrimes usually stray out of their reach and they could not be brought to justice for their crimes. Figures from the Bangladesh Telecommunication Regulatory Commission (BTRC) showed that there were 90.5 million active internet users in Bangladesh as of August 2018 [7]. In a single month, additional 1.8 million internet users are added to the country's network. Mobile internet is used by 84.7 million of them. Only 5.73 million of them use fixed broadband internet connections and the rest users depend on WiMAX connections. About 70 million active internet connections were recorded globally in April 2017, while 60 million in August 2016, 50 million in August 2015, and 40 million in September 2014 [10]. Bangladesh's banking industry is significantly prone to cyberattacks thanks to the nation's explosive growth in internet usage. Therefore, in such circumstances, robust built-in cyber protection is very crucial as well as experts are needed to safeguard knowledge and protect data from cyber-attacks.

First of all, there is a need for understanding the gravity of the cyber-related crisis and the challenges to safeguarding the whole cyber system. Usually, in terms of the nature of cybercrimes, there are four kinds of criminal activities that happened frequently in Bangladesh. The first of its kind can be termed cybercrimes against humanity. Such types of cybercrimes include hacking, unauthorized access, surveillance, data infiltration, email spoofing, spamming, fraud and forgery, slander, drug

trafficking, and virus transmission. The second type of crime is considered a property-related crime. It is a subset of cybercrime. Those include online time theft, intellectual property theft, and credit card fraud. The third type of crime is organized crime. Those are unauthorized access to and downloads from network resources and websites, the publication of pornographic and/or lewd content online, virus attacks, e-mail bombings, logic bombings, Trojan horses, data evasion, download blocking, theft of valuables, government terrorism against organizations, and vandalism of network infrastructure. Attacks on society or social values are another type of cybercrime in Bangladesh. Forgery, online gambling, prostitution, pornography (particularly child pornography), financial crimes, the degrading of young people through indecent exposure, web jacking, etc. are among the fourth type of cybercrimes [17]. However, pornography has now become a grave concern for Bangladesh. Regarding the social culture and moral principles of the nation, pornography is strictly forbidden here. One of the major reasons behind such crimes is the evolving growth of digital communication and information technologies. Since it is now feasible to quickly communicate with anyone, anywhere in the globe to anywhere else, paving the way for sharing and exchanging cultural values of a person or a nation more easily. This is why many damaging aspects of the culture or distorted contents of another country might readily contaminate their own culture as a result of the free exchange of ideas that might not be suitable for other cultures. Extremely repulsive aspects of perverse cultures, such as pornography, are not acceptable in Bangladeshi culture. Sexual harassment claims are frequently made to Bangladesh's law enforcement agencies. These sexual harassment tactics include sending obscene photographs and demanding significant sums of money in exchange for private nude film recordings or disseminating these images and videos for libel. This kind of crime is also sometimes done as a result of prior animosity, the unfulfilled desire for evil, or animosity toward a family member. These victims are mostly teenagers. Criminals do not, however, target ladies or children. Transborder crimes are those that are committed outside of the nation. When this occurs, it is a crime not just in the nation where it was committed, but also in the nation where it occurred. In the context of Bangladesh, it is not simple to legislate against cybercrimes like pornography because pornography is frequently not seen as illegal. It is not regarded as a crime in many nations, including the US. As a result, these transnational crimes must deal with a number of issues. Child pornography, however, is a global crime. Any nation would consider such a crime to be illegal. There is international support available to combat these crimes. The details of such a situation are provided below.

In June 2014, the Criminal Investigation Department (CID) police arrested master-minded of a gang TIM Fakruzzaman alias Tipu Kibria, chairman of Sudhui Bangla Photography Society, and two of his associates for child pornographic videos and photographs for foreign customers. The lawmen also rescued a 13-year-old boy from their capture. Kibria used to sell the videos and photos through the internet to his clients in Germany, the USA, Australia, the UK, and Saudi Arabia, CID officials said on June 12, 2014, at a press briefing in Dhaka. (Source: 'Child pornography gang busted in the capital' published on The Daily Star). Kibria used to bring male street youngsters to his house and lab where he would film pornographic material.

Before being caught, he attacked 400–500 children living on the streets [8]. With the assistance of his two assistants, Kibria used to perform all of these tasks. Later, the names of 13 foreign customers who frequently used online banking to pay Kibria to provide child pornography were discovered by the police. Meanwhile, law enforcement agencies in Bangladesh suspect that there may be other creators of this type of pornography in addition to Kibria. It follows that pornography poses a serious threat to Bangladesh's cyber security. Threats to online banking and cyber security are currently a major worry for all kinds of financial operations in Bangladesh. International criminal gangs conduct a variety of criminal operations here, including drug smuggling, trafficking, and terrorism, which pose a threat to the nation's cyber security due to the possibility of unlawful internet transactions. Due to inadequate cyber security measures in Bangladesh, the country's banking sector is one of the riskiest areas for cyberattacks. An incident of stealing \$101 million from Bangladesh Bank's account at the US Federal Reserve Bank is one of the most shocking events in the history of cyber heists across the world [11]. To commit the crime, hackers sent fake instructions from the Federal Reserve Bank of New York, using the Swift payment network, and eventually steal \$101 million from the account of Bangladesh's central bank. This is considered the largest cybercrime in history. This incident happened only for the fragility of cyber security and inadequate safety measures in the sector. There is a high possibility of the recurrence of such big incidents of cybercrimes in the financial sector of Bangladesh in the days to come if the country could not ensure cyber security. Due to the massive use of credit cards and electronic payment methods, there is a looming risk of sealing the personal information of users, bank account details, contact numbers, and e-mail IDs which may be appeared as a massive danger to the personal or organizational level. It is now very common for law enforcement agencies in Bangladesh to get complaints of victimizing by cybercrimes directly or indirectly due to digital financial transactions. Take the example of Eastern Bank and United Commercial Bank as evidence of stealing money from the account holders. 21 suspicious card transactions from Eastern Bank were discovered on February 12, 2016. The criminal used a fake card at the ATM booth at United Commercial Bank Limited to steal money. After an investigation by the Dhaka Metropolitan Police (DMP), the law enforcers On February 25 said they have found evidence against multiple hotels and travel agencies in connection with the ATM card scam. The DMP officials also found the involvement of some bank officials in the scam. A German national named Piotr was arrested on charges of ATM fraud. In the investigation, police found that a group of officials from City Bank, a private bank in Bangladesh, were involved in the scam. The lawmen also found that Piotr had the plan to conduct such financial scams in at least three other countries. Apart from the involvement of Piotr, the law enforcers also found that many other foreign nationals are also engaged in financial fraud in Bangladesh, including money laundering through ATMs. Bangladesh's insufficiently secure system for digital financial transactions is the reason they are focusing on it. Banks and customers are concerned as a result of the recent financial scandals. Lawyers for Bangladesh Bank and three other commercial banks discovered that at least Tk 2.5 million had been stolen in

February 2016 after reviewing the video of four ATM booths [11]. As per the information of a spokesperson of Bangladesh Bank, at least two foreign citizens were the mastermind of the ATM scams in the country. Similarly, United Commercial Bank Limited and Eastern Bank Limited fall prey to the ATM scam in the country. In the wake of the series of incidents of scams at the banks, almost all banks in Bangladesh, including both public and private commercial banks, have tightened their security measures. However, the matter is still concerning that whether those security measures taken by the banks are enough to face the challenges ahead. Where are the loopholes of technology to ensure security in the sector and why do such inadequate measures still exist? In addition, there is a huge lack of knowledge of security officials concerned. This is why the banking industry in Bangladesh is still having trouble ensuring cyber security. On top of that, many people fall prey to scams or phishing attempts in the country. The users are persuaded to provide private information like usernames, passwords, and credit card numbers via lucrative emails or alluring adverts in order to steal all of their money. The situation in Bangladesh is much critical as victims frequently lose \$100–500 through such digital crime but most of the victims are not willing to disclose the matter to the authorities concerned [14]. Cybersecurity risks in Bangladesh can include hacking or illegal access to a computer system without the owner's or user's authorization. Hackers always have a close watch on the financial operations of both governmental and non-governmental entities before conducting any financial crimes, and try to find out loopholes in the transaction system. The excessive reliance on foreign server system providers, insufficient cyber infrastructure networks, and lack of professional cyber security know-how have helped make Bangladesh a safe haven for cybercrimes, which is why it is very difficult to curb such crimes here. On top of the concerns over digital financial crimes in Bangladesh, the theft of data or information is another major concern here. There are multiple examples of leaking secretive information in the country which created public outcry or anarchy in the society. The leak of a verdict of the Bangladesh War Crime Trial Tribunal in 2014 was one such incident [8]. The tribunal's verdict on a war crime convict was released, leading to a significant backlash against the Bangladeshi government and exposing the country's lax cyber security. However, the most serious concern for the people in Bangladesh is the lack of cyber security on social media platforms. Thanks to inadequate security measures, unpleasant incidents are happening frequently on social media platforms, especially on Facebook, Twitter, and LinkedIn. The incidents of hacking social media accounts in order to extort money or attack any individual are also very common in the country. The hackers mainly targeted ladies and celebrities on social media and sought hefty sums of money in exchange. Without money, they threatened to damage the victims' reputations in society. Such incidents were so much repugnant that the law enforcement agencies in Bangladesh were compelled to form monitoring teams to tackle the arousing situations. Apart from that, Bangladesh Telecommunication Regulatory Commission formed a surveillance team in every district in the country to combat the crimes.

7 Recommendations to Raise Cybersucurity Awareness in Bangladesh

Ensuring cybersecurity should be the first and prime priority of every individual, entity, or country. There should have clear and bold strategies to make the cyber system secure and address the cyber challenges. As cyber-related issues are not only confined to a specific region and certainly a global issue, this is why interstate mutual cooperation and support are also necessary (public, corporate sectors, ISPs, etc.). So, it is not possible for a country to keep its cyber system safe with its own efforts alone. Meanwhile, increasing awareness about cybercrimes is the most important step which should be taken first of all other initiatives by the states or the authorities concerned. Public awareness of the issue will help lessen the severity of cyberattacks considerably. The matter of regret is that despite the way the use of the internet has increased, public awareness has not been raised to such an extent. In section addresses the steps that need to be made to increase awareness among the general public and government officials. If the guidelines or policies, which would be formulated or updated in order to contain the cybercrimes, are followed strictly, it would obviously help mitigate the cyber security risk of a country. The concerned stakeholders including the government should follow the following recommendations.

- To determine whether current legal frameworks are sufficiently applied against misuse in telecommunications and computer networks. This would eventually help investigate cybercrimes.
- Concerned counties should acknowledge issues relating to high-tech-related offenses in the right way based on mutual agreements.
- These countries must set up necessary protocols which will be tasked with collecting data from communication channels. Besides, they should develop appropriate strategies to expedite the process of international data transmission.
- The first task should be to enact legal frameworks and put them into effect in enhancing our cyber security. To this end, the governments should take appropriate legislative clearance. Under the framework, a dedicated institution-like National Cyber Security Council-should be established to implement the law.
- All concentration should be on safeguarding the whole national cyberspace and supporting the rights of internet users instead of ensuring the security of only some specific institutions.
- The endeavor to concentrate on defending cyberspace against potential assaults such as those involving cell phones, cloud computing, big data, etc.
- Review security plans periodically and stay updated with the latest technologies and stay alert against possible cyber threats.
- Consult with all relevant parties, including the government, military, telecommunications companies, financial institutions, the judiciary, civil society, religious organizations, and cyber security professionals, to get their opinions on the management or policies relating to cyber security.

- Create a comprehensive cyber management strategy with illustrations that accurately depict the interested parties, officials, transparency, objectives, investments, results, etc.
- Amend the national legal framework if necessary, to face cybercriminals and suspects as per their actions.
- Recommend launching necessary training and awareness campaigns on cyber security for users and concerned officials.
- Advise on increasing private–public partnerships to guarantee successful implementation of the nation’s cyberspace resilience.

8 Conclusion

Now cyber security has become the cause of global concern thanks to evolving communication and information technology. This study demonstrated that the level of cyber security awareness among Bangladeshi individuals is very insufficient to face the upcoming cyber security challenges and immediate steps are required to accelerate the level of awareness [12]. People would eventually find ways to safeguard themselves if they were aware of cyber security flaws beforehand. Bangladesh can reduce the danger of vulnerability by adopting the strategies used by the developed nations that have successfully been able to keep their cyber system safe and secure. Here are some particular suggestions which are related to the perspectives of Bangladesh. Having adequate knowledge and a better understanding the cyber-security issues and possible challenges are the main weapon to fight cybercrimes. Bangladesh’s government should regularly chalk out different programs on cyber security awareness and organize threat assessment seminars at government offices to raise the level of awareness among officials. The government also should take similar initiatives at the offices of private companies to mitigate security threats caused by internal workers. Another approach should be taken to persuade businesses to allocate a budget for data security that covers all of their cyber-related operations. The incorporation of cyber security awareness into curricula for school and college and tertiary education is also a useful tool to combat cybercrimes. It will increase student awareness of cyber security challenges as well as encourage them to pursue further study in the field and develop careers in the field that have increasing job prospects on the global job market. Another effective way of promoting cyber security awareness is to develop special programs or create special digital content like video games, particularly aimed at kids and college students.

References

1. Aziz A (2020) Challenges with digital inclusion in Bangladesh: the national ICT policy. *Int J Comput Sci Inf Technol Res* 9(1):88–94. www.researchpublish.com. Accessed 25 Apr 2023
2. Ahmed N, Islam MR, Kulsum U, Islam MR, Haque ME, Rahman MS (2019) Factors affecting Bangladesh's population's awareness of cyber security. In: 5th International conference on advances in electrical engineering
3. Alam S (2019) Cybercrime presents law enforcement with a new problem. *City Univ J* 2(1):75–84
4. Barua J (2019) The information technology and communications act should be amended. *The Daily Star*. <http://www.thedailystar.net>. Accessed 25 May 2019
5. Bleyder K (2012) Cybersecurity: the changing threat environment. BIPSS, Dhaka
6. BNWLA (2014) Women's psychological health survey. National Women Lawyers Association of Bangladesh, Dhaka
7. BTRC (2018) Subscribers to the internet. Telecommunications Regulatory Commission of Bangladesh. <http://www.btrc.gov.bd>. Accessed 10 Jan 2023
8. Editorial (2013) The 2013 draft ICT (Amendment) ordinance: a blackened law. *The Daily Star*. <http://archive.thedailystar.net/beta2/news>. Accessed 20 Feb 2022
9. Elahi SM (2022) Bangladeshi teenage porn addicts. Manusher Jonno Foundation, Dhaka
10. Greenemeier L (2022) China's cyber attacks indicate online as new front in conflict. <http://www.scientificamerican.com/article/chinas>. Accessed 15 Dec 2020
11. Hanif A (2020) Mass media and cyber laws of Bangladesh. National Law Book House, Dhaka
12. ICT Act (2006) The information & communication technology act. <http://www.prp.org.bd/downloads/ICTAct2006.pdf>. Accessed 1 Mar 2023
13. International Commission of Jurists (2021) Briefing Paper on the Bangladesh Information and Communication Technology Act Amendments. <http://icj.wpengine.netdna-cdn.com/2013.pdf>. Accessed 15 Apr 2023
14. Kabir AE (2022) Media and cyber laws in Bangladesh. Sufi Prokashoni, Dhaka
15. Karaman S (2017) In the face of online harassment, women support one another, but regulatory reform is required. Weblog for LSE Women, Peace, and Security. The London School of Economics and Political Science is located in London. <http://blogs.lse.ac.uk/wps/2017/11/>. Accessed 12 June 2022
16. Maruf AM, Islam MR, Ahamed B (2014) Bangladesh's emerging cyber threats: seeking effective legal redress. *NUB J Law* 1(1):112–124
17. Perlroth N, Gellesaug D (2014) Over a billion internet passwords are amassed by Russian Hackers. <http://www.nytimes.com/2014/stoleninternet>. Accessed 5 Mar 2023
18. Singer PW, Freidman A (2014) *Cyber security and cyber war: what everyone needs to know*. Oxford University Press, Oxford
19. Tikk E (2021) *Ten rules for cyber security-survival: global politics and strategy*. Routledge, London
20. USSD (2020) Human rights report for the country for 2016. United States Department of State. <https://www.state.gov/j/drl/rls/report/index.htm?ye>. Accessed 2 Apr 2023
21. Williams B (2014) Cyberspace: what is it, where is it and who cares? <http://www.armedforcesjournal.com/cyberspace>. Accessed 1 Mar 2022
22. Zaman S, Gansheimer L, Rolim SB, Mridha T (2017) Legal action against women cyber-violence. BLAST, Dhaka

Credit Card Fraud Detection Using Machine Learning



Berlin Srojila Manickam and Hamid Jahankhani

Abstract This research explores the application of Machine Learning (ML) algorithms in the detection of credit card fraud. Credit card fraudulence is a major concern that poses a significant threat to the financial industry, and machine learning has emerged as a promising approach to addressing this issue. The motive of this study is mainly to evaluate the efficacy of ML algorithms in the detection of credit card fraudulence and, to identify the key elements that impact the accuracy of these ML algorithms. The study utilizes a rigorous experimental design and a large sample size, which enables the analysis of the performance of diverse ML algorithms. However, the choice of ML algorithm and the performance evaluation metrics that is used are the important factors that influence the accuracy of the algorithms. This study also identifies the key factors that influence the accuracy of machine learning algorithms, including the choice of features, the quality of the data, and, also the parameter settings of the algorithms.

Keywords Machine learning · ML · Credit card · Fraud · Contactless · RoC curve · Artificial intelligence · XGBoost

1 Introduction

Over the span of years, there has been a significant rise in the use of credit cards globally. This trend is driven by several factors, including the rise of e-commerce, the increasing popularity of contactless payments, and the convenience and security that credit cards offer [24]. The tremendous growth of e-commerce is considered as one of the main drivers for the increase in usage of credit cards. In recent years, the online shopping has become increasingly popular, and credit cards are the most

B. S. Manickam · H. Jahankhani (✉)
Northumbria University, London, UK
e-mail: Hamid.jahankhani@northumbria.ac.uk

B. S. Manickam
e-mail: berlin.manickam@northumbria.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_15

275

commonly used payment method for online transactions. Credit cards offer a secure and convenient way to make purchases online, allowing consumers to shop from the comfort of their own homes. Another factor contributing to the rising use of credit cards is the increasing popularity of contactless payments.

Contactless payments feature allow consumers to make purchases so quickly and easily by simply tapping their mobile device or card on a payment terminal. This convenience has made credit cards a preferred payment method for many consumers, especially in countries where contactless payments are widely accepted. Credit cards also offer a range of benefits and rewards, such as cashback, points, and discounts, which incentivize consumers to use them for their purchases. Additionally, credit cards provide a layer of protection against fraud and unauthorized transactions, as consumers are not liable for fraudulent charges [27].

Unfortunately, as the usage of credit cards has increased, so has the incidence of credit card fraudulence activity. Credit card fraudulence occurs when the credit card information of someone else's is used by an individual without their knowledge to make unauthorized purchases or obtain funds. This can result in significant financial losses for the victims, as well as reputational damage to financial institutions. The existing fraud detection systems have limitations in detecting new and evolving fraud patterns. Traditional rule-based methods rely on predefined rules to detect fraudulent transactions, and they may miss new or sophisticated fraud schemes that do not fit into the predefined rules. Additionally, these systems can give rise to false positives in massive number, which might lead to unnecessary transaction declines and customer dissatisfaction [10].

In fraud detection, classification algorithms are commonly used to differentiate between legitimate and fraudulent transactions. However, detecting fraudulent transactions can be challenging as they are often rare, and financial institutions may be hesitant to share detailed transaction data due to privacy concerns. Consequently, when training a classification algorithm with limited information on fraudulent transactions, the overall performance of the model may suffer, resulting in many false positives [31]. A possible solution to address this issue is to apply sampling techniques to preprocess the data before feeding it into the machine learning algorithm. By doing so, the algorithm can have access to a balanced dataset, increasing its ability to identify fraudulent transactions accurately.

2 Artificial Intelligence

Sir Alan Turing asked, "Do computers think?" in his groundbreaking work *Computing Machinery and Intelligence*, and then proceeded to lay out a series of criteria, which he named *The Imitation Game*, to allow for the response to that question [7]. The Turing Test, which is based on this game, is the gold standard for determining whether or not a computer can convincingly simulate human intelligence.

Since then, the subject of Artificial Intelligence has expanded to include several specialised sub-disciplines that investigate narrower aspects of AI. Among them are the areas of Machine Learning (ML) and deep learning. According to a blog post by Nvidia employee Michael Copeland titled “What’s the Different Between Artificial Intelligence, Machine Learning, and Deep Learning?” [7] the three terms are associated in the following way.

AI, the original concept, would be at the centre of a set of concentric rings that also included machine learning, the next generation of this concept’s flowering, and deep learning, the engine of today’s AI boom [7].

Copeland gives a simple visual to help comprehend the interconnectedness of these three subfields of AI research, showing how Deep Learning is a relatively new subsection of the more general Machine Learning field, which is itself a subspecialty of AI study.

For the last several years, an unprecedented amount of progress has been made in the field of AI research, as reported by Copeland through Huang. The rise in popularity of AI and deep learning may be attributed in part to notable successes like Google’s DeepMind AlphaGo algorithm defeating South Korean Master Lee Se-dol in Go [7].

2.1 Machine Learning

Machine learning (ML) is a branch of Artificial Intelligence (AI) that mainly focuses on creating statistical models and, algorithms that allows the computers to learn from data and anticipate or take actions based on it. To make predictions or take action, entails the development of models and algorithms that can find patterns, correlations, and insights in huge, complicated datasets. Typically, supervised learning, unsupervised learning, and reinforcement learning make up the major three areas of machine learning. In supervised learning, a ML model is developed using a labelled dataset in which the correct answer for each input is predetermined. The model is then applied to forecast fresh, unforeseen data. Unsupervised learning involves training a ML model on an unlabeled dataset with the expectation that the ML model would recognise patterns or groups on its own. Reinforcement learning is a sort of ML in which the model learns by making mistakes, getting input from its surroundings, and modifying its behaviour as necessary [1].

2.2 Logistic Regression

Machine learning algorithms like logistic regression are frequently employed for classification issues. A set of input variables or features are utilised to estimate the likelihood of an outcome using this supervised learning process. A logistic function is fitted to the input data using the logistic regression procedure, which predicts the

likelihood of a binary result such as yes/no, true/false, or 0/1. The function sigmoid is an S-shaped curve that converts any input value to a value between 0 and 1, is the logistic function used in logistic regression. This means that, given a collection of input features, logistic regression can be used for estimating the likelihood of a binary outcome. The algorithm categorises the result as positive or negative, true or false, or 1 or 0, depending on whether the anticipated probability is greater than a predetermined threshold. There are several reasons why logistic regression is a popular approach. It is a straightforward algorithm that is simple to use and understand. It is a very effective algorithm that can deal with big datasets with lots of input features and it can manage missing data and resist outliers in logistic regression [17].

2.3 Decision Tree

Decision Tree (DT) is a widely used algorithm for classification and regression issues. This particular supervised machine learning technique constructs a model in the form of a tree structure that bases choices on the values of input features. This algorithm recursively divides the data into smaller subsets, depends on the values of the input features. At the root node, it starts with the complete dataset, and at each node, it selects the feature that offers the optimal split based on factors like information gain or Gini impurity. When the algorithm reaches the leaf nodes, which hold the ultimate judgements, it proceeds to divide the data into smaller groups. There are several benefits of using decision trees. They can handle both numerical and category data and are simple to interpret and comprehend. Moreover, they can deal with missing data, outliers, and overfitting [9].

2.4 Random Forest

A well-liked ML technique called Random Forest (RF) is used for classification, regression, as well as in other applications. This RF is an ensemble learning technique that incorporates various decision trees (DT) to construct a ML model that is highly dependable and accurate. During training, Random Forest builds massive number of decision trees, in which each DT is well trained using a random subset of the input data and with a random subset of the input characteristics. The programme then combines the result from each of these distinct trees' predictions. When compared to other ML methods, the RF machine learning model provides various benefits. The usage of bootstrapping and feature subsetting makes it less prone to overfitting than single decision trees. Moreover, it can handle datasets with missing values and high-dimensional data. It also can estimate the significance of each input characteristic, which is helpful for feature selection [32].

2.5 XGBoost

XGBoost stands for “Extreme Gradient Boosting”, which is a well-known open-source gradient boosting library that is used to solve supervised learning problems. It is based on decision tree ensemble methods and uses a gradient boosting framework to create a predictive model. XGBoost can handle a variety of data types, including numerical and categorical data, and can be used for both regression and classification tasks. It is known for its speed and accuracy in producing high-quality predictions, and these benefits makes it as a popular choice among data scientists and machine learning practitioners (Basabi et al. 2020).

3 Related Works

In the research work done by Sumant et al. (2022), they talked about the significance of a system for detecting the Credit Card fraud and suggested few Models based on ML and deep learning. They employed three crucial ML algorithms—Support Vector Machine (SVM), Deep Neural Network methods, and Naive Bayes—to create ML models using the data sets from the open source, named Kaggle. In their study, the prediction accuracy of the ML Model created using the Naive Bayes algorithm is 90%, that of the SVM is 97%, and that of the DNN is 99%. Unfortunately, they did not take precision, sensitivity, or time into account when analysing the results because they only used the parameter “accuracy” to evaluate the performance of the developed ML Model. Even though they mention specific pre-processing methods in their research paper, no information is provided about those methods in the paper, and there is little to no supporting data to be found. Therefore, the overall performance of the developed ML Model will then be evaluated in terms of its sensitivity and precision using the ROC curve in the proposed research work.

SVM, Decision Tree (DT), and Random Forest (RF) algorithms were used to construct a ML model by Saddam et al. (2021) in order to predict the fraudulent transaction that was associated with the Credit Card. They considered factors such as precision, sensitivity, and accuracy when assessing the performance of their Model. According to their findings, the Random Forest algorithm displays an accuracy of 85.5, while the SVM algorithm displays 84.1, and the Decision Tree algorithm displays 83.6. Their research work does not address the same issues as those that were discussed in the previous research work concerning their pre-processing techniques to avoid the noises from the data samples. They did not bring additional evidence in their research work to demonstrate that their results were more genuine.

In the research carried out by [15], the authors used the data of Credit Card fraud that was made available in the Kaggle data repository. The researchers found that the data set in the Kaggle repo contained a lower number of Credit Card fraudulent transactions in comparison to the benign transactions. They used this data set to train a variety of machine-learning algorithms, and then used sensitivity, precision, and

time to evaluate the performance of each newly created ML Model. They developed a ML Model by employing the following five Machine Learning algorithms: kNN, Naive Bayes (NB), Decision Tree (DT), Logistic Regression (LR), and Random Forest (RF). They concluded at the end of their work that the DT Model had a higher score of performance when compared to other Machine Learning Models. As time is a significant factor in preventing Credit Card fraud in a real-time environment, the greatest benefit of the DT Model is that it requires less time to make predictions on fraudulent transactions.

The greatest flaw of this research is that they directly process the data set without performing any activity to eliminate outliers and data errors. The algorithm for Machine Learning always produces better results based on the information that is fed to it during the training session. If it continues to make outliers and noises, this will eventually impact the prediction accuracy of the ML Model that was created. Because of this, the main purpose of the proposed research work is, to attempt to solve the problem. After performing several resampling techniques to remove outliers and noises from the input data, the Model will be created, the data will be fed into the ML Model, and the best ML Model will then be identified depending on its prediction accuracy.

4 Research Methods

The dataset for this research is obtained from Kaggle, an online open-source platform, which includes various credit card transactions information of European cardholders. The ULB (University Libre de Bruxelles) ML Group also uses the data sets for credit card-related studies. In this data set, there are 492 records for fraudulent transactions and 284,807 entries for genuine transactions. The data set is extremely unbalanced because there is a significant variance in data across the different classes [23].

Because of confidentiality issues, the data set that is collected does not hold all of the specifics of the credit card transaction. Principal component analysis (PCA), a method for transforming the data into another format with low information loss and great interpretability to increase confidentiality, converts 27 out of the 30 features in the data sets into another form. These features represent measurable data or a column for further analysis. The two parts, “Amount” and “time”, are left unaltered in the dataset because PCA does not transform them [30].

4.1 *Google Colaboratory*

Google Colab (short for Colaboratory) is a no-cost, cloud-based platform for AI research and data science on Google Cloud. It’s similar to Jupyter notebooks in that it lets users execute Python code in their browser on Google’s cloud. Users are able to create and run code, save and share their progress, and work together in real time

all inside the same environment thanks to Colab. As the Google colab gives free access to potent computing resources like GPUs and TPUs, it’s a perfect choice for individuals and companies who don’t have a lot of money to spend on processing power. It also interacts with a number of well-known ML libraries and frameworks including TensorFlow, Keras, PyTorch, and Scikit-learn, making it simple to create and train ML models [35].

4.2 Proposed Approach

The proposed solution presented in Fig. 1 consists of eight steps. The first step involves analyzing the data sets to identify their features and characteristics, which are then graphically represented. After analyzing the data, the next step is to transform it into a standard format to eliminate errors and ensure a consistent pattern across all data sets. The third stage is to prepare the data for use by machine learning algorithms by splitting it into training and testing sets. The best algorithm for spotting credit card fraud is then determined by comparing the results of many different algorithms.

The major highlights of the selected algorithms are explained in Table 1.

4.3 Data Analysis

Machine learning involves analyzing data to uncover patterns and insights for future decision-making. This process uses statistical and computational techniques to extract knowledge from datasets [6].

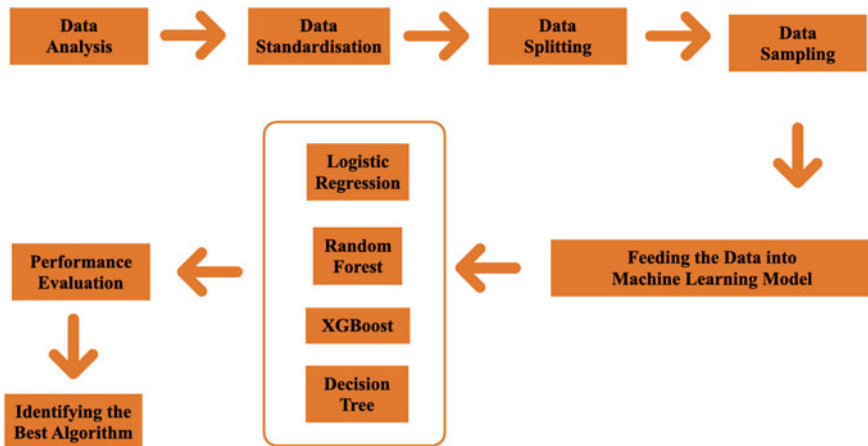


Fig. 1 The proposed solution to address the class imbalance problem

Table 1 Features of ML models (Feedzai, 2019)

Logistic regression	Random forest	XGBoost	Decision tree
<ul style="list-style-type: none"> ● Simple and Linear ● Reliable ● High Interpretability ● No parameters to tune 	<ul style="list-style-type: none"> ● Generalizes patterns well ● Robust to different input types (texts, number of scales, etc.) ● Robust to missing data ● Fast to train and score ● Trivially parallel ● Requires less tuning ● Probabilistic output (score) ● Can adjust threshold to trade-off between precision and recall ● Great predictive power ● Found to win many machine learning competitions 	<ul style="list-style-type: none"> ● Able to handle missing data and large datasets ● Highly Flexible and supports regularization ● In-built capability to handle missing values ● Cross Validation and effective tree pruning ● Parallel processing 	<ul style="list-style-type: none"> ● Requires less effort for data preparation during pre-processing ● Normalization and Scaling of data is not required ● Able to handle both numerical and categorical data ● Versatility and Non-linearity

In the context of ML, data analysis is a crucial step in the model-building process. Before developing a ML model, data analysts first explore and clean the data to ensure that it is accurate, complete, and relevant. They also use exploratory data analysis (EDA) techniques to understand the underlying relationships between the variables in the data. After the data has been thoroughly examined, data analysts use a variety of statistical and ML methods to construct predictive and prescriptive models (Zhang et al. 2021).

The code snippet (see Fig. 2) written in Python that imports various libraries commonly used in data analysis and visualization. Here’s what each line of code does:

`import pandas as pd`: This line brings in the Python package called pandas, which may be used to manipulate and analyse data. Because it is shorter and more convenient, the library is imported under the alias `pd` [5].

`import numpy as np`: A code line that brings in the Python- numpy package, which holds useful functions for manipulating arrays and doing mathematical calculations. The library is imported under the shorter and more memorable moniker `np` [12].

`import matplotlib.pyplot as plt`: Python’s matplotlib library’s pyplot module, which may be imported with this line, has the necessary tools for making static, interactive, and animated plots. The module is imported using the alias `plt` for brevity and convenience [21].

```
# Importing the libraries
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

import warnings
warnings.filterwarnings('ignore')
```

Fig. 2 Libraries used for data analysis

`%matplotlib inline`: This line is a magic command used in Jupyter notebooks to display matplotlib plots inline within the notebook. It allows for the automatic rendering of plots without the need to call the `show()` function [22].

`import seaborn as sns`: A line that brings in the library named seaborn, a Python toolkit, which is mainly used for building data visualisations. The library is imported using the alias `sns` for brevity and convenience [25].

`import warnings`: This code snippet includes an import statement for the warnings module, which is a feature in Python that allows for managing warnings [34].

`warnings.filterwarnings('ignore')`: This line of code changes the warning filter to suppress all warnings. It is commonly used to hide warnings that may interfere with the analysis or are not relevant. However, it is generally not advisable to ignore warnings since they can indicate significant problems with the code or data [21].

4.4 Data Standardisation

In machine learning, data standardisation is the process of converting raw numerical data into normal distribution with zero mean and, one standard deviation. In order to facilitate meaningful comparison and analysis, it is necessary to normalise characteristics that have previously been measured using distinct scales and units [4].

Many machine learning techniques, like neural networks and linear regression, are highly sensitive to the magnitude of the input characteristics, making standardisation essential. If the features have vastly different scales, some features may dominate the learning process, while others may have little to no effect. Standardization helps to mitigate this problem by ensuring that each feature contributes equally to the learning process [14].

The first step in the process of standardising a feature is to determine its overall mean and standard deviation. Then, we take each feature value and remove the mean before dividing by standard deviation. This produces a new feature distribution in which both the mean and standard deviation are set to zero [13].

4.5 Handling Missing Value

In order to prepare a dataset for machine learning, it is necessary to deal with any missing or incomplete data by detecting and filling in the gaps. There are several potential causes of missing data, such as mistakes during data input, unfinished surveys, or broken measuring tools.

Missing data may be dealt with in a number of ways. One method involves erasing the section containing the missing data. However, this method may limit the sample size of the dataset and cause the loss of important data.

One alternative is to make reasonable guesses based on the existing data to “impute” the missing numbers. The mean imputation, the mode imputation, and the regression imputation are only a few of the approaches used to fill in missing data. For missing data, mean imputation will substitute the average of all available data for that characteristic in place of the missing ones. With mode imputation, the most frequent value for a feature is substituted for a missing one. Using a Regression model and the other properties of the dataset, regression imputation attempts to infer the missing values.

The code snippet (see Fig. 3) provided calculates the missing values percentage in each column of a Pandas DataFrame `df`, sorts the values in descending order, and stores the result in a new DataFrame `df_missing_columns`. Here is a breakdown of the code:

`df.isnull()`: this code line returns a DataFrame with the similar shape as `df`, in which each element is displayed either as True or False; True indicates there is a missing value, whereas False indicates no missing values.

`df.isnull().sum()`: this code line returns a Series that contains the missing values count in each column of `df`.

`len(df.index)`: returns the number of rows in `df`.

`df.isnull().sum()/len(df.index)`: calculates the fraction of missing values in each column of `df`.

`round((df.isnull().sum()/len(df.index))*100,2)`: multiplies the fraction by 100 to convert it to a percentage and rounds the result to two decimal places.

`.to_frame('null')`: converts the resulting Series to a DataFrame with the column name 'null'.

`.sort_values('null', ascending = False)`: sorts the rows of the DataFrame in descending order of the percentage of missing values.

Overall, the resulting DataFrame `df_missing_columns` displays the missing values percentage for each column of the original DataFrame `df`, sorted in descending order.

```
#Handling the missing values in columns
# Checking percent of missing values in columns
df_missing_columns = (round(((df.isnull()).sum()/len(df.index))*100),2).to_frame('null').sort_values('null', ascending=False)
df_missing_columns
```

Fig. 3 Code snippet to handle the missing value

This information can be useful for identifying which columns have the most missing values and deciding how to handle those missing values in data pre-processing.

4.6 Splitting the Data

In machine learning, “splitting the data” is partitioning a dataset into two or more subsets such that each may be utilised independently. In this process, to train and test a ML model, the most typical data split is the training-validation-test split [36]. In a typical training-validation-test split, the dataset is divided into three subsets:

Training set: This subset is used mainly to train the ML model. It typically comprises the majority of the data, usually around 70–80% of the original dataset.

Validation set: This subset is used to evaluate the ML model performance during training and also to tune the hyperparameters. It is typically around 10–15% of the original dataset.

Test set: This subset is used to evaluate the ultimate performance of the ML model. It is a completely independent subset that was not used during training or validation. It is typically around 10–15% of the original dataset [29].

Splitting the data serves two purposes: avoiding overfitting and testing the ML model’s efficacy on unseen data. An estimate of the model’s performance on fresh data may be obtained by training it on a specific portion of the data and assessing it on another subset [3].

Depending on the data at hand and the nature of the issue at hand, many techniques for data partitioning are available, such as random sampling, stratified sampling, and partitioning based on time. If one want the model to be reliable and applicable to new data sets, one needs to carefully choose a data-splitting strategy [36].

The code shown in Fig. 4 provided how to split a dataset into two sets: the training and test sets, using the function `train_test_split` from the library `scikit-learn` in Python. The breakdown of the code is as follows:

`from sklearn.model_selection import train_test_split`: this line imports the function `train_test_split` from the library, named `scikit-learn`.

```
# Import library
from sklearn.model_selection import train_test_split

# Putting feature variables into X
X = df.drop(['Class'], axis=1)

# Putting target variable to y
y = df['Class']

# Splitting data into train and test set 80:20
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.8, test_size=0.2, random_state=100)
```

Fig. 4 Importing necessary library and performing splitting operation

`X = df.drop(['Class'], axis = 1)`: this code line creates a DataFrame X that contains all the columns of the original DataFrame df except for the column named 'Class'. This assumes that 'Class' is the target variable or dependent variable in the dataset.

`y = df['Class']`: creates a Series y that contains the values of the column named 'Class'. This assumes that 'Class' is the target variable or dependent variable in the dataset.

`X_train, X_test, y_train, y_test = train_test_split(X, y, train_size = 0.8, test_size = 0.2, random_state = 100)`: splits the dataset into training and test sets. This specific function takes the following arguments:

`X` and `y`: The features and target variables of the dataset, respectively.

`train_size`: represents the proportion of the dataset to include in the training set.

`test_size`: indicates the proportion of the dataset to include in the test set.

`random_state`: The random seed to ensure reproducibility of the split.

The function returns four subsets: `X_train` and `y_train` holds the training data, while `X_test` and `y_test` holds the test data.

The ML model may be trained on the training data and its effectiveness can be assessed on the test data after the dataset has been partitioned. This enables us to predict the ML model's performance on data it has never seen before. For reproducibility's sake, the argument `random_state` is utilised to generate the same random split each time whenever the code is executed [37–41].

4.7 Sampling Data

Sampling, in the context of ML, is the practise of drawing out a subset of a larger dataset for the purposes of further analysis or model training. The main target of sampling is to lessen the volume of data that are needed for analysis or training without compromising the accuracy or representativeness of the resulting subset of data [26].

4.7.1 Undersampling

In machine learning, undersampling is used to deal with unbalanced datasets in which there are disproportionately more samples from one class, named the majority class than the other, the minority class. The goal of undersampling is to get a more even distribution of classes within a dataset by randomly selecting and removing members of the dominant class (Brownlee et al. 2021).

For example, suppose there is a dataset with 1000 samples, of which 900 belong to class A and 100 belong to class B. This is an imbalanced dataset because class A is much larger than class B. To balance the class distribution, one could randomly select, say, 400 samples from class A and remove them from the dataset. This would result in a new dataset with 600 samples, of which 300 belong to class A and 300 belong to class B.

Undersampling can be useful when we have limited resources or time for model training and the majority class samples dominate the dataset, causing the ML model to be biased towards the majority class (Brownlee et al. 2021). Hence this dissertation makes use of under-sampling, a very simple technique.

4.7.2 Libraries Used for Undersampling

These two lines of code (see Fig. 5) import the necessary modules for performing random undersampling on an imbalanced dataset. The `RandomUnderSampler` class is part of the `imblearn` library, which provides a collection of methods for handling imbalanced datasets in machine learning. The `Counter` class is part of the `collections` module, which provides a container for counting the frequency of elements in a list or array.

`RandomUnderSampler` is a class that implements random undersampling by randomly selecting a subset of samples from the majority class in order to balance the class distribution. It takes several parameters, including `sampling_strategy` (this represents the desired ratio of minority class samples to majority class samples after undersampling), `random_state` (implies the seed for the random number generator), and `replacement` (whether to sample with or without replacement).

The `Counter` class is used to count the number of samples in each class, before and after undersampling. This can be useful for evaluating the effectiveness of the undersampling method and ensuring that the resulting dataset is balanced.

4.7.3 Instantiating the Random Undersampling

The code (see Fig. 6) performs random undersampling on the training set (`X_train` and `y_train`) using the `RandomUnderSampler` class from the `imblearn` library.

```
# Importing undersampler library
from imblearn.under_sampling import RandomUnderSampler
from collections import Counter
```

Fig. 5 Libraries used for undersampling

```
# instantiating the random undersampler
rus = RandomUnderSampler()
# resampling X, y
X_train_rus, y_train_rus = rus.fit_resample(X_train, y_train)
```

Fig. 6 Instantiating the random undersampling

The method `fit_resample` is called on the `RandomUnderSampler` object (`rus`) to fit the undersampling to the training data and apply the undersampling method to obtain a new resampled dataset, `X_train_rus` and `y_train_rus`.

The default behaviour of `RandomUnderSampler` is to achieve a balanced class distribution by randomly selecting the samples from the majority class. If the `sampling_strategy` parameter is not specified, the minority class will be sampled to have the same number of samples as the majority class.

5 Model Creation

There are two types of parameters that affect the Machine Learning Model. Both a model parameter and a model hyper parameter are examples of this. Parameters are variables that are estimated by the model during training, while hyper parameters are those variables that the model cannot predict. In ML, the model parameter may be estimated with the help of a model hyper parameter. For example, if a user alters the ML model's learning rate, the model parameter will likewise vary [37].

A method called `K-fold cross-validation` is for assessing the efficacy of a ML model on a sample of unknown data. A commonly used technique for hyperparameter analysis is `K-fold cross-validation`. It involves dividing the dataset into `K` subsets, which are randomly shuffled before the split. One subset is used for testing while the others are used for training, and this process is repeated with different subsets used for testing to obtain a range of hyperparameter values [37–41].

5.1 Hyper Parameter Tuning in Logistic Regression

This code snippet (see Fig. 7) demonstrates hyperparameter tuning using `GridSearchCV` for Logistic Regression (LR).

The first step is to initialize a `KFold` cross-validation object, `folds`, which is used to create 5 folds with shuffling and a random state of 4. This means that the data splits into 5 folds and shuffled before each split, and the random state is set to 4 for reproducibility purposes. Next, a dictionary of hyperparameters to be tuned is specified using `params`. In this example, we're tuning the regularization strength parameter `C` of a logistic regression model.

The next step is to create a `GridSearchCV` object, `model_cv`. the estimator is set to a LR model, and `param_grid` is set to the dictionary of hyperparameters specified earlier. `scoring` is set to `'roc_auc'`, a metric that is commonly used for binary classification problems. `cv` is set to `folds`, the cross-validation object we created earlier. `verbose` is set to 1 to print the progress of the grid search, and `return_train_score` is set to `True` to return the training scores as well.

Finally, the `model_cv` object fits to resampled training data (`X_train_rus` and `y_train_rus`) using the `fit()` method. This will run a grid search over all the possible

```

# Creating KFold object with 5 splits
folds = KFold(n_splits=5, shuffle=True, random_state=4)

# Specify params
params = {"C": [0.01, 0.1, 1, 10, 100, 1000]}

# Specifying score as roc_auc
model_cv = GridSearchCV(estimator = LogisticRegression(),
                        param_grid = params,
                        scoring= 'roc_auc', |
                        cv = folds,
                        verbose = 1,
                        return_train_score=True])

# Fit the model
model_cv.fit(X_train_rus, y_train_rus)

```

Fig. 7 Hyper parameter tuning in logistic regression

combinations of hyperparameters in `params`, and return the best hyperparameters that maximize the `'roc_auc'` score. Once the best hyperparameters are found, the LR model can be retrained on the full training data and then evaluated on the test set to assess its performance (scikit-learn, 2022).

5.2 Hyper Parameter Tuning in Decision Tree Classifier

This code (see Fig. 8) performs hyperparameter tuning using grid search cross-validation for a decision tree classifier. The hyperparameters being tuned are:

The parameter `max_depth`: represents the decision tree model's maximum depth.

The hyperparameter `min_samples_leaf`: indicates the number of samples that are required as minimum to be at a leaf node.

The parameter `min_samples_split`: specifies the number of samples that are required as minimum to split an internal node.

The possible values for each hyperparameter are defined in the `param_grid` dictionary. For example, `max_depth` can take on the values 5, 10, or 15. Then a `DecisionTreeClassifier` object is created without any specified hyperparameters, which means the default hyperparameter values will be used.

Then, a `GridSearchCV` object is created with the decision tree object, the `param_grid` dictionary, a scoring parameter of `roc_auc` (which specifies that part under the receiver operating characteristic curve is used for performance evaluation of the ML model), threefold cross-validation, and a `verbose` parameter set to 1 to display progress information.

Finally, the `fit` method of the `GridSearchCV` object, the `grid_serach.fit` is called with the training data: `X_train_rus` and `y_train_rus`, to train this DT model using

```

# Create the parameter grid
param_grid = {
    'max_depth': range(5, 15, 5),
    'min_samples_leaf': range(50, 150, 50),
    'min_samples_split': range(50, 150, 50),
}

# Instantiate the grid search model
dtree = DecisionTreeClassifier()

grid_search = GridSearchCV(estimator = dtree,
                           param_grid = param_grid,
                           scoring= 'roc_auc',
                           cv = 3,
                           verbose = 1)

# Fit the grid search to the data
grid_search.fit(X_train_rus,y_train_rus)

```

Fig. 8 Decision tree classifier-hyper parameter tuning

all combinations of hyperparameters specified in `param_grid`. The best hyperparameters found during the search are then saved in the `best_params_` attribute of the `GridSearchCV` object, and likewise the best model is stored in the `best_estimator_` attribute [37–41].

5.3 Hyperparameter Tuning in XGBoost

This code block as shown in Fig. 9 is performing hyperparameter tuning using `GridSearchCV` on an `XGBoost` classifier.

First, a `KFold` object with 3 folds are created.

Then, a dictionary `param_grid` is defined with a range of hyperparameters to be tested for the `XGBoost` model. The hyperparameters are `learning_rate` and `subsample`.

An `XGBoost` model is initialized with `max_depth` of 2 and `n_estimators` of 200.

A `GridSearchCV` object is initialized with the above-defined parameters including the estimator (`XGBoost` model), the parameter grid to be searched, scoring metric (ROC AUC), number of folds for cross-validation (3), verbose level and flag to return the train scores.

Finally, `model_cv.fit()` method is called to train the model with the hyperparameters defined in the `param_grid`.

The output of this code block will be the best hyperparameters for the `XGBoost` model based on the scoring metric (ROC AUC) and the specified range of hyperparameters [37–41].

```

# creating a KFold object
folds = 3

# specify range of hyperparameters
param_grid = {'learning_rate': [0.2, 0.6],
              'subsample': [0.3, 0.6, 0.9]}

# specify model
xgb_model = XGBClassifier(max_depth=2, n_estimators=200)

# set up GridSearchCV()
model_cv = GridSearchCV(estimator = xgb_model,
                        param_grid = param_grid,
                        scoring= 'roc_auc',
                        cv = folds,
                        verbose = 1,
                        return_train_score=True)

# fit the model
model_cv.fit(X_train_rus, y_train_rus)

```

Fig. 9 Hyper parameter Tuning in XGBoost

5.4 Hyper Parameter Tuning in Random Forest

In this code snippet (see Fig. 10), a random forest classifier is trained using GridSearchCV for hyperparameter tuning. The `param_grid` dictionary defines a grid of hyperparameters to be searched using cross-validation. The hyperparameters includes: 1. `max_depth`, 2. `min_samples_leaf`, 3. `min_samples_split`, 4. `n_estimators`, and 5. `max_features`.

This `RandomForestClassifier` object is instantiated as `rf`. The `GridSearchCV` object is instantiated with the following parameters: `estimator` set to the `rf` object, `param_grid` set to the `param_grid` dictionary, `scoring` set to `'roc_auc'` (the scoring metric for evaluation), `cv` is assigned with the value 2 (which indicates the number of folds in the process of cross-validation), `n_jobs` is assigned with the value -1 (which represents the number of CPU cores used for computation), `verbose` set to 1 (this indicates the level of messages in the search process), and `return_train_score` set to `True` (to also report the training score during cross-validation).

Finally, the RF model is fit to the training data: `X_train_rus` and `y_train_rus` using the function `fit()`. The best hyperparameters are selected by cross-validation, and the model is trained with these efficient hyperparameters [37–41].

```
param_grid = {
    'max_depth': range(5,10,5),
    'min_samples_leaf': range(50, 150, 50),
    'min_samples_split': range(50, 150, 50),
    'n_estimators': [100,200,300],
    'max_features': [10, 20]
}
# Create a based model
rf = RandomForestClassifier()
# Instantiate the grid search model
grid_search = GridSearchCV(estimator = rf,
                           param_grid = param_grid,
                           scoring= 'roc_auc',
                           cv = 2,
                           n_jobs = -1,
                           verbose = 1,
                           return_train_score=True)

# Fit the model
grid_search.fit(X_train_rus, y_train_rus)
```

Fig. 10 Hyper parameter tuning in random forest (Source Author Lab)

6 Model Evaluation

In machine learning, a model's efficacy is measured by how well it performs on a certain dataset, and this is called "model evaluation". This evaluation process is performed to test the ML model's ability to extrapolate to previously unknown information. Several measures are used to evaluate the model's efficacy. Metrics like the ROC-AUC curve, F1 score, and recall rate are often employed in machine learning evaluations. The challenge at hand and the intended solution should guide the selection of the measure.

Either the test dataset that was put aside during the data splitting process, or cross-validation methods like K-fold cross-validation, may be used for the evaluation. In order to acquire a full picture of the model's efficacy, it must be evaluated across a variety of measures. The assessment findings may either be utilised to fine-tune the model for optimal performance or to compare and choose the most effective model for deployment [19].

6.1 Confusion Matrix

The Classification model’s performance may be summarised in the tabular format which is named as confusion matrix (see Fig. 11), which does this by comparing the actual labels of a batch of data to the predicted labels provided by the model.

This confusion matrix consists of four quadrants, which are: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) [8]. Each quadrant represents a different classification outcome as in Table 2.

Accuracy, recall, precision and F1-score are only few of the performance measures of a classification model that might be computed using the confusion matrix. A model’s effectiveness may be gauged and improvement points pinpointed using this method.

Confusion_matrix is a function from the sklearn.metrics module in Python, which is used for the performance evaluation of a classification model by making comparison on the predicted output of the model with true output.

In the given code (see Fig. 12), y_test and y_test_pred are the true and predicted outputs of the classification model for a test set, respectively. The confusion_matrix function takes these two arrays as inputs and returns a 2 × 2 matrix confusion that represents the count of: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) of the ML model.

The values of TP, TN, FP, and FN are then extracted from the confusion matrix (see Fig. 13) and assigned to the variables TP, TN, FP, and FN, respectively.

Fig. 11 Confusion matrix



Table 2 Confusion matrix—quadrants outcome

True positives (TP)	The number of instances that were predicted as positive (in a binary classification problem) and are actually positive
False positives (FP)	The number of instances that were predicted as positive but are actually negative
True negatives (TN)	The number of instances that were predicted as negative and are actually negative
False negatives (FN)	The number of instances that were predicted as negative but are actually positive

```
# Confusion matrix
confusion = metrics.confusion_matrix(y_test, y_test_pred)
print(confusion)
```

Fig. 12 Defining confusion matrix values

```
TP = confusion[1,1] # true positive
TN = confusion[0,0] # true negatives
FP = confusion[0,1] # false positives
FN = confusion[1,0] # false negatives
```

Fig. 13 Creating the confusion matrix

These extracted values are then used to calculate various evaluation metrics such as accuracy, recall, precision and F1 score.

7 Analysis

7.1 Data Analysis

Data analysis is an essential component in credit card fraud detection using ML algorithms. It involves examining and processing massive amount of data in order to identify patterns and anomalies that might indicate fraudulent activities [2]. At the initial stage, the particulars of the classes in the dataset are analysed as shown in Fig. 14, where “0” is represented by the genuine transaction and “1” is represented by the fraudulent transaction. From the figure it is identified that there 284,315 transactions are genuine and 492 transactions are fraud. When comparing the two transactions the fraudulent transactions are very less. For obtaining more clarity on the data set’s distribution, the percentage of authorized transactions and fraudulent transactions are calculated as shown in Figs. 15 and 16. The obtained values from Figs. 15 and 16 is graphically represented in Figs. 17 and 18.

```
classes = df['Class'].value_counts()
classes
0    284315
1      492
Name: Class, dtype: int64
```

Fig. 14 Checking the class distribution

```
normal_share = round((classes[0]/df['Class'].count()*100),2)
normal_share
99.83
```

Fig. 15 Verifying the genuine transaction

```
fraud_share = round((classes[1]/df['Class'].count()*100),2)
fraud_share
0.17
```

Fig. 16 Verifying the fraudulent transaction

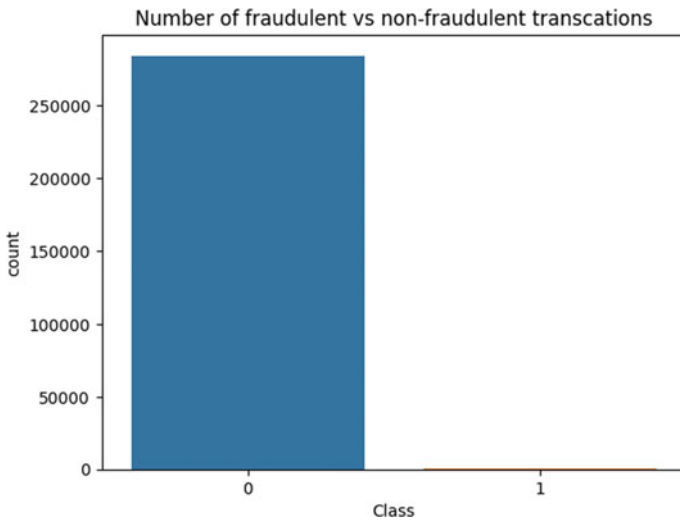


Fig. 17 Number of fraudulent versus non-fraudulent transactions

7.2 Model Evaluation

7.2.1 Analysing the Performance Metrics

After creating the Machine Learning model, the performance metrics of each model is evaluated using certain parameters like accuracy, specificity, sensitivity and f-1 score as shown in Figs. 19, 20, 21 and 22. The details of each metric are mentioned in Table 3.

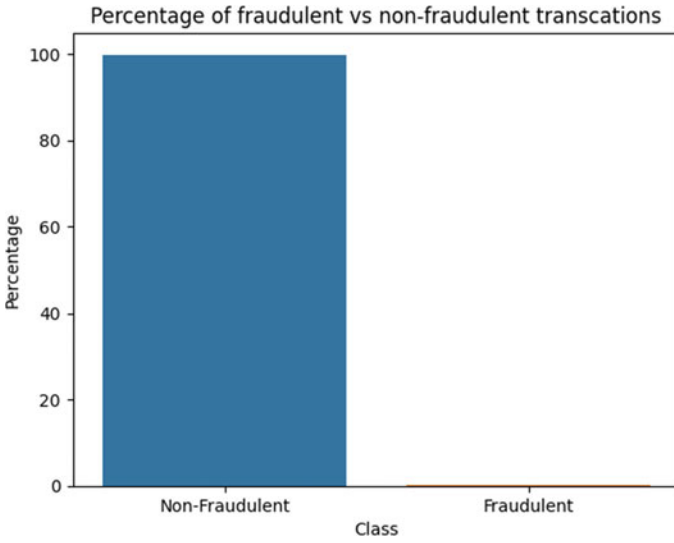


Fig. 18 Percentage of fraudulent versus non-fraudulent transactions

```
Accuracy:- 0.9545454545454546  
Sensitivity:- 0.9217171717171717  
Specificity:- 0.9873737373737373  
F1-Score:- 0.9530026109660575
```

Fig. 19 Performance metrics of logistic regression

```
Accuracy:- 0.5  
Sensitivity:- 1.0  
Specificity:- 1.0
```

Fig. 20 Performance metrics of XGBoost

```
Accuracy:- 0.9267676767676768  
Sensitivity:- 0.8686868686868687  
Specificity:- 0.9848484848484849
```

Fig. 21 Performance metrics of decision tree

```

Accuracy:- 0.9381313131313131
Sensitivity:- 0.8939393939393939
Specificity:- 0.98232323232324
F1-Score:- 0.9352708058124174
    
```

Fig. 22 Performance metrics of random forest

Table 3 Performance metrics comparison chart

	Metric	Logistic regression (Fig. 19)	XGBoost (Fig. 20)	Decision tree (Fig. 21)	Random forest (Fig. 22)
1	Accuracy	0.954	0.5	0.926	0.938
2	Sensitivity	0.921	1	0.868	0.893
3	Specificity	0.987	1	0.984	0.982
4	F-1 score	0.953	–	–	0.935

Logistic Regression

A Logistic Regression (LR) model accuracy of 0.95 in credit card fraud detection means that the model correctly identifies 95% of fraudulent transactions, as well as correctly identifies 95% of non-fraudulent transactions. In other words, the model is 95% accurate in predicting whether a given transaction is fraudulent or not. This level of accuracy is typically considered high for a credit card fraud detection model, as it means that the model is able to identify the vast majority of fraudulent transactions and limit the number of false positives (legitimate transactions that are mistakenly identified as fraudulent).

Sensitivity refers to the proportion of actual fraudulent transactions that are correctly identified as fraudulent by the logistic regression model [2]. A sensitivity of 0.921 means that the model correctly identifies 92.1% of all fraudulent transactions in the data set.

In the context of a LR model used for credit card fraud detection, a specificity of 0.987 means that the model correctly identifies 98.7% of non-fraudulent transactions as non-fraudulent. Specificity is one of the performance metrics used to evaluate the effectiveness of a binary classification model like logistic regression. It measures the proportion of true negative predictions (i.e., the number of non-fraudulent transactions correctly identified as non-fraudulent) out of all actual negative cases (i.e., the total number of non-fraudulent transactions) [20]. In this case, a specificity of 0.987 indicates that the logistic regression model is effective in identifying non-fraudulent transactions, and only misclassifies 1.3% of non-fraudulent transactions as fraudulent.

In the context of a LR model used for credit card fraud detection, an F1 score of 0.953 means that the model has a high level of accuracy in predicting both fraudulent and non-fraudulent transactions.

A performance metric, F1 Score is used to evaluate the effectiveness of a binary classification model like logistic regression. It is a combination of precision and recall, which are two other performance metrics that measure the LR model's ability to correctly identify true positive cases (i.e., the number of fraudulent transactions correctly identified as fraudulent) and avoid false positives (i.e., the number of non-fraudulent transactions incorrectly identified as fraudulent) [16].

The F1 score of 0.953 shows that the logistic regression model performs well in accurately identifying fraudulent transactions while minimizing false positives. Striking the balance between detecting fraud and avoiding false positives is crucial for credit card fraud detection to prevent customer dissatisfaction and revenue loss. With a high F1 score of 0.953 the logistic regression model proves, to be an efficient tool for detecting both fraudulent and non-fraudulent transactions, making it a valuable asset in credit card fraud prevention.

XGBoost

The accuracy score of an XGBoost model at 0.5 in credit card fraud detection implies that this ML model's performance is not better than random guessing in distinguishing fraudulent transactions. In other words, the model is making predictions at chance level, and its results are not useful for identifying credit card fraud. To assess this model's effectiveness, other evaluation metrics such as precision, recall, and F1 score should also be taken into account.

A sensitivity score of 1 for an XGBoost model indicates that the model can correctly identify all fraudulent transactions, which is an ideal scenario for credit card fraud detection. It enables timely interventions to prevent potential financial losses.

A specificity score of 1 for an XGBoost model implies that the model can correctly identify all non-fraudulent transactions as such, which is also an ideal scenario for credit card fraud detection. It helps prevent customer dissatisfaction and lost revenue that may occur from false positives.

An XGBoost model with both sensitivity and specificity equal to 1 can correctly identify all fraudulent and non-fraudulent transactions, respectively, with high precision. Such a model is highly effective in detecting credit card fraud and can be an invaluable tool for financial institutions and credit card companies. However, since the model's accuracy is only 0.5, it may misclassify a huge number of transactions, which is to be addressed.

Decision Tree

The accuracy of DT model is reflective of the percentage of correctly classified transactions in the dataset, with a score of 0.926 indicating a high level of accuracy in identifying fraudulent and non-fraudulent transactions. A sensitivity score of 0.868, suggests that the model can identify a significant portion of fraudulent transactions,

while a specificity score of 0.984 pointing that model is effective in distinguishing non-fraudulent transactions. These performance metrics suggest that the decision tree model can be a valuable tool for credit card fraud detection, balancing the identification of fraudulent transactions with avoiding false alarms for non fraudulent transactions.

Random Forest

When we discuss about credit card fraud, the accuracy of a Random Forest (RF) model refers to the proportion of correctly classified transactions, both fraudulent and non fraudulent, out of the total number of transactions. An accuracy score of 0.938, means that the random forest model correctly classified 93.8% of the transactions in the dataset. This is a relatively high accuracy score and recommends that the RF model is effective in distinguishing between fraudulent and non-fraudulent transactions.

A sensitivity score of 0.893, represents that the RF model is correctly identified 89.3% of the actual fraudulent transactions in the dataset. This indicates that the random forest model has a high ability to detect fraudulent transactions and is effective in identifying potential fraud.

A specificity score of 0.982; indicates that the RF model is correctly identified 98.2% of the actual non-fraudulent transactions in the dataset. This pointing that this ML algorithm has a high ability to accurately identify non-fraudulent transactions and is effective in avoiding false alarms.

Finally, a F1 score of 0.935, implies that the RF model has a high ability to correctly classify both fraudulent and non-fraudulent transactions, with a good balance between precision and recall. This suggests that RF model is effective in distinguishing between fraudulent and non-fraudulent transactions, and has low false positive and false negative rates.

7.2.2 Analysing the RoC Curve

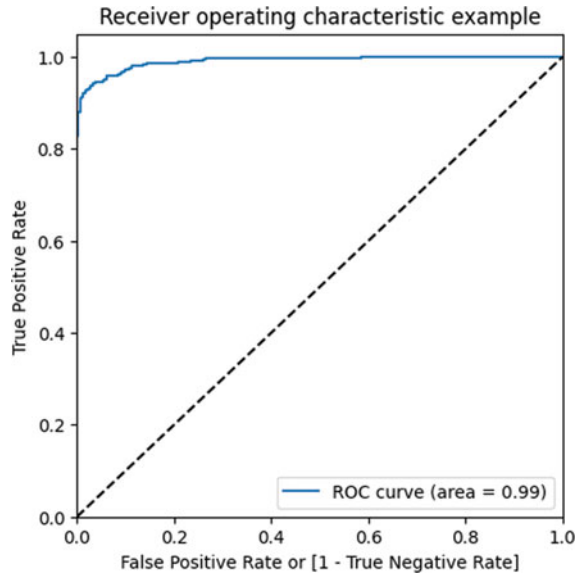
RoC Curve

If a user want to evaluate a predictive ML model's ability to discriminate between positive classes and negative classes, one may do so by plotting the model's ROC (Receiver Operating Characteristic) curve [28].

The ROC curve compares the TP rate (it is also known as sensitivity) to the FP rate, which is also known as specificity, for a variety of classification thresholds. The TP rate measures the proportion of actual positives that are correctly classified as positive by the model, while FP rate measures the proportion of actual negatives that are correctly classified as negative by the model [33].

A perfect model would have a ROC curve that passes through the top-left corner, representing a true positive rate of 1 and a false positive rate of 0. In practice, most models have a curve that falls somewhere below this ideal point [18].

Fig. 23 RoC curve for logistic regression



The portion under the ROC curve (AUC-ROC) is a widely used performance metric to quantify the overall performance of a binary classification model. The AUC-ROC ranges from 0 to 1, with higher values indicating better model performance. An AUC-ROC of 0.5 implies a model that performs no better than random guessing, while an AUC-ROC of 1 indicates a perfect model that can perfectly distinguish between positive and negative classes [11].

An AUC-ROC of 0.99 (see Fig. 23) indicates that the LR model has an excellent ability to differentiate between positive and negative classes, with a very high True Positive (TP) rate and a very low False Positive (FP) rate. In other words, the model correctly identifies the positive cases with a high probability and minimizes the risk of misclassifying negative cases as positive. A high AUC-ROC score of 0.99 indicates that the binary classification ML model is highly effective in differentiating between positive classes and negative classes, and it is a strong indicator that the model is performing well in its classification task.

An AUC-ROC score of 0.50 (as shown in Fig. 24) suggests that the binary classification model is performing poorly and is not able to accurately distinguish between positive and negative classes. The model's predictions are unreliable and do not provide any useful insights. A low AUC-ROC score of 0.50 represents that the model is ineffective in differentiating between positive and negative classes, and it is a clear sign that the model is not performing well in its classification task. In such situations, it is crucial to reassess the model and explore alternative algorithms.

An AUC-ROC score of 0.98 suggest that the given algorithm is good at distinguishing between positive and negative classes (see Fig. 25), with a high TP - true positive rate and a low FP—false positive rate. In other words, the model correctly

Fig. 24 RoC curve for XGBoost

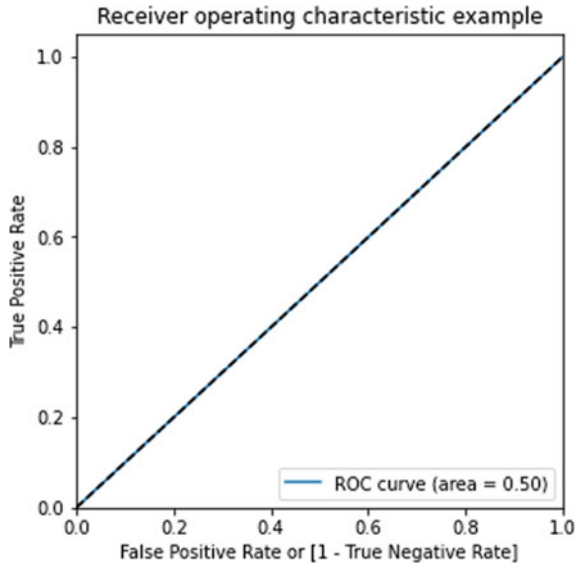
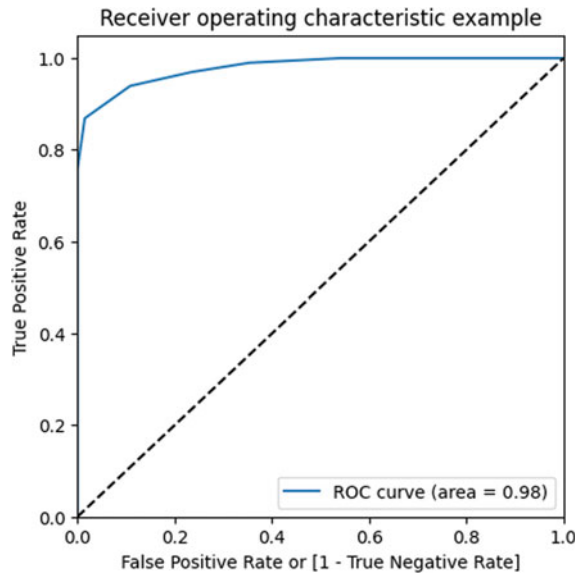
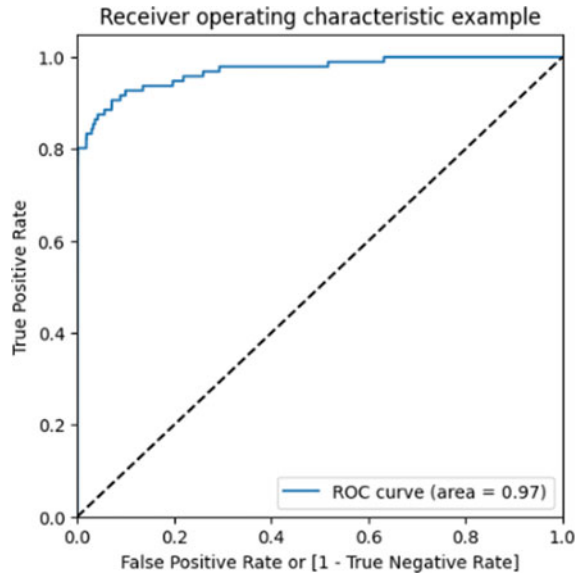


Fig. 25 RoC curve for decision tree



identifies the positive cases with a high probability and reduces the risk of misclassifying negative cases as positive. A high AUC-ROC score of 0.98 implies that the model is effective in differentiating between positive and negative classes, and it is a powerful indicator that the model is performing well in its classification task.

Fig. 26 RoC curve for random forest



An AUC-ROC with score 0.97 (see Fig. 26) suggest that the RF model has an excellent capability to differentiate between positive and negative classes, with a very high true positive rate and a relatively low false positive rate. In other words this particular ML model correctly identifies the positive cases with a high probability and minimizes the risk of misclassifying negative cases as positive. A high AUC-ROC score of 0.97 indicate that the RF model is good for predicting the fraudulent and non fraudulent transactions, and it is a strong indicator that the model is performing very well in its classification task.

8 Conclusions and Future Work

Credit card fraudulence is a never growing problem that has become a major concern for financial institutions and customers alike. Over the span of years, Machine Learning (ML) algorithms become an increasingly demanding tool for detecting and preventing credit card fraud. In this dissertation, the credit card detection system using ML algorithms are discussed along with its importance, and determined the best fit machine learning algorithm to predict the fraud transaction.

A credit card fraud detection system using ML algorithms is a software application that uses artificial intelligence (AI) algorithms in order to detect real-time fraudulent transactions. The detection system works by analyzing patterns in transaction data and identifying anomalies that are likely to be associated with fraud. These analyzed patterns are later then used to train the ML models to recognize and flag suspicious transactions. One of the key advantages of using ML algorithms for credit card fraud

detection is its potential to adapt and learn from new data efficiently. As new types of fraud emerge, the machine learning algorithm can be updated to detect these new patterns and prevent future fraudulent transactions.

This research used the open source credit card fraud detection data set that are available in the kaggle and during the first stage the data is preprocessed using undersampling technique. With the help of google collab and various python libraries different supervised learning algorithms are implemented use labelled data to train the algorithm to recognize specific patterns of fraud. The implemented algorithms are: Logistic regression (LR), XGBoost, Decision tree classifier (DT), and Random forest (RF). After creating the ML model, the data is fed into the model and its prediction accuracy is analyzed, in which it is identified that the logistic regression (LR) shows high performance than other Machine Learning algorithms.

In addition to machine learning algorithms, credit card detection systems can also use other technologies such as behavioral analytics, biometrics, and data visualization tools. Behavioral analytics can help to identify unusual patterns in user behavior, such as unexpected changes in spending patterns or unusual login activity. Biometric technologies such as facial recognition and fingerprint scanning can be used to authenticate user identity and prevent fraud. Data visualization tools can be used to help analysts and investigators identify patterns and trends in transaction data, allowing them to identify potential fraud more quickly and efficiently. By using a combination of these technologies, credit card detection systems can provide a comprehensive and effective solution for detecting and preventing credit card fraud.

In conclusion, credit card fraudulence is a growing problem that causes serious financial and personal consequences to victims. Machine learning algorithms provide a powerful tool to detect and prevent credit card fraud by analyzing the transaction data and identifying patterns that are likely to be associated with fraudulent activity. Credit card detection systems using machine learning, combined with other technologies such as behavioral analytics and biometrics, can provide a comprehensive and effective solution to detect and prevent credit card frauds in real-time. As technology continues to evolve, it is likely that credit card detection systems will become even more advanced, providing even greater protection for consumers and financial institutions alike.

References

1. Bengio Y, Bengio S (2020) The consciousness prior. <https://doi.org/10.48550/arXiv.1709.08568>. Accessed 25 Mar 2023
2. Bhandari A, Gupta S (2020) A review on performance evaluation metrics for machine learning models. *Int J Adv Sci Technol* 29(9):2843–2849
3. Bisong E, Bisong E (2019) Introduction to Scikit-learn. In: *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners*, pp 215–229

4. Cheema J, Raza K (2021) Data preprocessing techniques in machine learning: a comprehensive review. *Int J Comput Intell Syst* 14(1):944–971. <https://doi.org/10.2991/ijcis.d.210327.001>
5. Chen Y (2020) Credit card fraud detection using machine learning and data mining techniques: a systematic review. *J Financ Crime* 27(3):647–662. <https://doi.org/10.1108/JFC-03-2019-0044>. Accessed 05 May 2023
6. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP (2020) Introduction to machine learning, neural networks, and deep learning. *Transl Vis Sci Technol* 9(2):10–14
7. Copeland BJ (2020) Alan Turing and the mathematical obsolescence of intelligence. *Mind Mach* 30(2):231–254
8. Düntsch I, Gediga G (2019) Confusion matrices and rough set data analysis. *J Phys Conf Ser IOP Publishing* 1229(1):012–055
9. Fan W, Liu K, Liu H, Ge Y, Xiong H, Fu Y (2021) Interactive reinforcement learning for feature selection with decision tree in the loop. *IEEE Trans Knowl Data Eng.* <https://doi.org/10.48550/arXiv.2010.02506>. Accessed 03 May 2023
10. Ge D, Gu J, Chang S, Cai J (2020) Credit card fraud detection using lightgbm model. In: International conference on E-commerce and internet technology (ECIT), IEEE, pp 232–236. <https://doi.org/10.1109/ECIT50008.2020.00060>
11. Gohil S (2020) Comparative analysis of different performance evaluation metrics for breast cancer detection using ROC analysis. *Int J Adv Res Comput Sci* 11(3):107–114
12. Gulati M, Thakur M (2021) A comprehensive study on data analysis using Python. *Int J Recent Technol Eng* 9(3):9123–9130. <https://doi.org/10.35940/ijrte.C9799.129321>
13. Harvey H, Glocker B (2019) A standardised approach for preparing imaging data for machine learning tasks in radiology. In: Artificial intelligence in medical imaging. Springer, pp 61–72. https://doi.org/10.1007/978-3-319-94878-2_6
14. Huang J, Xie M, Wang Q (2021) A comparative study of data normalization techniques for deep learning. *Appl Sci* 11(6):2846
15. Khatri S, Arora A, Agrawal AP (2020) Supervised machine learning algorithms for credit card fraud detection: a comparison. In: 10th International conference on cloud computing, data science & engineering (Confluence), IEEE, pp 680–683
16. Kim M, Kim H, Kim J (2020) Credit card behavioral fraud detection using machine learning techniques. *Expert Syst Appl* 146:113–188. <https://doi.org/10.1016/j.eswa.2019.113188>
17. Kim M, Lee J, Ohno-Machado L, Jiang X (2019) Secure and differentially private logistic regression for horizontally distributed data. *IEEE Trans Inf Forensics Secur* 15:695–710
18. Kugler KG, Mueller KR, Mika S (2020) ROC curves: a review of methods with applications in machine learning. *Mach Learn* 109(1):45–80
19. Liang J (2022) Confusion matrix: machine learning. POGIL Activity Clearinghouse, vol 3(4)
20. Liu S, Zhang J, Wu X, Guo B (2020) A comparative study of performance metrics for imbalanced classification of COVID-19 data. *BMC Med Inform Decis Mak* 20(1):1–13
21. McNerney J, Astwood M (2021) A guide to data visualization in Python with Matplotlib. *J Open Source Softw* 6(61):3171. <https://doi.org/10.21105/joss.03171>
22. Mennella TA, Haynes RD (2021) Statistical analysis and data visualization using Python. *J Chem Inf Model* 61(4):1834–1840. <https://doi.org/10.1021/acs.jcim.0c01446>. Accessed 26 Apr 2023
23. Niu X, Wang L, Yang X (2019) A comparison study of credit card fraud detection: supervised versus unsupervised
24. Park M, Kim J (2019) Credit card fraud detection using a convolutional neural network. *Appl Sci* 9(21):4585
25. Raabe CA, Fritsch J (2021) Analyzing and visualizing data with python: a practical guide. Springer. <https://doi.org/10.1007/978-3-030-79701-5>
26. Ramezan C, Warner AT, Maxwell AE (2019) Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sens* 11(2):185
27. Razzak I, Choudhury SR (2019) Credit card fraud detection using machine learning: a systematic literature review. *Expert Syst Appl* 132:381–394. <https://doi.org/10.1016/j.eswa.2019.04.053>

28. Ristea A, Petrescu L, Grigoras G (2020) Performance evaluation metrics for classification problems in machine learning. *Adv Electr Comput Eng* 20(3):91–96
29. Roelofs R, Shankar V, Recht B, Fridovich-Keil S, Hardt M, Miller J, Schmidt L (2019) A meta-analysis of overfitting in machine learning. *Adv Neural Inf Process Syst* 32
30. Sailusha R, Gnaneswar V, Ramesh R, Rao GR (2020) Credit card fraud detection using machine learning. In: 4th International conference on intelligent computing and control systems (ICICCS), IEEE, pp 1264–1270
31. Santos MS, Embrechts MJ (2020) A review of methods for addressing class imbalance in machine learning. *J Big Data* 7(1):1–35. <https://doi.org/10.1186/s40537-020-00323-4>
32. Shaohui D, Qiu G, Mai H, Yu H (2021) Customer transaction fraud detection using random forest. In: International conference on consumer electronics and computer engineering (ICCECE), IEEE, pp 144–147
33. Singh RK, Dwivedi G, Singh SK (2020) Comparative study of performance metrics for machine learning models on sentiment analysis of social media data. *J Ambient Intell Humaniz Comput* 11(7):2985–2997
34. Subasi A (2020) Practical machine learning for data analysis using python. Academic Press
35. Tock K (2019) Google colab as a platform for Python coding with students. In: RTSRE Proceedings, vol 2(1)
36. Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. *PLoS ONE* 14(11):224–365
37. Yang L, Shami A (2020) On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing* 415:295–316
38. scikit-learn (2022) Decision Trees. [online] <https://scikit-learn.org>; <https://scikit-learn.org/stable/modules/tree.html>. Accessed 2023
39. scikit-learn (2022) Logistic regression. [online] <https://scikit-learn.org>; <https://scikit-learn.org/stable/modules/regression.html>. Accessed 2023
40. scikit-learn (2022) Random forest. <https://scikit-learn.org>; <https://scikit-learn.org/stable/modules/forest.html>. Accessed 2023
41. scikit-learn (2022) XGboost. <https://scikit-learn.org>; <https://scikit-learn.org/stable/modules/xgboost.html>. Accessed 2023

The Role of Trust and Ethics in IT Administration and Its Impact on an Organisation



Bijay Sunny George and Tasmina Islam 

Abstract Information Technology administration plays a crucial role in today's digital world, where devices are networked, and computer devices are used to accomplish tasks. A lack of policies and ethics fuelled by overtrust in this field can have a devastating impact on organisations and humans. The COVID-19 pandemic period was when organisations had to rush into technology solutions to ensure business continuity. Therefore, it is vital to understand the criteria businesses employ during solutions deployment. The interview questions primarily focussed on whether the trust was prioritised. If it was, questions were formulated in such a way as to understand the correlation between trust and ethics. This work further explored whether the trust of Senior management (elites) was exploited and whether it adversely affected the organisations. Through methodical data collection in the form of a semi-structured interview, meticulous qualitative work has been employed. Participants included Senior management (elites) and technicians. They primarily work in organisations in Seychelles, but their companies are headquartered in different countries in the Asian continent. Thematic analysis was used to analyse the responses, which generated three main themes: profile, solution deployment and trust.

Keywords Ethics · Information technology (IT) · Online trust · COVID-19

B. S. George (✉)

Department of Computing and Information Systems, University of Seychelles, Victoria, Seychelles

e-mail: bijay.george@unisey.ac.sc

T. Islam

Department of Informatics, King's College London, London, UK

e-mail: tasmina.islam@kcl.ac.uk

1 Introduction

The use of a computer device is inevitable for organisations. According to Petersen [1], most businesses use computers for communications, research, media production and more. These systems require to be periodically maintained and managed. Information Technology¹ administrators employed to manage and maintain computer systems are called System administrators, IT administrators, or IT technicians. They handle a diverse portfolio ranging from cabling, computer system setup, and solution deployment to handling the security of the entire network of their respective organisations. A day without the presence of an IT administrator would be unimaginable since not all end users are technology savvy. IT Administrators play a significant role in handling the day-to-day operations of an organisation's digital infrastructure.

In this technological era, information systems are an integral part of an organisation's operations. Technology plays a supportive role in almost all sectors, be it banking, tourism, education and more. Large, small, and medium Enterprises (SMEs), regardless of the industry, are in a state of digital transformation. There has been a digital transformation acceleration during the COVID-19 pandemic period [2]. According to Kutnjak [3], in the context of the Covid Pandemic, information and communication technologies (ICT) have been essential to ensure business continuity for individuals and businesses. He has also stated that the abrupt digital transition introduced by governments during the COVID-19 period had impacted businesses and individuals. An abrupt transition sounded like a lack of readiness to go digital, which therefore warranted future research to investigate. In this study, the role of trust in such a scenario will be explored, and the following research questions will be investigated:

- (1) Is the senior management's trust in IT administrators getting exploited and impacting the business adversely?
- (2) Is there a correlation between ethics and trust in IT administration?

First, a literature review will describe the relevance of IT administration in this digital era and the importance of ethics and trust in IT administration. Subsequently, the methodologies employed in this research will be explored. The results and analysis section will explore findings and critically evaluate them. Finally, discussion of findings, recommendations, and scope for future research will be explored. The essay is concluded under the final section, the conclusion section, where this study's relevance is again emphasised.

¹ For concision hereafter referred to as IT.

2 Literature Review

2.1 *The Importance of IT Administration in the Information Age*

According to De, Pandey and Pal [4], the COVID-19 pandemic drove the digital transformation for many organisations due to the social distancing standards and lockdown. Barrett et al. [5] have addressed system administrators as unsung heroes. To effectively run the digital operations of an organisation, the contributions of system administrators are significant. The same study mentioned the system administrator's behind-the-scenes work and its relevance in the success of a business, especially if it's a client-oriented organisation. This shows that an organisation's successful operation is directly linked to reliable IT administration.

2.2 *Trust*

Aljazzaf et al. [6], mentioned that the trustor could face undesirable consequences due to the trust laid on the trustee. Heavy reliability on anyone or any object comes with a risk. According to Forrest [7], system administrators can go rogue for various reasons, such as dysfunctional work culture, lack of gratitude, low salary packages, weak company policy, and lack of supervision and accountability. A rogue administrator could sell and leak the company's confidential data, delete user data, or even sell the credentials to a hacker giving him access to the company network. Trust and ethics become relevant in this context.

Proper monitoring policies allow a company to monitor the system administrator's activity. Marsan and Duffy [8] listed monitoring the system administrator's activity as one of the best practices to secure the system administrator's account, thereby reducing the risk of system administrator's going rogue. Mike Lorenzen, MCP, and VCP [9] identified that IT policies in place would set rules and guidelines for decision-making. When such policies guide a company, management tends to listen to policies than to system administrators and thus trust that the management develops in an employee is not misused. Fulmer [10] identified that little literature is available regarding trust across multiple organisational levels. He further mentioned that the trustee and trustor relationship is likely to take place once the trustee succeeds in creating positive perceptions and that trustors have concerns about the vulnerabilities that come with such a trust relationship [20, 21].

2.3 Ethics

Ethics is defined as the actions one undertakes from a good and evil perception [11]. The relationship between the trustor and trustee can worsen if the trustee does not adhere to the code of conduct or ethics, eventually impacting the organisation and its operation. LOPSA (LOPSA, n.d.) developed its code of ethics addressing system administrators that lists the following:

- (1) Professionalism
- (2) Personal Integrity
- (3) Privacy
- (4) Laws and Policies
- (5) Communication
- (6) System Integrity
- (7) Education
- (8) Responsibility to Computing Community
- (9) Social Responsibility
- (10) Ethical Responsibility.

These are well aligned with the IEEE Code of Ethics (Anon., n.d.).

3 Research Gaps

Mccarthy and Truchon [12] identified that organisational trust is a multidimensional concept and that trust culture is a reciprocal relationship. There have been numerous research conducted on organisational trust, which refers to the confidence of employees in the actions of their employer (Otto n.d.). However, there's little literature available that has researched the repercussions of trust in the IT administration context and its impact on the organisation. Substantial research has been conducted to study the response of IT administrators to business needs during the COVID-19 pandemic. Kaur et al. [13] studied the impact of COVID-19 on IT administrators and their day-to-day work. They identified that before the lockdown, certain processes relied on trust and procedures and rules were not strictly followed. However, there isn't any evidence of research conducted to identify the rationale behind implementing countermeasures and their repercussions. Within the IT administration setting, this study is aimed to identify a correlation between trust, ethics, and its impact [22].

4 Methodology

Quantitative research employed through questionnaires or surveys has its limitations since it doesn't provide room to filter out respondents with biases [14]. To establish the research questions, it was imperative to understand the perspectives of elites and technicians since they are the major actors in the IT administration scenario. Braun and Clarke [15], pointed out that thematic analysis is suited to a scenario where the perspectives of different participants need examination. Themes will be generated from the data collected. Hence qualitative research using an inductive thematic analysis has opted for this study.

4.1 *In-Depth Interviews and Recruitment*

Semi-standardized in-depth interviews were conducted with open-ended questions. An interview guide was prepared to ensure that all the required topics were covered. It also ensured that the interview was not driven by the respondents but by the researcher. Two categories of participants were interviewed: Senior Management (elites) and technicians, which included Senior System support officers and IT Managers.

Senior Management makes decisions based on the solutions and advice suggested by the technicians. As depicted in Fig. 1 (left), the gender distribution provides insights into the proportion of males versus females. This representation aids in understanding the gender dynamics within the dataset. Figure 1 (right) showcases the distribution of various organisation profiles or sectors. This visualization not only highlights the range of sectors represented but also indicates the dominance of specific sectors such as Wholesale, Education, Furniture, Finance, and Retail. Figure 2 (left) delves into the geographical spread of the organizations by portraying the distribution based on the location of their headquarters. Figure 2 (right) delineates the variety and frequency of different posts or designations within the organizations, from roles such as Director, IT Technician, IT Manager, Senior IT Manager, and Managing Director. Collectively, these charts (Fig. 1) serve as instrumental tools in summarizing and understanding the nuances of the dataset. The contact number of the participants were collected from their organisational websites, and they were informed of the interview for this study over a call. They were also informed that their participation is voluntary and doesn't involve incentives of any form.

Twenty-five participants were chosen; however, only 12 responded with their willingness to be part of this study. Participants provided verbal consent to participate in the interview. Out of the twelve, there had been an equal proportion of male and female participation, which happened by coincidence. Baker and Edwards [16] suggested using a sample size of twelve and sixty if time and interviewee characteristics pose a barrier to gathering a larger sample.

The interviews were scheduled for 1 h, with open-ended questions. Participants were recruited from Seychelles. However, some of the organisations they worked for

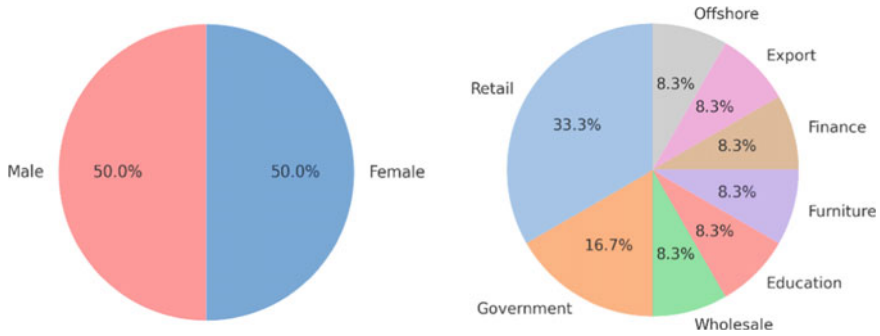


Fig. 1 Participants demographic distribution (left—gender, right—organisation profile)

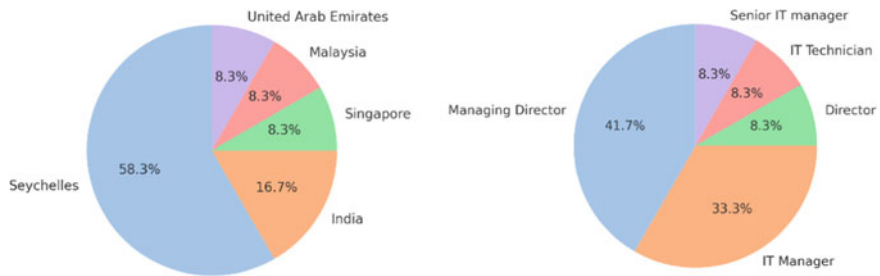


Fig. 2 Participants demographic distribution (left—organisation headquarters, right—position at work)

were headquartered in different regions of the Asian continent and hence their organisational goals, characteristics and policies were not restricted to the local context. This ensured that this study was wider than one country. One of the main reasons for recruiting participants from Seychelles is that it was easy to conduct face-to-face interviews with the elites. The interviews were conducted face-to-face in their respective offices due to the participant’s preferences. Only 2 out of 12 participants were agreeable to recording the interview, and hence the responses had to be transcribed. This had been a hindrance to the focus of the interview; however, the interview guide had helped to maintain the pace and to meet the goal.

5 Method of Analysis

The interviews were conducted, transcribed, and analysed by the primary researcher. Thematic analysis was employed to analyse the collected data as it provides a means to identify, analyse and report patterns in qualitative data [15]. They have summarised the phases of thematic analysis as follows:

- (1) Familiarisation with Data
- (2) Generating initial coding
- (3) Searching for themes
- (4) Reviewing themes
- (5) Defining and naming themes
- (6) Formulating the report.

The above steps were carefully followed during the analysis. It started with organising and structuring the collected data, followed by carefully coding the data from which specific patterns and themes were identified. There had been multiple mentions of the same code by different participants. This helped me to categorise the data with ease. The themes identified were reviewed and named, further which the report was created. This ensured that the analysis and findings were not driven by the researcher’s perspective but merely based on the gathered response from the respondents.

6 Results and Analysis

Transcriptions were carefully read and coded, and themes were identified. An overview of the code book can be found in Tables 1, 2 and 3. Analysing the interview transcriptions using thematic analysis revealed the following themes:

- (1) Profile
- (2) Solution Implementation
- (3) Trust.

6.1 Profile

The participant’s job profiles and their characteristics were collected. The interviewed participants included 6 elites and 6 technicians. It was found that 5 out of 12 participants undertook an all-in-one role with not just handling technology but other portfolios too. The rest of the 7 participant’s job responsibilities were

Table 1 Codebook overview

Profile	Responsibility	Qualifications	Operation style
	All-in-one roles (5)	Basic (3)	Rely on staff (4)
	Handles only technology (7)	Bachelors (8)	Self (6)
		No qualification (1)	Rely on consultants (2)

Table 2 Codebook overview

Trust	Belief	Factors
	Cognitive: positive prior experiences (8)	Job security (5)
	Mutuality: his job is also at stake if I get fired because of his acts (1)	Happiness with less pressure
	That solution was used by everyone, so I also use it because it must be secure (6)	Sense of responsibility and its after-effects
		Improved productivity

Table 3 Codebook overview

Solution implementation	Policy	Monitoring	Deployment type	Attacks
	All-in-one roles (5)	Basic (3)	Rely on staff (4)	Data leak (4)
	Handles only technology (7)	Bachelors (8)	Self (6)	Uninvited guests joining online meetings (5)
		No qualification (1)	Rely on consultants (2)	

restricted to overseeing only technology matters. This is very much aligned with the system administration responsibilities, which were mentioned in the literature review section.

“Participant ID² 1: I take care of the 360 aspects of the company, having a broader portfolio not just limited to technology. I devise and monitor strategies but do not operationalise the strategies.”

“PID 9: My responsibility includes handling, monitoring, and managing the tasks performed by my staff. However, since it is a very small company, I need to be responsible for other operational tasks too.”

The technicians, on the other hand, reported that they always perform hands-on duties. They are the real warriors that work behind the scenes.

“PID 8: I create users, automate tasks, ensure that the network is secure by applying policies at the firewall level and anything to do with network and computers.”

“PID 4: I manage the active directory, server upgrades, network upgrades and security. I also get the help of my colleagues. “In terms of qualifications, all the interviewed participants had qualifications in technology ranging from a diploma to Bachelor’s.

Five out of six elites responded that they fully rely on their staff to operationalise strategies. Based on the requirements and strategies, their staff create solutions.

“PID 1: There are limitations to myself and my team, and hence for specific projects, I bring in, consultants.”

“PID 5: Since I handle a wider portfolio and lack hands-on experience, I fully rely on my staff to full fill the requirements.”

² For concision, hereafter referred to as PID.

“PID 6: Well, My time is very limited to gather the necessary skills to attend to daily tasks, so I rely on my staff. Even during the COVID-19 lockdown, their swift actions by bringing in new solutions have ensured that the business wasn’t impacted.”

“PID 12: I am well experienced in the IT field to handle operations independently.”

“PID 3: I am lucky to have a director who gives complete freedom and approves the budgets as per my suggestion. I have never let him down.”

Understanding the profile of participants was imperative as it formed the baseline for understanding the gaps that have contributed to the exploitation of trust. It is very evident from the above statements that technology directors had to fully rely on their staff, either due to their lack of hands-on experience in the field or because they handle multiple portfolios and often tend to ignore their responsibilities related to technology. On the other hand, the technicians who are well experienced in handling technology get the complete freedom to implement the strategies which they think are the best.

6.2 *Solution Implementation*

Participants were questioned about the rationale they considered while implementing solutions during the COVID-19 pandemic and before the pandemic. Two elites mentioned that they go by the policies, and an IT technician mentioned that he used to refer to the policies while he decided on solutions but did not always abide by them. A few of them needed to learn about the importance of having policies.

“PID 7: During COVID, we had to bend our policies a bit; however, most of the solutions we used aligned well with the policy.”

“PID 11: I have read about policies, but we spent most of our time running the operations, and did not time get to devise policies.”

“PID 12: The management had sudden requirements which were almost impossible, so we only focussed on implementing solutions rather than prioritising policies.”

The two elites mentioned that they make a policy-driven approach. However they mentioned that there had been no quality or verification checks.

“PID 7: We rarely reflected on the implemented solutions. You know we need to trust the experts. We used to do that before COVID; however, due to the lack of a business continuity plan, at the time of COVID, we were in a rush and couldn’t do an audit.”

Eleven out of twelve participants mentioned that they used online platforms and relied on cloud-based solution providers rather than their own data centres. One of them mentioned that they are using a hybrid infrastructure.

“PID 2: Our data is stored in the cloud.”

“PID 4: We use cloud-based platforms such as WhatsApp, Zoom, and OneDrive.”

“PID 12: During the covid period, we moved from 100% on-premises to a hybrid server environment, where data critical servers are placed on-premises while less critical data is placed with cloud service providers. We also have tailor-made agreements with the cloud service providers.”

Nine out of twelve participants recalled that they suffered attacks. Majority experienced attacks during online meetings, with the presence of uninvited guests. One of them recalled that such an event affected their reputation and business.

“PID 5: It was devastating. We even lost a project. I realised at times I was blind and left the whole control of technical decisions to my staff.”

“PID 11: Further to data leaks, my line manager always questioned the security of proposed solutions.”

From the above responses, it is evident that most participants rushed into deploying solutions during COVID-19. An alarming fact is that over 90% of the participants deployed solutions that did not align with their company policies. The blind reliability on their staff is the main reason that forced elites to decide on using a particular solution. “We trust the experts” is a comment made by most elites. In the next sections, I have argued that it is such trusts that have formed the root cause of attacks.

6.3 *Trust*

The word trust repeatedly came up in the responses. Trust could be in people, platforms, or the technology itself. According to Cheshire [17], trust refers to credibility, security, surety, and reliability. There had been an indirect citation of some of these factors, such as job security and reliability, by the respondents. These factors were carefully extracted during this analysis process. Additional factors include happiness and a sense of responsibility. While elite respondents mentioned that reliability through trust makes them worry less and helps them to stay calm and happy, their staff mentioned that it comes with a sense of responsibility. On one side, trust lessens the pressure on elites, but on the other side, it adds pressure to their staff.

Elites used trust to stay within their comfort zone. They have created an imaginary world where they assume operations are implemented by their staff appropriately. They believe it gives job security and happiness. However, an attack could turn their life upside down as the impact of an attack could be devastating. Some of their staff confessed that they are forced to practice unethical ways to satisfy the requirements of their bosses because of the sense of responsibility that trust creates. This indicates that a correlation exists between trust and ethics. Nine mentioned that they were compromised further to solution deployments, out of which there were five elites. This proves that the trust of senior management is being exploited. An interesting observation is that eleven of them mentioned that “trust enhances productivity,” which is a contradiction. This is due to their lack of awareness that the root cause of those attacks is trust.

Cheshire [17] also identified the importance of online trust, which is the relationship between humans and computer systems. He further mentioned that trusting a

computer system or a technology is very similar to how humans trust each other. This could be based on certain behavioural and emotional characteristics. These characteristics were reflected in the responses, where the participants mentioned that they tend to believe in a system based on how effectively their staff have behaved in the past. Their prior interactions are considered, and this is one of the factors that form the basis of trust. One of them mentioned that their trust in the consultants heavily depends on the consultant's ratings and reviews on related platforms, for instance, gig platforms such as Upwork. This forms signals of trustworthiness. Such assurance and trustworthiness signals apply to system trust in any online service [17].

Mutuality is another factor observed from the responses, where one staff trust another staff due to the presence of common ends in their relationship, as mentioned by (PID4). Roderick et al. [18], identified that mutuality could be grounds for trust when both parties stand in common ends.

Eleven of them mentioned that trust enhances productivity, and one raised concerns that trust does not necessarily improve productivity.

“PID 1: I can cite numerous situations where my staff saved me and my companies reputation in the most turbulent period. There are situations where they have volunteered themselves and received no benefits. They were so reliable. In majority cases, they proved themselves as trustworthy.”

“PID 3: I do not want to complicate myself and take additional pressure. I would happily, trust my staff as it keeps me healthy without pressure.”

“PID 4: A wrong decision from my staff can affect themselves. I may lose my job for taking a wrong decision, but my staff's job, too, is at stake if it happened due to the suggestion that he gave me. I do not rely on staff who depend on side income. They may not take their primary job seriously.”

“PID 11: I look for companies and platform that has a lot of subscribers.”

“PID 6: There have been situations where my staff hasn't performed to the expected level, but due to the lack of trust in me as the manager, my boss hasn't taken any actions up till now. There had been a one-off case where my actions resulted in my boss's distrust.”

“PID 8: My manager sometimes asks me to submit the online activities performed by his colleague unofficially. At times I tend to avoid official gatherings and parties since he ask the activities of his colleagues during those times, which I am know is unethical. My friends have same experiences and I feel it is normalised in this country.”

7 Discussion and Future Recommendations

This section briefs the themes and codes and relates them to the relevant literature. The researcher will also argue points to answer the research questions mentioned in the introduction.

As mentioned in the above sections, a dependency factor was observed during the analysis, and it has been found that it is this dependency that has contributed

to the trust being exploited and ethics getting bypassed. While analysing the theme profile and related codes, a dependency on staff by the senior management has been observed mainly because the majority does an All-in-one role not necessarily limited to technology but also other roles. Their lack of time, policy unawareness and hands-on experience made them fully rely on their staff. Lack of necessary monitoring and verification of the suggested solution, misaligned with the company policies, contributed to organisations being vulnerable.

Most participants mentioned that there had been a rush to implement the technical solution during the COVID-19 pandemic. The lack of preparedness or a business plan contributed to this rush. From the responses, it has been observed that with no policies and tailor-made agreements, organisations were inclined towards using cloud platforms. The rationale that the majority chose a particular solution was that a bigger population was using it, creating a sense of online trust. Cognitive, behavioural, and emotional characteristics too added to this trust. Nine out of twelve participants revealed that they were compromised in different ways.

It can therefore be strongly argued that the rationale for an organisation to choose a solution shouldn't be based on online trust; instead, they should be guided by respective policies. COVID-19, triggered in one country, impacted the whole world. Similarly, the reason that billions use a solution does not necessarily mean it is the best solution. On the contrary, when billions use one solution, that company is more visible and makes itself a recognisable target for attackers. When the popular app Zoom became the de facto standard for teleconferencing during the COVID period, it became an interesting target to the attackers resulting in zoom bombings [19]. Further to such incidents, the senior management realised that they should have been vigilant with solution deployments and its administration, emphasising policies rather than trust in the staff. It can be deduced from the responses that reputation damage resulting in business loss is one of the major impacts that overtrust in system administration has contributed to the affected organisations, especially for client-oriented businesses. This also proves that interpersonal trust and trust in systems can get exploited.

The recommendation would be to audit the existing systems and implement policies. For small and medium organisations, hasty audit and policy implementation could be challenging; however, stringent monitoring mechanisms should be enforced. Monitoring the monitorer should be put in place to ensure ethics are observed. All system administration staff, including the senior management, must follow the SAGE code of ethics (LOPSA, n.d.).

8 Limitations and Future Research

Even if the minimum sample requirement of 12 was met, the output of this research could have been enhanced with the help of a larger sample of company representatives from diverse cultures. A general observation from the responses was a dependency on the staff. The dependency factor and its relation to trust requires to be explored further.

Policy unawareness, too, was noted. Cultural factors also could have been a reason for this kind of attitude, which requires further study. Participants represented companies headquartered in India, Seychelles, Singapore, Malaysia, and United Arab Emirates (UAE). They constitute samples from the Asian region. To establish the research questions globally, future research that contains samples from diverse regions would benefit. The study lacked the support of a second coder, which would have helped during the theme review process. It would have also enhanced the credibility of this study.

9 Conclusion

This research has explored issues that trust in the system administration context can create. There seemed to be little research available regarding how trust can get misused in a system administration context, and hence this study occurred. Twelve participants from multiple organisations in Seychelles but headquartered in surrounding Asian countries were selected and interviewed. The transcriptions were analysed using thematic analysis, which was coded and organised into three themes such as profile, solution implementation and trust. A new factor which wasn't referred to in any related studies has been identified during the analysis, which is the dependency factor. The study also revealed that trust could create a sense of responsibility overhead, leading to a lack of ethics. The responses showed that there exists a correlation between trust and ethics. Trust gives the liberty to act without ethics, at times not necessarily on purpose, but due to the responsibility overhead that trust creates on IT administrators. The study concluded that with policy and ethics observed, the trust could result in a positive impact on the organisation else; this could adversely affect the organisation.

Ethics Ethics clearance had been granted by the ethics office of King's College and registered with reference number MRSU-22/23-34,724.

References

1. Petersen L (2019) Importance of computers in business. <https://smallbusiness.chron.com/importance-computers-business-4012.html>
2. Stackpole B (2021) Digital transformation after the pandemic. <https://mitsloan.mit.edu/ideas-made-to-matter/digital-transformation-after-pandemic>
3. Kutnjak A (2021) Covid-19 accelerates digital transformation in industries: challenges, issues, barriers and problems in transformation. Research Gate, May
4. De R, Pandey N, Pal A (2020) Impact of digital surge during Covid-19 pandemic: A viewpoint on research and practice. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7280123/>. Accessed 23 Nov 2022
5. Barrett R, et al (2004) Field studies of computer system administrators: Analysis of system management tools and practices. s.l., s.n.

6. Forrest A (2021) Do you trust your System Admins?. <https://intrix.com.au/articles/do-you-trust-your-system-admins-3-types-of-rogue-system-administrators/>. Accessed 17 Nov 2022
7. Marsan CD (2010) 6 tips for guarding against rogue sys admins. <https://www.csoonline.com/article/2125997/6-tips-for-guarding-against-rogue-sys-admins.html>
8. Lorenzen M (2018) The benefits of documented IT policies & procedures. <https://www.linkedin.com/pulse/benefits-documented-policies-procedures-mike-lorenzen-sphr-mcp-vcv>
9. Fulmer CA, Gelfand MJ (2012) At what level (and in whom) we trust: trust across multiple organizational levels. *J Manage* 1167–1230
10. Bartneck C, Lütge C, Wagner A, Welsh S (2021) What is Ethics?. s.l.:s.n.
11. Mccarthy V, Truhon S (2016) A primer on organizational trust: how trust influences organizational effectiveness and efficiency, and how leaders can build employee-employer relationships based on authentic trust. https://www.researchgate.net/publication/294722472_A_Primer_on_Organizational_Trust_How_trust_influences_organizational_effectiveness_and_efficiency_and_how_leaders_can_build_employee-employer_relationships_based_on_authentic_trust
12. Kaur M, Parkin S, Janssen M, Fiebig T (2022) “I needed to solve their overwhelmness”: How system administration work was affected by COVID-19. Association for Computing Machinery, 11 November, pp 1–30
13. Andrade C (2020) The limitations of online surveys. *Sage J* 13 October, 575–576
14. Braun V, Clarke V (2006) Using thematic analysis in psychology. *Research Gate*, January, pp 77–101
15. Baker SE, Edwards R (2012) How many qualitative interviews is enough. https://www.researchgate.net/publication/277858477_How_many_qualitative_interviews_is_enough. Accessed 11 Nov 2022
16. Cheshire C (2011) Online trust, trustworthiness, or assurance?. *Daedalus*, 11 October
17. Weigert JDLAA (1985) Trust as a social reality
18. Lorenz T (2020) ‘Zoombombing’: when video conferences go wrong. <https://www.nytimes.com/2020/03/20/style/zoombombing-zoom-trolling.html>
19. Aljazzaf ZM, Perry M, Capretz MA (2010) Online trust: definition and principles. <https://ieeexplore.ieee.org/document/5628836>. Accessed 19 Nov 2022
20. Anon (n.d.) IEEE code of ethics. <https://www.ieee.org/about/corporate/governance/p7-8.html>
21. LOPSA (n.d.) Code of ethics. <https://lopsa.org/CodeOfEthics>. Accessed 18 Nov 2022
22. Otto W (n.d.) What is organizational trust (and how to build it)?. <https://www.predictiveindex.com/blog/what-is-organizational-trust-and-how-to-build-it/>. Accessed 19 Nov 2022

How Explainable Artificial Intelligence (XAI) Models Can Be Used Within Intrusion Detection Systems (IDS) to Enhance an Analyst's Trust and Understanding



Chelsea Shand, Rose Fong, and Usman Butt

Abstract An intrusion detection system (IDS) is a fundamental tool when deploying cyber defence within an organisation. The ever-evolving landscape of cyber threats has pushed an advancement in the application of artificial intelligence (AI) within such tools to pioneer more sophisticated detection techniques (Wang et al. in *IEEE Access* 8:73,127–73,141, 2020). Cyber security analysts rely on these technologies to make critical decisions and correctly identify and prevent malicious threats to their organisation. It is therefore imperative that analysts can understand, trust, and have confidence in the IDS decisions (Neupane et al. in “Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities”, arXiv, Ithaca, NY, 2022). However, the advancement of these technologies has led to complex AI systems that lack transparency and are difficult for human analysts to comprehend. This research explores how these issues can be elevated by using explainable AI (XAI) to help make complex AI models more understandable (Kelley and George, “How to solve the Black Box AI problem through transparency,” 16 August 2021. [Online]. Available: <https://www.techtarget.com/searchenterprisecai/feature/How-to-solve-the-black-box-AI-problem-through-transparency>) and add clarity and context to their decisions. Ways in which trust can be measured, and the factors affecting trust, are identified through this research to analyse how the perceived ease-of-use, trust, and confidence of an analyst can be increased through the adoption of XAI. Key findings from this have been demonstrated through a recommended implementation approach for an XAI model, and proof-of-concept user-interface (UI) design. This research brings recognition to the need for explainability within IDSs and provides a user-centric approach to doing so.

Keywords Artificial intelligence · Explainable AI · Intrusion detection · Explainability

C. Shand (✉) · R. Fong · U. Butt

Department of Computer and Information Sciences, Northumbria University, London, UK

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_17

321

1 Introduction

With cyber-attacks continuing to grow, in complexity and numbers [4], it's important organisations have visibility over their network to ensure effective and timely detection of potentially malicious activity [5]. The dependence on intrusion detection systems (IDSs) is therefore increasing within organisations, as is the need for these systems to continually enhance their capabilities. AI-embedded IDSs are proving more effective at detecting intrusions than standard signature based IDSs [1]. They are therefore becoming more widely adopted [2]. However, as these AI-embedded IDSs become more advanced, as does their complexity, which leads to growing issues for human analysts to be able to understand the decisions and reasonings behind IDS actions [1]. In turn this affects the transparency and trust in an IDS. The increased awareness of such issues has led to a drive in the research and development of XAI to help make complex AI models more explainable [3].

This research explores the benefits of model explainability within IDSs, and addresses methods for implementing XAI in order to recommend the best suited approach to provide local explainability. The aim is to lessen the issues brought about by complex AI models and allow security analysts to triage alerts more effectively and efficiently, enhancing their trust and understanding. Existing challenges have been identified to bring recognition to the need for explainability within IDSs. Interview data collected from a sample of cyber security analysts has been analysed to assess what is lacking from current intrusion detection explanations and what is therefore needed to make a good explanation. The success of the proposed methods will be measured against key findings from interview data to determine whether XAI can provide the missing factors. This research will explore the following questions:

1. What makes a good explanation in the context of intrusion detection alerts?
2. Can XAI improve the useability for security analysts when triaging IDS alerts?
3. How can we build trust and provide transparency in the decisions made by IDSs?

1.1 Background

Explainable AI integrates techniques to generate human-understandable explanations of AI decisions and models [6]. XAI is an under researched area within intrusion detection [1], with most studies focusing on XAI within fraud detection technologies or medical advancement technologies. On top of this, existing research tends to focus more on how well the explainable model works, as opposed to how successfully it enhances a human's understanding and trust of the model outcomes. Intrusion detection is a key part of a robust cyber security strategy, allowing malicious network activity to be identified before it compromises the confidentiality, integrity, and availability (CIA) of an organisation's assets [7]. Traditional IDSs require knowledge of known attacks to be able to compare behaviour in the network against those attacks and then determine if something is a threat. With the emergence of new, rapidly

evolving, cyber threats and their sophisticated tactics, it is difficult for these traditional systems to protect against the unknown. AI technologies have therefore been embedded into IDSs to provide this level of capability by using machine learning algorithms which can learn anomalous behaviours and detect emerging threats as they try to traverse the network [8]. Human analysts are therefore using IDSs to make critical decisions and take appropriate actions against detections. But with limited understanding of why or how the IDS has come to that decision it can be difficult for humans to take the best course of action. Therefore, reviewing the explainability of an AI system should be just as important as reviewing the accuracy of its detections [1]. XAI can be used to alleviate the discussed issues, and in turn shed light into the complex AI models, to enhance human understanding of intrusion detections. It is therefore critical to contribute to this research area to improve the relationship between humans and AI and create a shift in the way organisations can successfully defend against cyber-attacks [9]. This research is significant for the development and acceptance of AI-embedded IDSs.

2 Related Work

2.1 AI Technologies

Artificial Intelligence (AI) can be defined as the development of smart technologies that allow machines to simulate human intelligence. Machine learning (ML) is a component of AI which allows systems to process large data sets to find patterns, provide insights, make informed decisions, and find connections across different data points through trained models and algorithms [10]. The complexity of these algorithms and models mean that most AI technologies are considered to be ‘black boxes’. This is because the AI models don’t typically share their inputs or operations with the end user, causing a lack of transparency and significantly complicating a user’s trust in the AI algorithms [3]. The black box nature of these AI algorithms also presents an absence of explainability, which makes them less comprehensible to humans, and in turn, leads to uncertainty from a user’s perspective [11]. AI models can be used to help humans make fair and accurate decisions [12], but without trust, certainty, and understanding, it’s difficult for humans to do so.

The adoption of AI technology is ever increasing, with organisations implementing digital transformation plans to keep up with the changing landscape we are in. At the start of the covid pandemic PwC reported that 52% of companies accelerated their AI adoption plans to boost productivity, manage disruption, alleviate skills shortages, and deal with the change in working patterns [13]. Within security, digital infrastructure is more interconnected than ever. Security focused AI technologies are being adopted to help organisations mitigate attacks more efficiently, allowing them to operate with minimal impact to their business. Intrusion Detection Systems (IDSs) are an example of such technologies. AI-embedded IDSs work by learning

patterns of behaviour in an organisation's network, including user activity, servers, and devices. The IDS can then learn to recognise anomalous behaviour and detect threats such as network intrusions, DDoS attacks, and data breaches. This level of advanced technology is critical in being able to protect against the rapidly evolving cyber threats [8]. However, successfully integrating these technologies into an organisation relies heavily on the end user's acceptance of it. There are many factors that can affect user acceptance, with trust considered a significant factor [11]. Recent research assessed trust as a predictor for technology acceptance [14]. When it comes to designing trustworthy AI, input is required from not only the people designing it, but from the people using it [12].

2.2 *Measuring Trust in AI Technologies*

With more organisations adopting AI-embedded technologies to perform critical tasks, and with the acceptance of such technologies relying heavily on trust, the need to therefore assess what influences trust is imperative [11]. Trust plays a critical role in the perceptions and acceptance of AI technologies [11] and is a fundamental trait of human nature. It allows us to cope with vulnerability, uncertainty, and complexity [11]. Trust is complex to define and is influenced by many factors, particularly when looking at human trust in AI technology. Glickson et al. review trust factors across two domains, cognitive trust, and emotional trust. They identified that tangibility, transparency, reliability, and immediacy were most important for cognitive trust, and the attribution of human characteristics was most important for emotional trust [14]. Choung et al. similarly performed an analysis of how human-like capabilities influenced trust, as well as general functionality of the AI technology. They identified that perceived ease of use, reliability, transparency, ability, integrity, and helpfulness are key factors [11]. Assessing the presence and importance of these factors for human analysts, in relation to their experience with AI-IDSs, is an integral part of this research.

Ashoori et al. performed a qualitative analysis on the different factors influencing trust in AI-embedded technologies, and how these factors were affected in different scenarios. The seven factors assessed were decision stakes (how critical is the decision being made), decision authority (who makes the final call, AI vs. human), model trainer, model interpretability, social transparency, and model confidence (how confident is the AI model in its decision). They found strong statistical evidence that the use of interpretable models, as well as transparency in the how the models were trained and tested, leads to an increase in trust between users and the AI technology [12]. From their testing, interpretable models were consistently preferred over black box models and were rated more reliable and trustworthy [12]. Glikson et al. had similar findings, reflecting that explanations on the rationale behind decisions or mistakes made by the AI algorithms positively impacted a user's trust significantly [14]. Certainty, or model confidence, is a key factor to consider when looking at

building trust. If explanations can reflect how confident the model is in its prediction, this leads to greater transparency and in turn, increases trust [15].

It is clear from research that XAI brings an advancement to AI technologies that could mean an increase in trust and transparency.

2.3 *AI-Embedded Intrusion Detection Systems (IDS)*

CheckPoint reported that 2021 saw an “unusually high” volume of attacks [16], cemented by Forbes who reported that the average number of cyber-attacks in 2021 increased by 15.1% when compared to 2020 [4]. Both these reports highlight the need for organisations to implement able technology to provide resilience and protect against these growing threats [4]. Intrusion detection systems were developed in response to the increasing number of cyber-attacks [7].

There are two main types of IDSs, signature-based (SIDS) and anomaly-based (AIDS). SIDS use knowledge-based methods, relying on signatures of known attacks and using pattern matching techniques to detect those attacks [17]. This presents significant challenges and limitations for IDSs by making it hard to detect the unknown [17]. Evolving network-based threats can be missed, which could cause significant damage to an organisation if not detected. Therefore, AIDSs are being implemented and favoured over SIDSs [1], to avoid the limitations of relying on known attacks [17].

AIDSs use machine learning methods and AI to improve detection rates [1] by detecting when traffic deviates from what it has learnt to be ‘normal’ for that particular organisation [18]. AIDSs are therefore complex in nature, which provides drawbacks for human analysts who are required to interpret the decisions made by the complex models [1]. This leads to a lack of transparency and trust [7]. Therefore, more researchers are looking into ways in which XAI models can be used to alleviate these issues.

2.4 *Explainable AI (XAI)*

A lack of explainability can restrict end users from being able to fully participate in the decision-making process [19]. XAI aims to give humans the ability to understand and interpret how an AI model arrives at certain decisions [20]. Explainability has been defined as “the degree to which a human can understand the cause of a decision” [15]. AI models are therefore more explainable if a human can comprehend their decisions or predictions and extract the relevant information and recognise indicators for certain predictions [15]. Research has found that explanations are more valuable to humans when they provide transparency [19]. Transparency should not only provide clarity on why a decision or action has been made but also the reason why a model may have failed. Knowing why it has succeeded or failed can help humans learn more

about the data itself and therefore make informed decisions [15]. For an analyst triaging IDS alerts, this can help them mitigate false positives. The goal of XAI is to provide useful and transparent explanations as well as provide accountability, fairness, and trust [19]. XAI can be achieved through various methods, for example, Perez et al. assessed how XAI could be used to make complex deep-learning models more explainable by looking for uncertainties in the data [21]. Other methods look to highlight elements of the data that have contributed to a model's prediction [18]. Explainability methods can be implemented in two ways, locally or globally, and can be model-specific or model-agnostic. Local XAI methods look to explain specific predictions, whereas global methods explain the entire model behaviour [15]. Model-specific explainability methods are specific to a model and therefore can't be applied to different methods. In contrast, model-agnostic explainability methods can be used on any machine learning model and are applied after the model has been trained [15]. This research focuses on local explanations that are model-agnostic.

Figuring out the best method for implementing and achieving XAI relies on understanding what kind of explanations are needed and which are important, what goals it needs to achieve, and who it is benefitting [19]. For this research, the goal of implementing XAI within IDSs is to enhance an analyst's ability to understand the decisions made and why certain actions have been taken. Explanations that aid investigation by making it clear how the threat was detected, what has been affected, and hence what steps need to be taken to help resolve or mitigate such a risk, are key. This directly benefits the analyst, as well as the organisation whose network is being protected.

2.5 *Using XAI Models Within IDSs*

A review of existing frameworks and research into how XAI has been used within IDSs, is significant to this research. Wang et al. present an XAI framework for IDSs based on SHapley Additive exPlanations (SHAP). Barnard et al. also propose a solution using SHAP to leverage explanations from the supervised IDS model, Extreme Gradient Boosting (XGBoost). SHAP is a method which focuses on explainability of individual predictions. It computes the contribution that each element has to a prediction [15], which provides insights on the features that give rise to different cyber-attacks [18]. Both Wang et al. and Barnard et al. agree that SHAP offers a "state-of-the-art approach" for obtaining explanations and "has solid theoretical foundation", which has provided significant advances in XAI [1]. SHAP was able to provide both local and global explanations for Wang et al. and has the advantage of being able to be applied to any model [1]. Barnard et al. compared and evaluated the use of SHAP for unsupervised (AIDSs) and supervised (SIDSs) models and found similar accuracies across both [18]. In contrast, Mahbooba et al. present a solution using decision tree algorithms and evaluated the accuracy, precision, and recall of their method in comparison to other supervised algorithms. A decision tree builds predictive models from a given set of observations and demonstrates the chance of

event outcomes [7]. Although the models presented across the three literatures differ slightly, all share a common goal in helping to improve the explainability of intrusion detection models with the aim to enhance trust and transparency.

To test the proposed solutions discussed above, researchers used the NSL-KDD dataset. This dataset is widely used as a benchmark for intrusion detection data [17]. This dataset was one of the first to contain a wide range of realistic network-based cyber-attacks and can be used to evaluate the performance of IDSs [18]. With this dataset being so widely used to test evaluations of proposed methods, it provides consistent comparable results across different research and therefore enables informed decisions to be made on what approach works better [17]. For example, Bernard et al. was able to compare their solution directly with the Wang et al. framework and find that theirs performed higher on accuracy, recall, and precision [18]. The NSL-KDD data set contains a high volume of records, with over 20 different types of intrusion attacks, making it sufficient for testing against [17]. However, it has some limitations in capturing more modern intrusion attack types [18]. Despite this limitation, this dataset has been able to measure the feasibility of proposed XAI solutions on IDS traffic with a high level of confidence.

3 Research Methodology and Design of Practical Work

3.1 Methodology

This research has been conducted through an interpretivism paradigm, analysing the benefit of explainability within IDS models through the perspective of a human analyst; applying meaning to the data being analysed through the reasoning of individuals [22]. Inductive reasoning has been applied to allow for valuable insights to be gained from the data collected, via interviews and through literature reviews. Only literature that has either been peer reviewed or published within the past 4 years by accredited authorities, was used to ensure relevance, validity, and reliability. Descriptive research methods were applied to the findings with the aim to accurately describe a population through measuring observations and allowing for frequencies and trends to be identified to help draw conclusive theories [23]. This research has produced a normative study, providing recommendations based on comparisons and analysis of secondary and qualitative data. Using normative methods meant that present state could be evaluated, and improvements could then be recommended, as well as future developments. The aim of this research is to improve the usability of IDSs by enhancing analyst's capability to interpret and understand IDS decisions, hence a normative approach. Qualitative analysis has been used to provide credible insights within the specific research area [24]. It is hypothesised that XAI will improve the relationship between human analysts and AI-embedded IDSs, in turn improving the usability of such tools, by increasing trust through more transparent, reliable explanations.

3.2 *Design of Practical Work*

This research involved two practical stages, the first being a set of Interviews and the second a Proof-of-Concept (PoC) involving a recommended XAI model and user interface (UI) design.

Interview Strategy. Interviews were semi-structured, allowing for set questions to be asked within the research framework, whilst also leaving room for flexibility in the questions and answers. Existing questionnaires assessing human perceptions of AI were reviewed to aid with producing suitable questions for the interview. As well as this, factors considered to influence trust were incorporated into the interview questions to measure whether those factors exist for the users. A pilot interview was conducted to test out the validity of the interview questions before undertaking the main interviews. Feedback was gathered from the pilot and was used to enhance the interview strategy and adjust the questions as necessary. Interview answers were analysed using a qualitative data analysis software to extract useful insights and build helpful visualisations of the data. A codebook was built and used to conduct the analysis. Due to a lack of existing research and limited theories on this topic, a blended method of Framework Analysis and Grounded theory was used to formulate a theory from this analysis on why explainability matters within IDSs and why it's important for cyber security analysts to understand the reasonings behind the IDS actions.

To ensure reliable insights could be gained, the interviews were conducted with ten cyber security professionals, who have familiarity with IDSs and analysing their outputs. Their data was pseudonymised to protect their identities and to avoid collecting any personally identifiable information (PII). Consent for each participant's involvement in the interviews was obtained, as per the ethical obligations followed by this research.

Proof-of-Concept (PoC) Strategy. Existing methods and frameworks for implementing XAI models were analysed to explore opportunities, challenges, and limitations related to this research. From this, a suitable XAI method was recommended that achieves the desired level of explainability within IDSs and meets the end user's needs. On top of this, a recommended PoC UI prototype was built to include relevant features that help accommodate explainability within the UI. Key findings from the interview data, such as missing elements in current AI embedded IDSs, were incorporated to ensure it aligned with the aims of this research study, and most importantly catered to the needs of the user, aligning with the user-centric approach of this research study.

4 Data Analysis and Evaluation

This analysis explores which factors influence an analyst’s trust in an IDS, how trust further influences factors like confidence and immediacy, and how XAI can help enhance these factors as well as an analyst’s understanding of how the AI works within the IDS.

Of the 10 participants, 6 had over 3 years of experience working with an IDS. Going forward, experience will be discussed in two groups, the less experienced participants (0–1 years and 1–3 years), and the more experienced participants (3–5 years and 5+ years). These groups have been used in Fig. 1 to compare how AI understanding differs between each.

Over 80% of the more experienced participants showed more understanding of AI than those with less experience, with whom only 25% showed some understanding. This indicates that AI understanding could improve with experience, as well as other factors which will be explored further in this chapter. The participants that expressed a lack of AI understanding were asked why they believe that could be the case and there were four main themes, including Outside scope of role, Lack of information, Model complexity, and Lack of training.

When comparing participants with at least some AI understanding to their perceived reliability of IDSs, over 65% of them showed a positive perception towards reliability. This could indicate that where users have more understanding of AI within IDSs, they are more likely to have positive perception towards the reliability of the IDS. Of these participants, 75% of them are also within the high confidence group. Reliability can therefore in turn help to build trust by making users more confident in the decisions the IDS makes, and how it makes them. Comparing experience with confidence, of the 6 more experienced participants, 50% were somewhat confident, and 50% were highly confident. None of the participants showed no confidence. Of those with less experience, 75% were highly confident and 25% were somewhat confident. From these results, it is hard to tell whether experience has a direct influence on a user’s confidence in the IDS. The three more experienced participants who demonstrated high confidence discussed that the amount of activity being detected

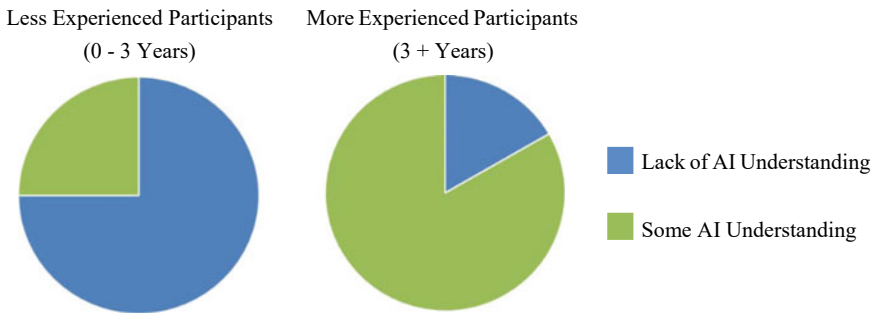


Fig. 1 Comparison pie charts for experience versus AI understanding

within the IDS gives them the confidence that it is working as it should, and detecting what it needs to.

Comparing experience and trust, none of the more experienced participants showed low trust, 83% demonstrated somewhat trust and 17% demonstrated high trust. Among the less experienced users however, 25% showed low levels of trust, and only 50% demonstrated somewhat trust, and 25% demonstrated high trust. Overall, of the 7 participants that somewhat trust an IDS, 3 participants explicitly said they trust the IDS more when humans are involved, showing a strong indication and desire for immediacy. In the case of this study, this would mean direct involvement with the IDS through decision making and implementing exclusions, etc. This demonstrates the importance immediacy has on a user’s trust in IDS decisions and backs up existing research. Leading on from this, decision authority was explored and 70% of the participants said they preferred the decision authority to be human based, which corroborates with the immediacy statistics. This was affected by things such as severity of an alert, model confidence, and trust in the AI. When exploring participants with high and somewhat trust in the IDS, common factors were trust in the vendor, understanding the model logic, the level of human interaction, and how well the tool proves to work from the decisions it makes. Figure 2 shows a word cloud generated from high and somewhat trust nodes, demonstrating the distributed weighting of words associated with those themes. This indicates a strong relationship between trust and these factors; where there are more of these factors present, an analyst is more likely to have higher trust in the IDS.

Factors that contribute to the ease-of-use (EOU) of an IDS were also explored. Of the participants that demonstrated a positive perception of EOU, 4 also discussed having perceived high model transparency. This would indicate that where there is higher model transparency within an IDS, users are more likely to find it easier to use. Three of these participants also conveyed high confidence in the decisions made by the IDS. XAI can help increase model transparency, adding further clarity to the explanations provided by the models by highlighting the steps a model has taken

Fig. 2 World cloud—high and somewhat trust



in a decision. For example, pulling information from the model decision tree. This can highlight what caused the model to breach so an analyst can understand the potentially malicious factors. This can in turn increase transparency in the alerts and provide analysts with more detail around what has happened [2]. For example, it could provide information such as the confidence of the model. From the literature review, model confidence can have an impact on trust. When asking participants explicitly whether model confidence would be helpful or contribute to perceived EOU, there were mixed opinions. Most of the participants find it helpful for triaging alerts, however it's not something they solely trust when making decisions as to whether something is a false or true positive.

The availability, perceived usefulness, transparency, and importance of explanations were explored within the interviews, as well as desired types and what potential qualities or factors are missing from current explanations. This was to help understand where XAI could be beneficial and explore what users perceive as helpful. When looking at answers related to the importance of explanations, a key theme was transparency. It was clear from participant's answers that having detailed explanations is important to provide transparency and build trust. Users mentioned that having as much information as possible helps them to make informed decisions, something which is critical when triaging IDS alerts. The word cloud in Fig. 3 is generated from answers relating to important features of explanations. The most common words were information, activity, events, detail, and caused. These indicate the importance of having detailed information within the explanations.

Similarly, when looking at what participants believed to be missing from explanations, there were clear crossovers, again highlighting the importance of contextual information. More than 60% of the participants mentioned a lack of information around why an alert has been triggered, therefore leaving them wanting more context and explainability. This was mentioned by both experienced and less experienced participants. Users also perceived a lack of human understanding, where the IDS

Fig. 3 Word cloud—important features of explanations



is unable to fully understand the nuances of human actions, which leads to a high number of false positives occurring. As a result, analysts need as much contextual information as possible so they can use the information to make informed decisions on the activity seen.

When asked explicitly whether participants think there is a need for more explainability within an IDS, 9/10 answered yes. The participant who answered no is the least experienced out of all 10 participants and demonstrated a lack of AI understanding, which could have influenced their answer. Participants who answered yes were further asked why they think more explainability is needed. The main reasons were to add value, provide helpful information for easier triaging of alerts, and add clarity.

The more information that can be provided within an explanation, the more confidence in the IDS a user will have. This can in turn increase trust. Out of the 6 participants that exhibited high confidence in the IDS decisions, 2 expressed high trust in the IDS, and 3 expressed somewhat trust in the IDS. There is a mix of experience levels in these participants; 2 are less experienced and 3 are more experienced.

On top of more information and explainability, high confidence and trust also came from factors such as the number of alerts seen, vendor reputation, and how the IDS gets implemented and configured in the first place.

Leading on from this, participants were also asked what features of an IDS they believe to be most important. Comparing the distributed weighting of their associated answers, the main themes were User Interface design, Transparency, Integrations, and Customisation. For UI design, many participants highlighted that it needs to be easy to use. Integrations linked into this, with participants expressing that integrations would improve EOU, as well as IDS functionality and value added. Other popular themes included Data, Adaptability, Training, Reliability. A miscellaneous sub-node was included in the codebook in relation to important features of an IDS to account for any answers which didn't fit within the main categories. There were 4 answers coded as miscellaneous, which included features such as filtering, Multi-Factor Authentication (MFA), proactive blocking, and vendor support.

Finally, when asked how important participants feel an IDS is in helping protect an organisation, everybody expressed a need for them and discussed their criticality. In fact, 'critical' was the most used word from answers associated to the importance with participants noting that organisations without an IDS as part of their defence in-depth strategy are much more powerless and vulnerable.

It's clear that IDSs are a critical piece of software within any security team, and therefore it's incredibly important that the people using them to protect our organisations can trust them and use them to their full extent. IDSs need to be easy to use, provide enough information and context for an analyst to be able to investigate alerts efficiently, and display that information in a way that is simple to comprehend.

5 Key Findings and Recommendations

5.1 Key Findings

This research analysis aimed to get a better understanding of what factors influence an analyst's trust and confidence in an IDS, to determine whether XAI can help enhance these factors and improve an analyst's understanding of AI-embedded IDSs. Key findings have helped answer the research questions.

Experience was explored first, and it was unclear whether experience had a direct effect on analysts' confidence levels. However, there were higher levels of trust amongst the more experienced participants, which would indicate that experience can help build trust. When looking at participant's understanding of AI, the more experienced groups had higher levels of AI understanding. Therefore, with more experience an analyst's AI understanding can increase, which can positively influence their confidence and trust [2]. When exploring participants with a lack of AI understanding this seemed mainly due to them perceiving it as something outside of the scope of their role and a lack of available information to help them build their knowledge of it. When looking at participants with some AI understanding they also demonstrated a positive perception of the reliability of the IDS. Therefore, helping to build a user's knowledge and understanding of AI could help increase their understanding of the reliability of the IDS [2] and avoid uncertainties [11].

Where users perceived positive reliability in the IDS, they also showed high confidence levels in the decisions made by the IDS. This indicates that improving the reliability of IDSs through the integration of XAI, can increase the confidence users have in the tool and the decisions being made by the AI. Increasing this confidence is vital in improving the relationship between humans and AI-embedded IDSs [25].

Trust was a big area to explore and analyse, and there were a few key factors that seemed to influence trust. Firstly, immediacy and the level of human interaction played a big part in participant's trust [14], with the majority of participants preferring the decision authority to be human based. This was particularly significant for alerts of higher severity [12]. Other factors that were significant to trust were reputability of the vendor themselves. This is particularly important for software such as IDSs, which are implemented to provide protection. User's need to be able to trust the vendor in order to build initial trust in their product [26]. Understanding of model logic and having model transparency was also prominent in factors of trust [12]. Analysts rely on being able to understand the outcome of an IDS to make accurate decisions, therefore having transparency in the models and their logic through explainable models can provide that understanding and increase trust [2]. The number of blocks seen within the IDS was indicated as a factor which helps users have faith and trust that the IDS is performing well. Where users perceived high model transparency, they also perceived positive ease-of-use of the IDS, as well as high confidence levels in the IDS, which can help increase trust [11]. Model confidence was also something participants said would be helpful for triaging but wasn't something they solely rely on.

Explanations were explored to find out what participants perceived as important and missing from current explanations. Key findings for importance were providing context and being transparent. XAI can provide context through detailed explanations, which help to provide transparency, increase a user's understanding, and enable them to make informed decisions and prioritise critical alerts; giving analysts the confidence that they are taking the correct actions [2]. When exploring what's missing, users noted explanations aren't detailed enough, contextual information is missing, and there is a lack of understanding of human patterns from the AI. When asked if participants believe there needs to be more explainability within IDSs, almost all said yes. They believed this would add value, help their triaging process, and add clarity. Explanations are more valuable if they are clearer [19].

Factors that affect confidence in an IDS were explored. Similar to trust, a key theme here was having more information in explanations. This would increase an analyst's confidence, and in turn their trust, by providing human comprehensible information through more explainable models [2]. How the IDS is implemented on the network also influenced confidence, as well as the number of alerts seen.

Finally, when exploring important features of an IDS in general, the user-interface was key. An easy to use and simple to navigate UI was essential for participants. Transparency and the ability to integrate with investigation tools were two other key themes. Amongst the rest were adaptability and reliability.

Every participant believed an IDS is a critical tool for an organisation. Therefore, helping to increase trust and understanding is essential. If users can't aid their investigation and triage alerts because they don't have enough context, or they don't trust the tool, then critical detections could be missed or not dealt with in the correct way. This could in turn lead to bigger problems, such as critical incidents or major breaches [2].

It was hypothesised that XAI could improve the relationship between human analysts and AI-embedded IDSs, in turn improving the usability of such tools by increasing trust through more transparent, reliable explanations. I believe key findings from this research concur with this statement (Table 1).

5.2 Recommendations

It is recommended that XAI models are used within AI-embedded IDSs to improve usability, understanding, and trust. Improvements to an IDS user-interface (UI) are also recommended through a PoC design, which will include desired components and overcome challenges highlighted from key findings in interview data. This research focuses on a user-centric approach, and therefore these recommendations have been designed and assessed with the user's needs in mind.

XAI PoC Model recommendation. There are many models that can be used to achieve explainability, however, to achieve the aims of this research study, it is recommended that model-agnostic methods are used to provide local explainability. Using model-agnostic approaches means that the explainable AI methods can be used

Table 1 Summary of key findings against the research questions

Research questions	Key findings
What makes a good explanation in the context of intrusion detection alerts?	<ul style="list-style-type: none"> • Contextual information around why the model has breached, and what that means • Transparent explanations • Detailed information
Can XAI improve the useability for security analysts when triaging IDS alerts?	<ul style="list-style-type: none"> • XAI can help increase model transparency, improve explanations provided by the AI, and add more valuable information, which will all in turn increase the EOU
How can we build trust and provide transparency in the decisions made by IDSs?	<ul style="list-style-type: none"> • Increasing confidence by adding more valuable explanations can in turn increase trust • Increasing AI understanding will help to increase reliability and therefore confidence and trust • Providing more detailed contextual information within alerts will increase transparency around the decisions the IDS has made, by explaining elements such as why the model has breached, what features of the activity detected made it trigger an alert, etc.

after a model has already been trained and can work with any AI model [15]. This means that IDS developers can use their pre-existing models and then advance them through model-agnostic approaches post-hoc, to add a layer of explainability on top. Post-hoc methods have been found to be more appropriate for intrusion detection data [2]. Local explainability aims to explain specific predictions, which combined with a model-agnostic XAI approach, will be able to explain the existing predictions of the IDS [15]. This approach will achieve the aims of this research; to aid investigation and enhance an analyst’s ability to understand why and how decisions have been made by the IDS, and what factors contributed to the decisions and actions taken, through clear, reliable, and transparent explanations.

Key findings from this research study present the following key requirements for what the XAI model is expected to achieve:

- Provide contextual Information.
- Provide details as to why an alert has been picked up and why the model breached.
- Produce transparent explanations.
- Clearly detail suspicious factors from the alert.
- Highlight the decision process, by providing key decision information.

It is recommended to use explainable methods, such as SHapley Additive exPlanations (SHAP), on decision tree-based models to achieve the outlined requirements. SHAP is a local, model-agnostic, approach which intends to explain the outcome predictions of a model by looking at the contributions of the individual features [27]. This can highlight which nodes in the decision tree factored into the final prediction of the model, providing explanations which highlight the reasons why a decision was

made [1]. An example decision tree is shown in Fig. 4, by R. Staudemeyer, where the answer is either it was (YES) or wasn't (NO) an attack. If the final prediction for this example was 'YES' (it was an attack) and the decision path was as highlighted, then SHAP would be able to highlight which of those factors from the path contributed to the final prediction and provide an explanation which conveys that information.

SHAP has been successfully tested and used by many researchers to achieve model explainability and would be a suitable approach to meet the aims of this research. However, it is important to be conscious about the limitations which may exist for this method to ensure it can be implemented in a way that works for the individual tool being developed. For example, SHAP is dependent on the model it is explaining because it works by explaining how important the features were to that model. Therefore, if the original model is incorrectly developed or trained, then this will negatively impact the results SHAP outputs [28]. Training is an essential part of model development and should not be overlooked when introducing XAI methods.

UI PoC Recommendation. Based on this study's research and the key findings from interview answers, the following recommendations have been made for UI improvements (Table 2).

The recommended UI design has been split into three pages to demonstrate the recommendations fully.

Figure 5 demonstrates the dashboard homepage. This includes an overview of the connected devices on the network and their locations through a world map visual. Insight graphs show the number of alerts, breaches, and blocks that have occurred over a set timeframe.

A filter feature has been included to allow for adjusting the timeframe on the dashboard. On the right-hand side, there is a visual representation of the types of MITRE ATT&K techniques that have been detected on the network, within the filtered timeframe. As demonstrated within the image, these can be interacted with so that clicking on the number within one of the attack types, will spin to show a breakdown of that figure by the number of breaches, blocks, or critical events

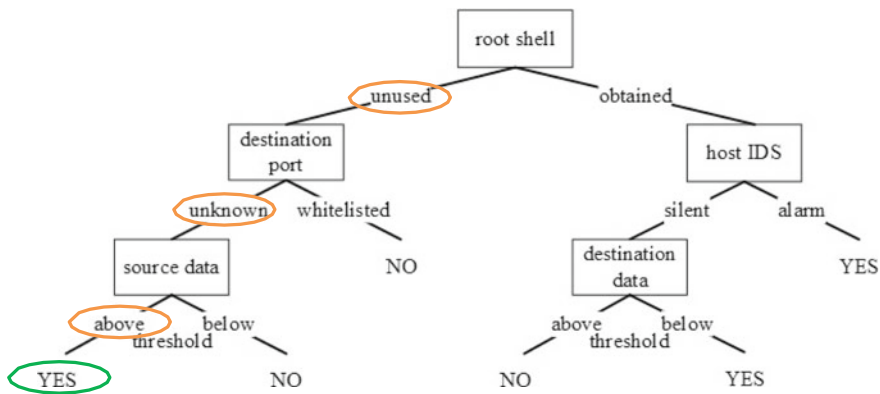


Fig. 4 Example decision tree

Table 2 UI improvement recommendations

Key requirements	Recommended UI improvements
Simple and easy to navigate	<ol style="list-style-type: none"> 1. Clean and simple UI, with search bar navigation from the homepage and filters available 2. Avoid confusing visuals, or too many pop-ups
Available con- textual information	<ol style="list-style-type: none"> 3. More information within the explanations through use of XAI
Visual indicators for alerts	<ol style="list-style-type: none"> 4. Explanations around why the model breached, the model logic, threshold scores, etc. 5. Traffic light system for severity to help with prioritisation 6. Graphs for useful insights 7. Visual representation of activity flow
Ability to correlate events	<ol style="list-style-type: none"> 8. Button to link correlated alerts together. When looking at a linked alert, highlight related events and have clear indications of where they relate
Dashboards	<ol style="list-style-type: none"> 9. Homepage with a dashboard for useful insights on the threats seen, alerts, blocks, and detections
Useful features for easier triaging of alerts	<ol style="list-style-type: none"> 10. Ability to mark alerts as false positives/true positives 11. DNS response feature to provide information on domains and suspicious IPs
Integrations and interoperability with other vendors	<ol style="list-style-type: none"> 12. Ability to link to a ticketing system 13. Integrate with OSINT tooling, like domain scanners, for investigation aids and have the ability within the UI to link out to the third-party 14. Integrate with threat intelligence feeds to provide useful insights

associated to that attack type. This is a feature that many participants mentioned, and something that will add context to the alerts. This homepage is clean, simple, and provides useful insights at a glance, which would be beneficial for the analysts. This could be expanded further to include regional filtering on the map.

Figure 6 shows the recommended alerts overview page, which includes details of open alerts and acknowledged alerts, within a timeframe specified within the available filter. A traffic light system gives visual representation of the severity.

On the right-hand side is a priority matrix, which would provide analysts with a visual representation of the impact and severity of each alert so they can easily prioritise events at a glance. This page is clean and simple, and avoids an overload of irrelevant information, allowing analysts to prioritise their work conveniently and with ease. This could be developed further to allow organisations to decide which tables they want to view an overview for. For example, instead of acknowledged alerts, a table for high priority ones can be included instead.



Fig. 5 PoC UI homepage dashboard

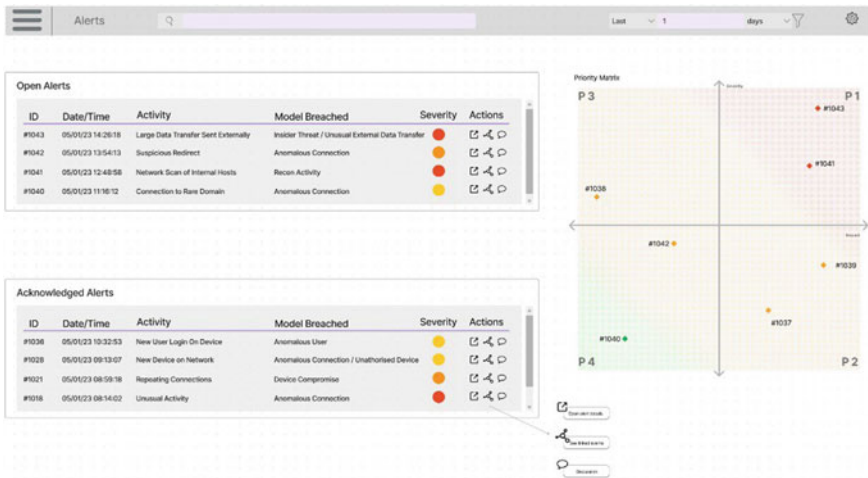


Fig. 6 PoC UI alerts overview page

Figure 7 is the recommended page for alert triaging. It demonstrates what it would look like when an alert is opened for triage/investigation. It is recommended that each element can be opened as a new window within the UI, as demonstrated, and can be closed or opened as needed to avoid busy visuals.

On-top of the desired features from key findings, the ability to collaborate with other analysts has been included. This would allow analysts to see if outcomes/actions are agreed. The relevant information is presented in a simple format, highlighting key details. The use of integrated features allows analysts to utilise tools and aids from within the IDS that will enhance their investigation. An addition that

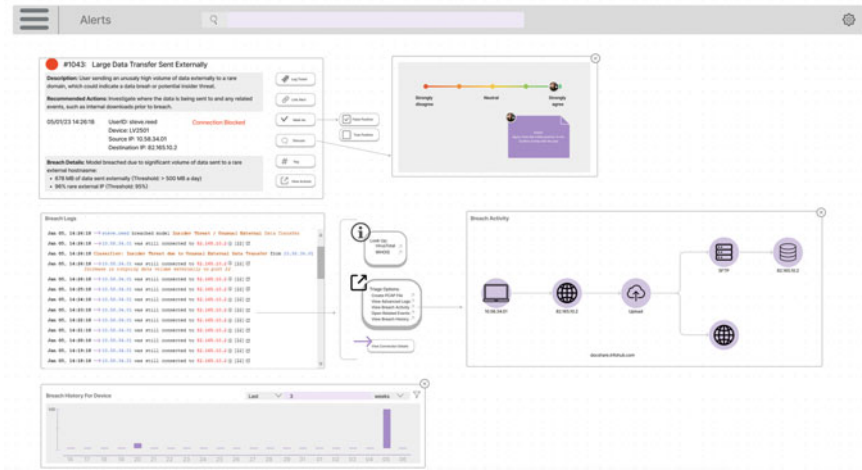


Fig. 7 PoC UI alerts triage page

could be included in further development is including a playbook of recommended investigation steps so less experienced analysts feel more comfortable when triaging alerts.

It is recommended to include these features when designing an XAI embedded IDS, to enhance the ease-of-use and allow for features to be included which complement the enhancements provided through XAI. A well-designed UI offers users simple usability through clean, engaging visuals that allow for easy navigation, allowing users the ability to do their job with convenience. This is essential for building trust and reputation in a product [29].

6 Conclusions and Future Developments

6.1 Conclusion

This research study aimed to explore the benefits of model explainability within IDSs, using a user-centric approach to investigate how it can be used to enhance an analyst's trust in the technology, as well as enhance their triaging capabilities. Based on research of existing methods and key findings from the qualitative analysis of interview data, it can be concluded that enhancing the explainability of IDSs can have a positive impact on an analyst's understanding and help build trust between humans and AI-embedded technologies. Explainability provides transparent, human comprehensible, explanations that provide analysts with the required information to help them understand what activity has happened, and why the IDS has alerted them to it. It is evident from this research's findings that understanding the decision process

of the AI helps to build a user's trust in the product and gives them confidence that the IDS has made a correct decision.

This research aimed to explore appropriate XAI methods and was able to recommend an approach for implementing such methods on intrusion models to achieve the desired factors that were found to help build trust. Unfortunately, the limitations of this study meant that a PoC model couldn't be tested on intrusion data to review the outcomes compared with non-XAI models. However, from research it is clear that the outcomes from XAI explanations would align with the key findings in this research study. A UI PoC was built to demonstrate how IDS technologies can be innovated through the use of XAI and increase the ease-of-use for users. Feedback obtained on this PoC would conclude that this was a successful approach and was able to provide analysts with the correct level of detail, present the information in a user-friendly manner that enhances ease-of-use, and in turn positively affect their triaging capabilities.

Through qualitative analysis, this research has been able to answer the research questions. It can be concluded from this that providing better explanations, improving useability, providing transparency, and building trust can all be achieved through the use of XAI. This in turn can positively impact the adoption of AI-embedded IDSs, by improving the relationship between human analysts and AI, as hypothesised. The importance of XAI has been successfully demonstrated through this research and has therefore made a significant contribution to research in this field.

This research has successfully achieved the aims and objectives, and presented an approach for not only enhancing the technology but also a way in which we can analyse and improve trust.

6.2 Future Developments

It would be beneficial to expand this project further and develop a test environment to run an XAI model on intrusion data to monitor the outcomes and demonstrate the differences that it can make. This could then further be expanded to build a UI prototype that uses the XAI model and feeds in real intrusion data to create a demo environment in which tests could be run and participants could be invited to use.

It would also be beneficial to increase the number of interview participants for this research, and ensure they have a mix of experience across different IDS tools, to allow for more in-depth insights and analysis. These participants can then be split into smaller groups, between those with high confidence in the AI-IDS and those who are not so confident, to enable further analysis on how trust factors differ between those groups. It would also be of interest to delve deeper into the ethical considerations around AI-embedded technologies. This could explore how unconscious bias is embedded at the core of these technologies, and the effect this has on society and the way in which we as humans adopt and trust such technologies.

This project has provided a great foundation for exploring XAI within IDSs and can be expanded to further develop and contribute to this area of research.

6.3 Ethical Conclusions

High standards of ethical practice were upheld throughout this research, following core ethical concepts, including honesty, transparency, non-discrimination, fairness, and informed consent. No PII has been collected from participating individuals, therefore there are no concerns for their data protection. All participating parties had the option to pull out of the research at any time and were invited to review the findings to confirm their answers had been used fairly and with confidence. Professional research practices have been followed, as well as common practices relating to AI, including using a human-centered approach. This research intends to positively contribute to existing research in the area of XAI and aims to enhance studies in the field of intrusion detection.

References

1. Wang M, Zheng K, Yang Y, Wang X (2020) An explainable machine learning framework for intrusion detection systems. *IEEE Access* 8:73127–73141
2. Neupane S, Ables J, Anderson W, Mittal S, Rahimi S, Banicescu I, Seale M (2022) Explainable intrusion detection systems (X-IDS): a survey of current methods, challenges, and opportunities. *arXiv*, Ithaca, NY
3. Kelley K, George B, How to solve the Black Box AI problem through transparency, 16 August 2021. <https://www.techtarget.com/searchenterpriseai/feature/How-to-solve-the-black-box-AI-problem-through-transparency>
4. Brooks C, Alarming cyber statistics for mid-year 2022 that you need to know, 3 June 2022. <https://www.forbes.com/sites/chuckbrooks/2022/06/03/alarming-cyber-statistics-for-mid-year-2022-that-you-need-to-know/?sh=174b40547864>
5. Kleinman L, Cyberattacks: just how sophisticated have they become?, 3 November 2020. <https://www.forbes.com/sites/forbestechcouncil/2020/11/03/cyberattacks-just-how-sophisticated-have-they-become/?sh=5eaa9bc44c3e>
6. Das A, Rad P (2020) Opportunities and challenges in explainable artificial intelligence (XAI): a survey, Ithaca. *ArXiv*, NY
7. Mahbooba B, Timilsina M, Sahal R, Serrano M (2021) Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, p 11
8. Darktrace (2022) 5 AI and cybersecurity predictions for 2022
9. Darktrace (2022) Darktrace AI: combining unsupervised and supervised machine learning [White paper]. <https://darktrace.com/resources>
10. Brown S, Machine learning, explained, 21 April 2021. <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>
11. Choung H, David P, Ross A (2022) Trust in AI and its role in the acceptance of AI technologies. *Int J Human-Comput Interact* 1–13
12. Ashoori M, Weisz J (2019) In AI we trust? Factors that influence trustworthiness of ai-infused decision-making processes, *arXiv*
13. McKendrick J, AI adoption skyrocketed over the last 18 months, 27 September 2021. <https://hbr.org/2021/09/ai-adoption-skyrocketed-over-the-last-18-months>
14. Glikson E, Woolley A (2020) Human trust in artificial intelligence: review of empirical research. *Acad Manage Ann* 627–660
15. Molnar C (2022) *Interpretable machine learning*. Independent, Munich

16. CheckPoint (2022) Check point 2022 cyber security report. CheckPoint, Tel Aviv- Yafo
17. Khraisat A, Gondal I, Vamplew P, Kamruzzaman J (2019) Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity* 1–22
18. Barnard P, Marchetti N, DaSilva L (2022) Robust network intrusion detection through explainable artificial intelligence (XAI). *IEEE Netw Lett*
19. Colaner N (2022) Is explainable artificial intelligence intrinsically valuable?. *AI and Society* 231–238
20. Sutton D, Deep learning and the new frontiers of model explainability, 15 November 2021. <https://www.featurespace.com/newsroom/deep-learning-and-the-new-frontiers-of-model-explainability>
21. Perez I, Skalski P, Barns-Graham A, Wong J, Sutton D (2022) Attribution of predictive uncertainties in classification models. In: *Proceedings of the thirty-eighth conference on uncertainty in artificial intelligence*, pp 1582–1591
22. Nickerson C, Interpretivism paradigm & research philosophy, 5 April 2022. <https://simplysociology.com/interpretivism-paradigm.html>
23. McCombes S, Descriptive research design|definition, methods & examples, 5 May 2022. <https://www.scribbr.co.uk/research-methods/descriptive-research-design/>
24. Fujs D, Anže Mihelič SV (2019) The power of interpretation: qualitative methods in cybersecurity research. In: *ARES '19: proceedings of the 14th international conference on availability, reliability and security*, pp 1–10
25. Gillath O, Ai T, Branicky M, Keshmiri S, Davison R, Spaulding R (2021) Attachment and trust in artificial intelligence. *Computers in Human Behavior*
26. Nayyar S, Why you need to build trust with your security vendor before signing the purchase order, 8 September 2022. <https://www.forbes.com/sites/forbestechcouncil/2022/09/08/why-you-need-to-build-trust-with-your-security-vendor-before-signing-the-purchase-order/>
27. Sukumar R, SHAP Part 3: Tree SHAP, 30 March 2020. <https://medium.com/analytics-vidhya/shap-part-3-tree-shap-3af9bcd7cd9b>
28. Durgia C, Using SHAP for explainability—understand these limitations first, 31 December 2021. <https://towardsdatascience.com/using-shap-for-explainability-understand-these-limitations-first-1bed91c9d21>
29. Vee A, Importance of a good User Interface (UI), 27 February 2020. <https://www.linkedin.com/pulse/importance-good-user-interface-ui-anna-v/>

Could Adam Smith Live in a Smart City?



Ian Toft and Tasmina Islam 

Abstract In this paper we look at how Smart Cities can accelerate trends in society initiated by the information economy that run counter to traditional western norms. These include the right to privacy, reliance on trust (whether personal or institutional) and emphasis on individual rights, potentially going as far as weakening legal principles such as the presumption of innocence. We investigate the contribution of behavioral science in reinforcing these trends. We look at how such societal norms are transmitted through both economic and ethical frameworks such that the transparency championed by the information economy affects everything from the wealth divide to our attitude toward consequentialist and data-driven decision making. We see that transparency and surveillance can disrupt the process by which we build trust, learn right from wrong and develop as individuals who can then contribute to a pluralistic society. We explain the existence of feedback loops that exacerbate the replacement of trust with assurance and look at the potential for misuse of surveillance, particularly when combined with social media and behavioral science. We conclude that making such rapid changes to our societal norms risks upsetting the cooperative balance we have evolved between the needs of the individual and the collective and that we need to slow down in order to allow a more extensive public discussion about how this technology is implemented and whether current legislation is adequately protecting the principles that are most important to us.

Keywords Smart city · Smart cities · Ethics · Information economy · Transparency · Privacy · Decision making · Consequentialism

I. Toft (✉) · T. Islam
Department of Informatics, King's College London, London, UK
e-mail: ian.toft@kcl.ac.uk

T. Islam
e-mail: tasmina.islam@kcl.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_18

343

1 Introduction

The central idea of this paper is that smart cities are a manifestation of a greater phenomenon in which big data, AI and digital surveillance are diminishing the role of the individual as the core unit in western society. In fact, there has been a coincidence of timing, in which developments in neuroscience and psychology are undermining the idea of the individual as a rational decision-making entity at the same time as the information technology ecosystem has reached a point where it is able to take over the roles relinquished by responsible individuals. The purpose of the paper is to present ideas from a range of disciplines in order to provide a technical audience with hopefully a wider appreciation of how smart cities and the information economy in general interact with principles underpinning our societal structure. In fact, the information economy introduces some inherent incompatibilities with fundamental western ethical, legal and philosophical positions. This unresolved tension will surely lead to trouble if not addressed in the public arena. Therefore the second purpose of the paper is to encourage wider public engagement and suggest a slower and more careful pace of introduction, particularly in order to provide enough time for our legal framework to adapt to the new technology.

Traditional capitalist economies are centered around the individual as the main responsible decision maker both in an economic and ethical sense. In Adam Smith's 'wealth of nations' he considers that it is the cumulative effect of free individuals, each acting in their own best interests that produces the optimum division of resources and labour and is most beneficial to society as a whole [1]. Given that the individual is the most important decision maker, this economic system exists in symbiosis with ethical frameworks that provide each individual with personal understanding of how they should act in society, whether that be via deontological rules or virtue ethics. It also encourages a legal system that enshrines individual rights such as free speech and privacy as core legal principles. The information economy in general and smart cities in particular, have a radically different underlying philosophy which we will explore throughout the paper. Transparency is valued over privacy. Economic efficiency is derived by algorithms processing vast amounts of data rather than by cumulative individual preference. The ethical framework of consequentialism, which in any case doesn't purport to provide a basis on which to make real-time individual decisions [2], sits better in this system [3] than the individual guidance provided by deontology or virtue ethics.

Despite operating under such profoundly different principles, the infrastructure for smart cities and digital economies is being rolled out with inadequate public consultation [4], often in the form of public-private partnerships [3], with private partners standing to make large profits as a result [4]. This ought to be ringing a few alarm bells and indeed the Canadian Civil Liberties Association has sued the Canadian government for "outsourcing government responsibilities to a private corporation" in relation to the Toronto smart city project [3]. Of course, the question of whether smart cities should be constructed and run by public or private entities is distinct from the question of whether they are philosophically a good idea in the first

place, but the Canadian example highlights how these projects are being implemented too quickly (particularly considering a lack of robust legal structure for the handling of personal data [5]) and with little public awareness.

Smart cities inherently require a multidisciplinary approach to understand how they will impact society. We cannot focus just on the technological details, or we will miss the potentially more important sociological and political ones. In this paper we present ideas from neuroscience and behavioral science in Sect. 2, ethical frameworks in Sect. 3, the impact of big data on modelling and data driven decision-making in Sect. 4 and the impact of transparency on societal trust in Sect. 5, before concluding in Sect. 6.

2 Neuroscience: Are We Capable of Acting in Our Own Best Interests?

It is necessary to describe briefly some of the results from the field of neuroscience in order to understand what is currently happening to the concept of individual responsibility. According to Camerer et al. [6], our actions are the result of competing processes in the brain, some controlled and some automatic, some cognitive and some driven by emotional systems. More than this, we often incorrectly attribute decisions to cognitive processes when in reality they were subconsciously executed. As an example, people tend to agree more with an editorial if they listen while nodding their head as opposed to while shaking it. The brain in this case is reconciling available information coming from different internal systems, one of which is driven by the movement of the head, even though this is arbitrary in this case. Other experiments show evidence of self-manipulation, for example when people cheat during their own medical tests, suggesting that a cognitive wish to get an accurate result is being overridden by other mental processes, the net result being a situation that might not be in the individuals' best interests at all. Because people lack access to the systems that produce such internal contradictions and biases, it is hard to correct for it or even know that correction is needed [6].

The ability to exert conscious self-control also seems to be a limited resource. Experiments show that participants asked to remember a large number of digits exhibited lower self-control when subsequently offered either a healthy or unhealthy snack. Other research suggests that decision making is often an exercise in pattern matching rather than careful deliberation. Neuroscientists therefore consider preferences to be "transient state variables that ensure survival and reproduction" [6], rather than the traditional economic understanding of preferences as a stable construct that maximize current and future wellbeing. Another discovery undermining our sense of cognitive control is the fact that electroencephalogram (EEG) readings show that brain activity occurs 300 ms before we are aware of an intention to perform an action. We attribute the action to our own intention, but in fact both are preceded by activity we are unaware of [6].

Given that decisions are the result of competing internal processes, many of which are inaccessible to our conscious mind and which start before we are even aware of any intention to act, leads to questioning the idea of there being a decision making entity that we can call an individual at all, and certainly not one that can be relied upon to act in a responsible manner with a view to overall long term best interests [7]. The impact of this thought process on governance is clearly visible and has been present for some time. ‘Paternalistic’ policies to restrict unhealthy foods are an example of government stepping in to act in citizens ‘best interests’ since they can’t be trusted to do it themselves. More than this though, the implication is that the ‘greater good’ is not something that can be discovered by simply asking people what they want or observing their revealed preferences, but is something that must be measured by other means [7]. As Saint-Paul discusses [7], the concept of fundamental human irrationality invites government intervention and observation into all areas of life, including into areas that were once strictly private. He also suggests that if the individual is composed of competing and transient decision-making processes, to what extent can someone be held accountable for past actions?

In Fig. 1, we show the results of a literature search using the ACM digital library for publications containing references to paternalistic government in the last ten years. Many of the results focus specifically on the interaction between government and society which is an interesting development in a primarily computer science oriented library. Other results focus on paternalism in content moderation, identification of misinformation, energy conservation, waste management, smart home technology, behavioral ‘nudging’ and collective decision-making. The results show an increasing interest in these topics amongst computer scientists in recent years.

Following this line of thought leads to the idea that if there is no consistent and rational entity that can fairly be held accountable, then law and order should be based more on prevention than on punishment [7]. Indeed, many would no doubt welcome some kind of surveillance-based system that raises an alert when antisocial behavior crosses some threshold. Big data is already used in predictive policing [8] and pre-emptive solutions to the current knife crime epidemic in London are suggested by Smart et al. [9]. In Fig. 2, we show the annual frequency of literature that contains references to “predictive policing” over the last ten years in the ACM digital library, which indicates an increase in interest particularly in the last five years.

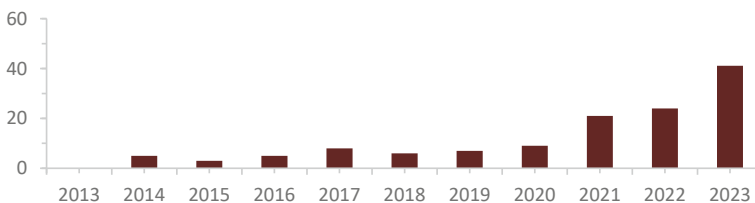


Fig. 1 ACM digital library search for literature containing the terms: paternalistic AND (government OR governance) AND (citizen OR “civil society”), anywhere in the text over the last ten years (with pro-rata adjustment for 2023)

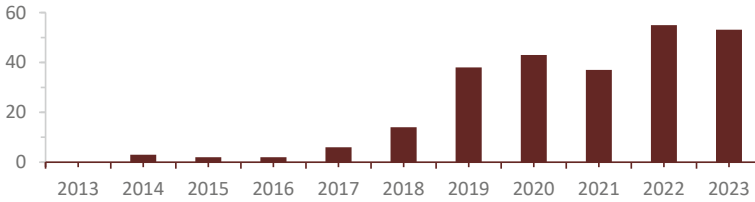


Fig. 2 ACM digital library search for literature containing the term: “predictive policing”, anywhere in the text over the last ten years (with pro-rata adjustment for 2023)

But we don’t know the ultimate effect of changing our policing to be preventative rather than holding people to account after committing an offense. It certainly upends the principle of innocent until proven guilty and goes hand in hand with a surveillance network that would be worrying in the wrong hands. Writing in *The Guardian* after the pre-emptive arrests made during the recent coronation [10], Monbiot suggests that “the police can do what they want to us now”. This is possibly an overreaction, but is indicative of the impact on societal trust that big changes such as a shift from reactive to pre-emptive policing can have. More than this, it starts to interfere with our evolved ethical system of being brought up to follow a code of good behavior by observing examples of good and bad behavior around us as we become responsible members of society. This is the topic of the next section where we look at the main three ethical frameworks and the impact of the information economy upon them.

3 Moral Frameworks

At this point we should remind ourselves that whilst modern techniques have generated new insight into how the brain works, nothing has actually changed about how humans function and have functioned for thousands of years. We have built successful societies despite our cognitive biases and one of the ways we have been able to do this is by constructing frameworks of moral norms that define acceptable behavior and dampen some of our irrational impulses. The three main ethical frameworks are consequentialism, deontology and virtue ethics. It is useful to consider a couple of hypothetical scenarios that highlight the strengths and weaknesses of the different frameworks and how they interact. This thought process leads us to see the value of virtue ethics as an important component of our behavioral systems and it is this framework that is most compromised by the introduction of ubiquitous surveillance, data-driven decision making and pre-emptive policing. We will also highlight some of the weaknesses of the consequentialist system, which is the framework most suited to the moral underpinning of a smart city.

The first scenario is that of the terrorist who has planted a nuclear device and an agent who must decide whether it is morally acceptable to use torture to force the terrorist to disarm the device. The situation is so extreme that the natural reaction

of many would be that it is acceptable in this case even if totally unacceptable in other situations [11]. A strict deontologist on the other hand, perhaps following the tradition of Kant, who said:

“better the whole people should perish than injustice be done”

would take a different view here [11]. In the real world, when situations become extreme, we would apply a different legal and moral framework, either by way of military involvement, or emergency regulation [12, 13]. The moral framework turned to here is consequentialism; the total amount of good or bad is assessed for both options and the decision is made to follow the path that creates the most good or least bad [2].

We can then ask, why is this logic not applied during normal times? In fact, even consequentialists would say that their framework is more appropriately used to retrospectively determine ‘right action’, rather than as a real-time guide to decision making [2]. This is because the world is far too complex for individuals to consider the total impact of their actions, including indirect effects. This analysis changes during situations that are so extreme that it becomes possible for human decision makers to see that one course of action will, with a high degree of certainty, produce a better outcome than another [13].

Strict consequentialism also runs into difficulties when applied to thought experiments such as ‘transplant’ in which a doctor must decide if it is acceptable to transplant organs from one patient without consent into five other patients, thereby saving five lives at the expense of one. A consequentialist has to turn to an indirect form such as rule-based consequentialism to explain why most people find this unacceptable [2, 11]. This modification looks at the consequences of following certain rules and whether these rules lead to optimum outcomes [2]. If we live in a world where we thought doctors might sacrifice us for the greater good, then we may avoid hospitals which overall makes the world less good. In this way, the rules-based consequentialist and the threshold deontologist, whilst approaching from different angles, arrive at a quite similar result: we apply a rules-based approach most of the time (or a virtue ethics approach as we will discuss later) until it becomes very clear that the rules will lead to extremely bad outcomes at which point the framework changes.

If the world is too complex to directly apply consequentialism to decision making, then it seems unlikely we could ever come up with a set of rules that would reliably lead to good consequences either; our deontological rule making will be challenged by the same complexity. In addition, if the outcomes of deontological rules are in any case tested against our pre-existing morals in the real world, then the rules themselves would seem secondary to something else. Instead, we can turn to a much older ethical framework. The virtue ethical solution is to say that we make our decisions by cultivating our virtues and our understanding of how to apply them in the world [14]. A virtue ethicist would say that it is not the “inculcation of rules but...the training of character” that will enable right action [14]. We do this by observing ‘exemplars’ of virtuous behavior who are the people that come closest that humans can, to approximating what Plato would call the ‘form of the good’ and what a religious person might call God [14]. However, according to Zagzebski [15],

“we do not have criteria for goodness in advance of identifying the exemplars of goodness”.

Thus, we learn goodness from examples we see around us, but it is not explicitly and rigidly defined. A less esoteric, evolutionary interpretation of this is that the virtues are the behaviors that we evolved to allow us to cooperate and balance the needs of individuals against the needs of society [16]. Virtue ethics can coexist with a certain amount of consequentialism by specifying that in order to prevent good intentions from resulting in bad outcomes, the application of virtues requires practical wisdom [14]. However, most virtue ethicists would draw the line at thinking that any finite being such as a human can possibly succeed in fully optimizing consequences and the religious among them might consider that this is similar to raising man to a god-like status [14]. We mention the religious angle because it is an important ethical system that has enabled the construction of successful and stable societies, irrespective of whether one believes the details or not. In fact, the religious, philosophical and evolutionary interpretation all support the idea of virtues as behaviors that allow the individual to balance their selfish needs against their collective responsibilities. It is important to note here that if we learn goodness from examples around us without necessarily defining “criteria for goodness”, we risk disturbing this process by introducing constant surveillance and preventative intervention. This then potentially upsets the balance between the individual and society as well as upsetting the process by which we all learn to distinguish good from bad as we grow up.

The other general point to make is about the rate of introduction of new technology relative to the size of its impact on our society. As discussed previously, when we compare the outcomes of moral thought experiments to the real world to decide what is right, we are comparing against our societal moral norms which take time to evolve [16]. Our legal systems in turn evolve relatively slowly. Woods [17] states that:

“Our policy and legal apparatus stabilizes society by setting a standardized expectation and reducing harm. It cannot keep pace with the rate of change of our technology”.

He makes the point that only specialized experts can understand the technology properly, for everyone else it is changing faster than our cultural norms can adapt to it, thus “policies and laws fall out of sync with reality”. Therefore, by rapidly introducing disruptive new technology, we are both weakening our evolved moral frameworks and challenging our legal frameworks whilst turning at speed toward behaviors that are not grounded in our social norms. In other words, the modern world has moved beyond the ability of our legal and ethical frameworks to guide us reliably. In such a situation, the virtue ethical concept of searching for ‘exemplars of goodness’ without explicitly defining them is important, as it allow our frameworks to evolve over time. When the world is changing rapidly, it is tempting to instead be guided by expert led consequentialism, but the disadvantage of this framework is that evolution becomes difficult, since what is right and wrong has been decided in advance. This situation is potentially amplified when combined with surveillance and preventative intervention. It is therefore a recurring theme in this paper that more time must be given for our naturally evolving ethical and legal frameworks to adapt to the new digital world.

4 Big Data Versus the Individual

In the deontological and virtue ethical frameworks, morals are personal. Each person is required to keep their “own agency free from moral taint” [11] with the implication that the morals of society as a whole will then look after themselves. This resonates with the capitalist idea that the cumulative effect of individuals acting in their own best interests is actually beneficial to society as a whole and far more efficient than if a centralized body tries to decide the optimum division of resources and labour [1]. We could then ask the question: is there an information level at which a centralized authority could surpass the efficiency of a capitalist economy? In a similar way, is there an information level at which a group of decision makers could make decisions based on a reasonable prediction of the greater good? In fact, the smart city is attempting to do both of these things; improving efficiency caused by market failures and also directing activity in a beneficial direction, for example in relation to climate change.

Ananny and Crawford [18] look at what it means to understand a complex system. They argue that the Enlightenment concept that

“observation produces insights which create the knowledge required to govern”

is not appropriate when applied to complex systems that do not “contain complexity” they “enact complexity”. You cannot understand elements of a complex system in isolation from their interactions with the rest of the system, thus there is often a confusion between being able to see something and being able to control it. In fact,

“what systems are or mean depend upon the tools and perspectives people employ while looking [18]”.

Big data however addresses this issue to some extent. Rather than working with samples which are subject to biases of collection, one is working with the entirety of the data, essentially looking at the whole of a complex system at once [19]. This removes the task of trying to simplify and model it, which is what proved impossible in the past and led to the reliance on cumulative individual contributions as a means to handle complexity. Kitchin believes however that “data are never raw” and “do not exist independently of the ideas, instruments, practices, contexts, knowledge and systems used to generate and process them” [19]. Even with relatively raw big data, there is still a philosophy behind its widespread gathering, perhaps without specific consent, perhaps with subsequent repurposing and sale for future profiling or behavioral nudging, perhaps with insufficient anonymization [19]. The very existence of this data can then have a subsequent impact on a complex system, for example due to surveillance and a loss of individual privacy which may cause significant and unforeseen changes in behavior. Therefore, the data can never be considered pure and able to “speak for itself”.

In addition, whilst we have huge amounts of data on some things, others remain qualitative or ambiguous and we always will need human decision makers to balance these inputs. These human decision makers have biases in their cognitive processes

just like anyone else. For example, it can be shown experimentally that people are averse to ambiguity as distinct to risk (unknown vs known probability distributions) [20]. Similarly, experiments conducted by Van Dijk et al. [21] show that ambiguous information is discounted during decision making. In these experiments, participants disregarded ambiguous information even though they were told it lay within a range that was shown to change a decision if revealed precisely. In other words, any possible precise value that the ambiguous information could have had would have changed the decision, but the ambiguity alone caused the information to be ignored.

Using the recent pandemic as an example, there were many qualitative and ambiguous areas of importance such as the impact of lockdowns on education or suicide rates or missed cancer diagnoses which are accumulating data in retrospect [22], but at the time were hard to compare to the data rich topics of transmission rates and case numbers. This is not to say that focusing control on what can be easily seen, such as transmission rates means that secondary effects were not understood or predicted to some extent, but just that since they could not be presented in the same data-driven manner, the comparison becomes difficult. Governance itself then becomes increasingly data driven [18] (which has the additional benefit of making the moral implications of decisions seem accountable).

In Fig. 3, we look at the annual frequency of publication in the ACM digital library of work containing references to data-driven decision making together with some mention of ambiguity, qualitative information or bias. It is interesting to compare this to the frequency of publication without selecting for ambiguity and bias etc. Results that satisfy the extra criteria comprise roughly half (minimum 31%, maximum 67%) of the articles without the criteria for any given year. This is of course not to say that these articles are either for or against data-driven decision making, but just that it is a topic of increasing interest and there is at least considerable awareness of the challenges posed by qualitative or ambiguous information.

Collecting results from over 100 studies for a cost-benefit analysis, Allen concludes that “it is possible that lockdown will go down as one of the greatest peacetime policy failures in modern history” [22]. It is likely too early to draw a

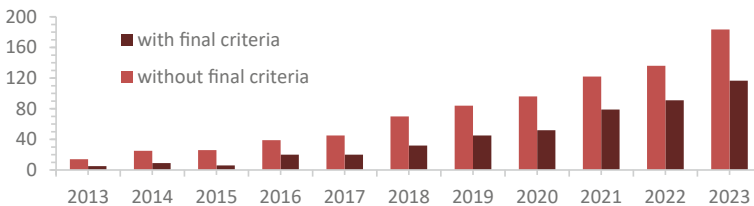


Fig. 3 ACM Digital Library search for literature containing the terms: (“decision making” OR “decision-making” OR “decision maker” OR “decision-maker”) in the abstract together with (“data driven” OR “data-driven”) and either with or without the final criteria: AND (ambiguous OR ambiguity OR qualitative OR bias) anywhere in the text, over the last ten years (with pro-rata adjustment for 2023)

firm conclusion on this, but we can at least question whether confidence in data-driven governance was misplaced during the pandemic and proceed with caution in applying this consequentialist approach elsewhere. In a smart city it seems similarly unlikely that using big data to guide decision-making will account for all the cultural, social and political factors that exist in a city and there seems to be a similar risk that data heavy factors may outweigh ambiguous factors when decisions are being made. Coming back to the example of predictive policing, a smart city will no doubt be able to direct police resources to the statistically most likely areas of trouble, but whilst this may suppress crime in those areas it is not clear whether this data will allow the city to better understand the cause of the problem in the first place. If a problem is not actually solved, then it can result in further unforeseen consequences in the future. In fact, there are many examples in which the outcome of policy ends up being the opposite of that which was intended. For example, recent studies have shown that the introduction of the sharing economy to transport by companies such as Uber can actually increase congestion, depending on the density of the city [23]. This is not to say that data should not play a part in problem solving, but just that it must not be allowed to dominate the decision-making process and that the idea that a large enough quantity of data can replace the need to understand a system should be treated with caution.

5 The Information Economy: Trust Versus Transparency

Bundi and Pattyn [24] have shown that the public attitude toward technical, data-driven decision making depends on the alternatives available. They looked at the correlation between trust in government, trust in other citizens and attitude towards evidence-informed policy making during covid. What they found using data from six western countries was that the lower the public trust in other citizens and in governments, the more positive was the attitude towards evidence driven decision making. Public trust in each other and in political institutions is something that evolves over time and is intertwined with all aspects of how society functions. It rests on expectations of how people will behave due to their ethics, their economic condition, the laws and policies that govern them and many other factors such as interaction frequency and rapid change [25]. The information economy is indeed rapidly changing how people interact, how they are incentivized economically or otherwise and as discussed, is also possibly changing the moral landscape as well in a consequentialist direction. It is therefore not unlikely that core concepts such as societal trust can be undermined or at least altered by this process.

One of the central components of the information economy is the philosophy of transparency. Clearly those involved in the accumulation of data for packaging, profiling and ultimately selling would prefer not to be restricted at each step by stringent privacy laws. But transparency is also being presented as a social good. The CEO of Facebook, Mark Zuckerberg put a moral spin on it when he said:

“the days of you having a different image for your.co-workers and for the other people you know are probably coming to an end pretty quickly. Having two identities for yourself is an example of a lack of integrity [26]”.

Whilst this may not be completely wrong, it is important to note that this kind of transparency is a radical change from what we had before. Zuckerberg is implying that personal exploration and expression should be done in public. Peppet [27], in contrast, states that individuals need to be able to construct a self in privacy, which is also important to the whole society given that a democracy is a collective of individuals with a plurality of ideas. Excessive transparency can discourage honest conversation and debate [18] and homogenize thought and behavior [28]. This is another example of how “information capitalism” with its philosophy of transparency and its consequentialist moral framework is fundamentally different to traditional capitalist societies with their focus on the individual, on privacy and deontology and virtue ethics.

This is not to say that the old industrialists were being purely altruistic. A strong middle class with rising wages was a necessary part of the manufacturing economy [29]; democracy and individual rights were part of this successful and profitable system. In the information economy however, the link between revenues and the public is indirect; profits are achieved by scale and a close to zero marginal cost of expansion. Rather than fostering a robust middle class, the result is a wealth divide that mirrors the information divide; those with access to big data and the means to monetize it accumulate more of the wealth and the power [30].¹ Of course, the information economy is not the only driver of wealth inequality, factors such as globalization are also clearly involved and there are areas in which previously disadvantaged groups can benefit from the digital economy (for example mobile banking). The main point to highlight however, is that the emphasis on transparency over privacy in a digital world, results in a significantly different economic structure, with far reaching effects.

A good example of how transparency starts to impact our societal norms is the remote monitoring of workers to ensure they are doing what they are contractually obliged to do. Dubai airport for example digitally tracks over nine thousand of its workers whilst security firm CityWatcher has offered subcutaneous RFID tags to its employees [27]. In Fig. 4 we look at the annual frequency of literature in the ACM digital library that references workplace or employee surveillance. (It was necessary to include a long list of alternative but specific word pairings here in order to ensure article relevance). Increasing interest in this subject seems to be relatively recent, starting a significant increase in 2021–2022.

This form of assurance is in fact a replacement of trust in responsible individual action. Another example of this is continuous monitoring of vehicle speed, acceleration and braking to qualify for cheaper car insurance [32]. Instead of augmenting the traditional contract however, such assurances transform it from a social arrangement

¹ Public and legal opinion has begun a lagging response to the rapid development of the information economy. The EU in particular has put in significant protection for its citizens with the 2018 GDPR [31].

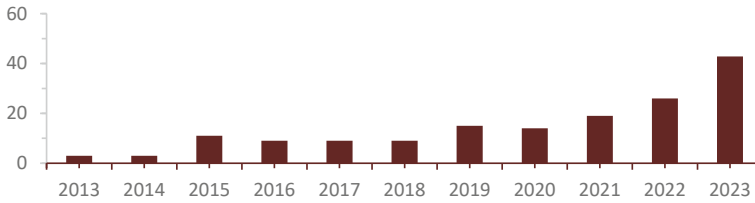


Fig. 4 ACM Digital Library search for literature containing the terms: “workplace surveillance” OR “employee tracking” OR “tracking employees” OR “tracking of employees” OR “monitoring employees” OR “monitoring of employees” OR “employee surveillance” OR “surveillance of employees” OR “tracking workers” OR “tracking of workers” OR “worker surveillance” “surveillance of workers” OR “monitoring workers” OR “monitoring of workers”, anywhere in the text over the last ten years (with pro-rata adjustment for 2023)

into a mechanical process [29]. Contracts are the way in which trust is built in the face of uncertainty. They may not represent the personal level of emotional trust that small groups of people exhibit, but they nevertheless represent a cognitive, system level of trust in institutions [25]. If you remove the opportunity to develop trust, then people start to operate not according to how they know they *should* behave but according to how they *must* behave. Experiments have shown that assurances solve short-term problems at the expense of long-term trust [33]. There therefore seems to be a risk that we have started a self-reinforcing cycle in which societal trust is reduced the more that assurance methods are put into place, but then as trust in each other is reduced, the more people turn to data-driven governance and assurance [24].

Covid lockdowns and contact tracing schemes represent a different example of assurance, providing transparency as to where people are and who they are with. This transparency is useful to decision makers as it allows modelling and control of transmission rates, but it is also useful to the members of society who feel that they don’t trust other members to behave in a sensible way regarding transmission. But if people are increasingly behaving according to automated or legislated punishment and reward incentives then it becomes difficult to either witness and develop virtues or pass on evolved behaviors such as trust. It becomes hard to break the cycle of decreasing public trust and increasing technical assurance. It is worth considering where this cycle ends; something similar to the Chinese ‘social credit’ pilot schemes,² in which people take behavioral cues from a digital surveillance system seems a possible destination [36]. The impact of doing this for a prolonged period of time is unknown; it is the governing according to a shared set of societal values that provides authority with a sense of legitimacy, losing this leaves behind only the mechanical aspects of power [29].

There is of course another key benefit of the traditional individual centered, privacy focused society which we have not yet touched upon, which is resistance to tyrannical overreach. Ideological transparency and widespread surveillance in the wrong

² The Chinese authorities have more recently limited the scope of such schemes [34], for example pledging to ban AI in social scoring systems [35].

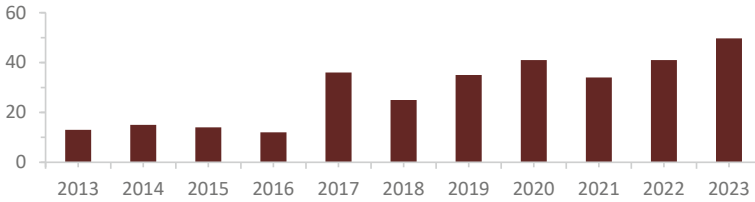


Fig. 5 ACM digital library search for literature containing the terms: “social media” AND (“political campaign” OR “election campaign” OR “electoral campaign” OR “political advertising”), anywhere in the text over the last ten years (with pro-rata adjustment for 2023)

hands is clearly a frightening concept and in fact there have already been misuses, particularly of social media, that should be making us uncomfortable. As advertising on digital platforms increased, it became clear that the response of the public could be measured, and real-time experiments could be done as to what made the advertising effective [29]. It was then a small step further to consider how political messaging can influence behavior; the 2012 Obama campaign and the scandal involving Cambridge Analytica during the 2016 Trump campaign, both used Facebook data and messaging to influence voters [37, 38]. This actually represents a reduction in transparency, since an activity that was previously conducted in the public sphere, that of trying to influence voters, can now be done out of public view. Again, the information economy is turning on its head what went before; that which was public is now private and that which was private is now public. Figure 5 shows the annual frequency of reference in the ACM digital library to political campaigning in combination with social media over the last ten years. It shows a significant increase since 2016.

Perhaps even more questionable was the Facebook experiment that investigated to what extent emotional contagion could spread through social networks by increasing sad posts in some peoples feeds and happy posts in others [38]. We are now in highly dubious ethical territory and indeed bringing in the other ideas from behavioral science, the full potential of the information economy to be a tool of social or political persuasion starts to be revealed. Tufekci uses the term “computational politics” [38] which he says is:

“informed by behavioral sciences..refined using experiments..and is often used to profile people..and to develop methods of persuasion and mobilization which, too, can be individualized”.

This is individually targeted messaging based on psychological profiling with real-time feedback on its impact.

Behavioral techniques are not just used in experiments or for political influencing. They were also turned to by governments during Covid; in the UK, a sub-group of SAGE called the ‘Scientific Pandemic Insights Group on Behaviors’ recorded in the minutes of their 22 March 2020 meeting [39] that:

“the perceived level of personal threat needs to be increased among those who are complacent, using hard hitting emotional messaging”. “Messaging needs to take account of the different motivational levers and circumstances of different people”.

Of course, use of behavioral techniques by government during an emergency is a different prospect to the use by corporations in normal times, but the ethical appropriateness of these techniques still needs to be questioned. Andric et al. [40] argue that “governmental online targeting” is inconsistent with “fundamental democratic doctrines and values”. This is because of a lack of transparency, given that in general only the targeted group will see communication that is aimed at them, making it unavailable for public debate. There are also issues with privacy violation and potential discrimination due to algorithmic profiling. The authors argue that the “governments collection and analysis of individuals’ data gives them a greater ability to influence and control their citizens” than is “desirable in a democratic society” and “raises questions about who has the authority to shape society and determine its priorities”.

As discussed in section two, behavioral science no longer considers the human as a rational being but as a potentially irrational being who can be induced into different behaviors by the right cues [38]. The inducing of behavior that an individual would not arrive at rationally, in order to achieve the goal of the greater good, is highly consequentialist in nature. Increased use of ‘nudge’ techniques goes hand in hand with consequentialist ethics and clearly becomes more powerful the more we are monitored and the more time we spend using digital media. In figures six and seven, we look at the annual frequency of publication in the ACM digital library of work containing references to nudging and psychology, in combination with either social media (Fig. 6) or surveillance (Fig. 7). The purpose of splitting the search like this is that there are two parts to psychological nudging, which are now able to work together in real time: the presentation of the nudge via social media and then the surveillance of its effectiveness, for example via sensors in a smart city.

Smart cities thus potentially provide an ideal test bed for these types of techniques and experiments due to the rapid cycle of optimization that would be possible in such an environment. It would seem necessary to evaluate an appropriate ethical framework for this and obtain meaningful consent from citizens, as well as a means of declining consent that doesn’t require moving somewhere else. An independent regulatory body similar to that recommended by Andric et al. in the context of governmental online targeting [40], would seem to be a minimum requirement. In Josh O’Kane’s book “Sideways: The City Google Couldn’t Buy”, the author describes

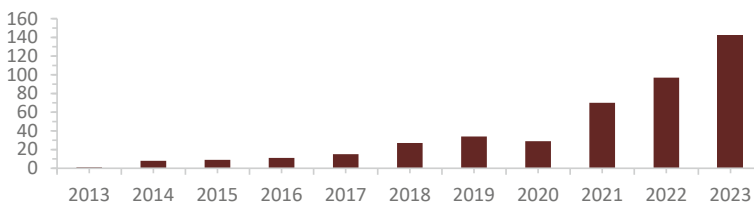


Fig. 6 ACM Digital Library search for literature containing the terms: nudge AND (psychometric OR psychology OR psychological) AND (“digital media” OR “social media”), anywhere in the text over the last ten years (with pro-rata adjustment for 2023)

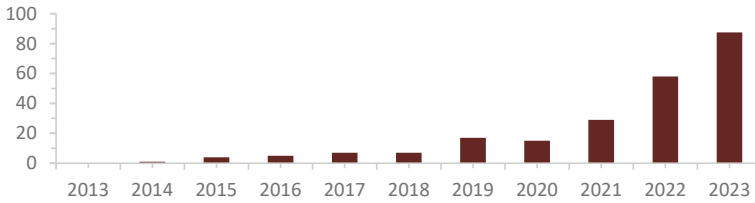


Fig. 7 ACM digital library search for literature containing the terms: nudge AND (psychometric OR psychology OR psychological) AND surveillance AND (NOT “VR”) anywhere in the text over the last ten years (with pro-rata adjustment for 2023). Final exclusion included in order to remove a large number of references to a single book of limited relevance in 2018

the controversy that built up around Toronto’s smart city project as being caused, in part, by the inadequacy of legal structure around data and privacy in Canadian law at that time. The Alphabet subsidiary Sidewalk Labs was suggesting the creation of a new category of data they referred to as “Urban Data” which is not a category recognized under Canadian law. In that sense the organizations making plans for Toronto were operating in a policy vacuum which should have been recognized and addressed by Government. One of the central recommendations of this paper therefore is to hold back on further roll out of this technology until a suitably robust legal and ethical framework has been built for it by lawmakers and lawyers in consultation with the public. This is surely necessary in order to protect the public from psychological manipulation and experimentation and to define clearly what is and what is not acceptable when it comes to ‘nudging’ and online targeting.

6 Smart Cities: The Conclusion?

The Smart City is a melting pot and magnifier of everything discussed so far. The type of real-time experiments discussed in the previous section could become routine as the amount of data collected on everything and everyone increases. Of course, much of this would have purely functional goals such as minimizing electricity and water waste and increasing transport efficiency. But even the functional aspects of this process should be questioned if they are taken to the extreme. In that case we would effectively be turning the economy of a city from being something like a human machine to being a data-driven machine, optimized not by the individuals who live there through their collective activity, but by algorithms designed to optimize what is happening according to a goal that has not yet been publicly discussed and agreed upon. From Amsterdam and London to Singapore and Yokohama, policy makers have laid out goals for their respective smart cities, ranging from carbon reduction and energy efficiency, pollution and congestion reduction, development of IT infrastructure and IT know-how, traffic and transportation efficiency, health services, smart economy investment and public behavior modification [41]. Dubai’s

smart city has an overarching “happiness agenda” which aims to “equip policy makers with the data and tools to evaluate peoples’ happiness in real time and empower a data-driven movement towards achieving better quality of life” [42]. These are all commendable aims. The concept of quantifying and monitoring happiness in real time is perhaps one of the more ambitious and experimental goals and would surely be highly susceptible to data-driven factors outweighing qualitative input. But even the simpler goals, must be held up to genuine public debate and scrutiny. The recent attempt by five councils to block the expansion of the ultra-low emission zone in London [43] is indicative of a lack of widespread public acceptance of some of these policies. Of course, there will always be people who disagree, but these are debates that can be had before policy execution, not during and after.

There are of course many ways that smart cities could be implemented, perhaps using less of the available technology and less surveillance, perhaps with a more restricted role for private corporations, perhaps with modest goals or with ambitious goals. However, the aim of this paper is more generally an attempt to convey that the smart city, as a prime example of the information economy in general, represents a significant alteration in our core societal values and norms. This therefore needs detailed public debate before careful introduction, instead of which we are getting rapid roll out in certain jurisdictions and insufficient public engagement. The information economy exchanges the realms of the public and the private and values transparency, assurance and consequentialism as a replacement to trust, individual privacy and personal responsibility. Once in process there are feedback loops that will tend to accelerate this transition. This approach is bolstered by developments in behavioral science that can be used to suggest that humans are not inherently responsible, despite having built successful societies for millennia. In the words of Crawford, the emphasis on the irrationality of humans could create a tendency to think of the public:

“not as citizens whose considered consent must be secured, but as particles to be steered [44]”.

This view of the public as “steerable particles” would potentially accelerate in a smart city. We can look to the recent pandemic for examples of data driven decision making which led to ideas such as lockdowns and vaccine passports, which were never a feature of pandemic planning in the past. Gloude-mans et al. [45] write that “in a smart cities context, it becomes possible to measure inter-personal distance using networked cameras and computer vision analysis” in their experimental study on “small behavioral interventions” “aimed at increasing distancing compliance”. Again, these are assurance techniques that leave no room for individual decision making. This removal of agency has consequences, since public trust decreases when people do not make their own behavioral decisions [33], which then in turn further increases the preference for data-driven, consequentialist decision making [24]. The exchange of the public and the private sees transparency expected of citizens, while large companies promoting it are able to keep methods, algorithms, and databases private [26]. The reach of the information economy into the political landscape is not inherently a feature of a smart city, but is something that could be magnified

there if not regulated. Clearly, taking voter influencing campaigns from the public sphere into the private sphere is an interference in the democratic process and is detrimental to institutional trust. Algorithmic control of content visibility on platforms such as Facebook and Twitter provide another example. Twitter particularly is used for “information sharing among politicians, journalists and citizens” [38]; such communication would be considered part of the public sphere and yet the visibility of posts can be controlled by algorithms on a privately owned platform.

On a positive note, if we slow down, there is nothing preventing us from creating smart cities that protect privacy and individual rights whilst increasing efficiency and reducing waste. Any smart city that manages to comply fully with the EU GDPR will have significant privacy protections in place including that of the ‘right to be forgotten’ unless data is truly anonymized. There is nothing preventing us from understanding that complex systems and complex problems cannot be understood by data alone, that decision making must give adequate weight to ambiguous factors. There is nothing preventing us from protecting citizens from psychological manipulation and preventing political campaigning or content manipulation to be conducted on private platforms. We can create a more positive smart city by being aware of the risks of creating a bad one. It is possible that the smart city could be a place devoid of public and institutional trust, governed by assurance and pre-emptive policing; “the mechanical aspects of power” in the words of Zuboff. But by understanding how trust can be eroded we can instead try to reverse the process by encouraging individual responsibility and agency wherever possible. We can think carefully about how the wealth divide exacerbated by the information economy can be reduced without further reducing individual agency in the process. We can slow down so that our legal system and non-expert members of society can catch up with the new technology and be included in deciding how it is implemented. We can, in short, learn how to integrate the individual back into the information economy, such that Adam Smith could indeed live in a smart city.

References

1. Smith A (2014) *The wealth of nations*. Sheba Blake Publishing, New York, New York
2. Sinnott-Armstrong W (2019) Consequentialism. In: Zalta EN, Nodelman U (eds) *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Summer 2019 edn
3. Malek JA, Lim SB, Palutturi S (2021) The ethics of smart city planning: examining post-utilitarianism in Malaysian blueprints. In: 2021 international conference on ICT for smart society (ICISS). IEEE, pp 1–5
4. Balsillie J (2018) Sidewalk Toronto has only one beneficiary, and it is not Toronto. *The Globe and Mail* 5
5. O’Kane J (2022) *Sideways: the city google couldn’t buy*. Random House Canada
6. Camerer C, Loewenstein G, Prelec D (2005) Neuroeconomics: how neuroscience can inform economics. *J Econ Literature* 43(1):9–64
7. Saint-Paul G (2013) Liberty and the post-utilitarian society. *Econ Aff* 33(1):119–126
8. Hardyns W, Rummens A (2018) Predictive policing as a new tool for law enforcement? recent developments and challenges. *Eur J Crim Policy Res* 24:201–218

9. Smart-Akande S, Pinney J, Hewage C, Khan I, Mallikarachchi T (2022) Preemptive policing: can technology be the answer to solving london's knife crime epidemic? *Artif Intell Natl Secur*, pp 205–230
10. The coronation arrests are just the start. police can do what they want to us now, *theguardian.com*. <https://www.theguardian.com/commentisfree/2023/may/12/coronatio-nprotest-arrests-police/>. Accessed 17 June 2023
11. Alexander L, Moore M (2016) Deontological ethics. In: Zalta EN (ed) *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Winter 2016 edn
12. Ginbar Y (2008) Why not torture terrorists?: moral, practical, and legal aspects of the 'ticking bomb' justification for torture. Oxford University Press, Oxford [UK]
13. Rebera AP, Rafalowski C (2014) On the spot ethical decision-making in CBRN (chemical, biological, radiological or nuclear event) response: approaches to on the spot ethical decision-making for first responders to large-scale chemical incidents. *Sci Eng Ethics* 20(3):735–752
14. Hursthouse R, Pettigrove G (2018) Virtue ethics. In: Zalta EN, Nodelman U (eds) *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Winter 2018 edn
15. Zagzebski LT (2004) *Divine motivation theory*. Cambridge University Press, Cambridge
16. Kumar V, Campbell R (2022) *A better ape: the evolution of the moral mind and how it made us human*. Oxford University Press, New York
17. Circle city con keynote, *iamthecavalry.org*. <https://iamthecavalry.org/2014/08/28/circle-city-keynote-text/>. Accessed 18 Nov 2022
18. Ananny M, Crawford K (2018) Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc* 20(3):973–989. <https://doi.org/10.1177/1461444816676645>
19. Kitchin R (2016) The ethics of smart cities and urban science. *Philos Trans Royal Soc A: Math, Phys Eng Sci* 374(2083):20160115
20. Camerer C, Weber M (1992) Recent developments in modeling preferences: uncertainty and ambiguity. *J Risk Uncertain* 5(4):325–370
21. van Dijk E, Zeelenberg M (2003) The discounting of ambiguous information in economic decision making. *J Behav Decis Mak* 16(5):341–352
22. Allen DW (2022) Covid-19 lockdown cost/benefits: a critical assessment of the literature. *Int J Econ Bus* 29(1):1–32
23. Li Z, Liang C, Hong Y, Zhang Z (2022) How do on-demand ridesharing services affect traffic congestion? the moderating role of urban compactness. *Prod Oper Manage* 31(1):239–258
24. Bundi P, Pattyn V, Trust, but verify? understanding citizen attitudes toward evidence-informed policy making. *Public Administration* n/a(n/a). <https://doi.org/10.1111/padm.12852>, <https://onlinelibrary.wiley.com/doi/abs/https://doi.org/10.1111/padm.12852>
25. Lewis JD, Weigert A (1985) Trust as a social reality. *Soc Forces* 63(4): 967
26. West SM (2019) Data capitalism: redefining the logics of surveillance and privacy. *Bus Soc* 58(1):20–41. <https://doi.org/10.1177/0007650317718185>
27. Peppet SR (2011) Unraveling privacy: the personal prospectus and the threat of a full-disclosure future. *North Western Univ Law Rev* 105: 1153. <https://scholar.law.colorado.edu/faculty-articles/177>
28. The moral character of cryptographic work, *web.cs.ucdavis.edu*. <https://web.cs.ucdavis.edu/rogaway/papers/moral.pdf>. Accessed 18 Nov 2022
29. Zuboff S (2015) Big other: surveillance capitalism and the prospects of an information civilization. *J Inf Technol* 30(1): 75–89. <https://doi.org/10.1057/jit.2015.5>, <https://doi.org/10.1057/jit.2015.5>
30. Fairfield J, Shtein H (2014) Big data, big problems: emerging issues in the ethics of data science and journalism. *J Mass Media Ethics* 29(1):38–51
31. Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a "right to explanation." *AI Mag* 38(3):50–57
32. Usage based insurance, *the-digital-insurer.com*. <https://www.the-digitalinsurer.com/wp-content/uploads/2015/09/572-EY-usage-based-insurancethe-new-normal.pdf>

33. Cheshire C (2011) Online trust, trustworthiness, or assurance? *Daedalus* (Cambridge, Mass.) 140(4): 49–58
34. China's social credit score—untangling myth from reality, *merics.org*. <https://merics.org/en/opinion/chinas-social-credit-score-untangling-mythreality>. Accessed 18 Nov 2022
35. China backs un pledge to ban (its own) social scoring, *politico.eu*. <https://www.politico.eu/article/china-artificial-intelligence-ai-ban-socialscoring-united-nations-unesco-ethical-ai/>. Accessed 18 Nov 2022
36. Central planning, local experiments, *merics.org*. https://merics.org/sites/default/files/2020-04/171212_China_Monitor_43_Social_Credit_System_Implementation.pdf. Accessed 18 Nov 2022
37. Leaked: Cambridge analytica's blueprint for trump victory, *theguardian.com*. <https://www.theguardian.com/uk-news/2018/mar/23/leaked-cambridge-analyticas-blueprint-for-trump-victory>. Accessed 18 Nov 2022
38. Tufekci Z (2014) Engineering the public: big data, surveillance and computational politics. *First Monday* 19(7)
39. Options for increasing adherence to social distancing measures, *assets.publishing.service.gov.uk*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/887467/25-options-for-increasingadherence-to-social-distancing-measures-22032020.pdf. Accessed 18 Nov 2022
40. Andrić K, Kasirzadeh A (2023) Reconciling governmental use of online targeting with democracy. In: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency. FAccT '23, Association for Computing Machinery, New York, NY, USA, pp 1871–1881. <https://doi.org/10.1145/3593013.3594133>
41. Duan W, Nasiri R, Karamizadeh S (2020) Smart city concepts and dimensions. In: Proceedings of the 2019 7th international conference on information technology: IoT and smart city. ICIT '19, Association for Computing Machinery, New York, NY, USA, pp 488–492. <https://doi.org/10.1145/3377170.3377189>
42. Zakzak L (2019) Citizen-centric smart city development: The case of smart Dubai's "happiness agenda". In: Proceedings of the 20th annual international conference on digital government research. dg.o 2019, Association for Computing Machinery, New York, NY, USA, pp 141–147. <https://doi.org/10.1145/3325112.3325236>
43. London mayor has no legal power to expand Ulez zone, high court told, *theguardian.com*. <https://www.theguardian.com/politics/2023/jul/04/london-ulez-expansion-tory-councils-legal-challenge>. Accessed 14 Aug 2023
44. Covid was liberalism's endgame, *unherd.com*. <https://unherd.com/2022/05/covid-was-liberalisms-endgame/>. Accessed 18 Nov 2022
45. Gloudemans D, Gloudemans N, Abkowitz M, Barbour W, Work DB (2021) Quantifying social distancing compliance and the effects of behavioral interventions using computer vision. In: Proceedings of the workshop on data-driven and intelligent cyber-physical systems. DI-CPS'21, Association for Computing Machinery, New York, NY, USA, pp 1–5. <https://doi.org/10.1145/3459609.3460523>

An Analysis of Blockchain Adoption in Zimbabwean Mining Land Title Management (MLTM) Using NVivo



David V. Kilpin, Eustathios Sainidis, Hamid Jahankhani, and Guy Brown

Abstract Zimbabwe is home to an expansive ore reserve estimated at 13 million metric tonnes. Despite its long mining history spanning a century, only 580 metric tonnes have been extracted, indicating a massive reservoir of untapped potential [Goosen M (2023) Top 5 minerals produced in Zimbabwe. Energy Capital Power]. However, this potential is hindered by significant challenges, the foremost being the threat of mining land dispossession. This paper seeks to pave a way forward for Zimbabwe's mining sector by highlighting blockchain's transformative potential, provided robust governance supports this transition. The findings underscore that Zimbabwe's manual mining title system is outdated and rife with inefficiencies. The qualitative insights gathered from the diverse interviewees revealed a consensus: the present manual system is susceptible to manipulation, leading to title disputes that hamper operations and dent investor confidence. Furthermore, the pervasive sense of insecurity stemming from potential land dispossession discourages investment and disrupts community relations, exacerbating the challenges within a dynamic sector. Utilising NVivo for analysis, this research delves deep into the challenges inherent in the manual system of mining titles.

Keywords Cadastre · Blockchain · Self-Sovereign Identity (SSI) · Mining Land Title Management (MLTM) · NVivo

1 Introduction

In recent decades, Zimbabwe has emerged as a major mining country. Global extractive logic, which has a colonial history, is reanimated as investors from around the world seek investment opportunities in the booming mining industry in the country. However, there are issues with tenure, bureaucratic bottlenecks, reliance on manual systems, and a lack of proper demarcations that are a significant problem for local

D. V. Kilpin · E. Sainidis · H. Jahankhani (✉) · G. Brown
Northumbria University, London, UK
e-mail: Hamid.jahankhani@northumbria.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_19

363

and foreign investors. These issues only benefit a small number of miners who can rely on connections and patronage to access mining rights. The problem that arises in this context is that weaker miners, indigenous communities, and local farmers are at a disadvantage as they find the system working against them. Overall, such a condition also leads to decreased productivity due to delays and sunk costs associated with getting mining rights. Such a context makes it necessary to understand the implementation of blockchain in the Zimbabwean mining industry, including perceptions towards digitisation and the benefits miners and other stakeholders think they may be able to accrue with the implementation of blockchain in mining titles. Overall, blockchain implementation holds the promise to reduce the possibility of tampering with records and ensure the retrievability of data so that any undue external influences can be reduced, as these influences have been the cause of many disputes in Zimbabwe.

A blockchain is a growing list of records or blocks linked using cryptography. The purpose of having information about the previous block is that the preceding one reinforces each additional block. The fundamental advantage offered by blockchains is that they are secure by design, as data in any block cannot be retroactively altered without altering the subsequent blocks. The disjointed data repositories in Zimbabwe have been riddled with numerous errors and inconsistencies in administrative data, spelling errors, and discrepancies such as duplicated records, multiple copies of the same database with different contents, and absent data. Therefore, e-government is beneficial to economise and improve government service operations, including efficiency, reduced transaction costs, transparency, and improved citizen services.

Numerous policies have been related to digitisation in Zimbabwe, such as the Zimbabwe National Policy for Information and Communication Technology of 2015, which was informed by numerous national and international documents. However, the full fruits of the transformation have yet to be reaped, especially in the mining sector. Due to the convergence of technology platforms, multiple services previously available on separate platforms can now be obtained on a single platform or network. It is no longer necessary to have numerous institutions overseeing the development of electronic communications in any given country. As a result, through blockchain technology, a check and balance mechanism can be applied to each and every aspect of governance, thus eliminating the possibility of corruption.

As has been shown in the case of post-colonial societies, the extensive paperwork used for land ownership can be manipulated to make claims to land, such that the government's claim may differ from those of individuals. The problem is particularly acute in the mining sector, as ownership can be contested between governments, local landowners, and mining companies. Existing research on mining in Africa has shown that private companies have far greater power over local governments in shaping mining regulations, and this situation has resulted in the marginalisation of local owners. Land use conflicts between large-scale miners and community members have been frequent worldwide. As a result, it has become increasingly difficult for miners who demand a significant area to coexist with community members, especially as they rely on these lands for their livelihoods. Another reason why mineral rights are such a contested issue is because of the so-called rule of capture, according to

which minerals that have the capability of migrating beneath the Earth's surface can be extracted even if the original source was someone else's property. Given the concept of split ownership, this situation forces us to consider what possibilities and challenges exist in relation to the implementation of blockchain.

A major corruption problem persists across much of Africa, resulting in a different development impact than in other places. It is common for the "resource curse" concept to be applied to large industrial mining operations. Still, miner tensions and violence are often caused by the economic stakes associated with exploiting mineral resources. In Zimbabwe, IFFs were prevalent in the mining sector, with US\$ 2.7 billion lost due to illicit flows. The contested nature of landownership in the context of mining and the great potential for corruption using traditional cadastres raise the need to consider whether the implementation of blockchain in land registration in the mining sector can reduce corruption.

2 New Institutionalism (NI), Actor Network Theory (ANT) and Implications for Governance

Actor Network Theory (ANT) serves as the theoretical foundation for numerous e-government investigations. In their study of the technology used in border control, Whitley and [41] used the ANT notion of symmetry [19] between human and non-human actors. They concluded that it is helpful to analyse both groups of players simultaneously since border control issues relating to people (human actors) and products (technology) are comparable. Heeks and Stanforth offered an alternative viewpoint, claiming that an ANT analysis of the politics of processes in public sector systems is a great way to comprehend project trajectories and network formation among organisational actors, including the distinction between global and local actors. They noticed that, whether these procedures are successful or not, players' emergent power plays are crucial to understanding such processes. With a focus on networking success and failure, as well as global and local actors, [36] employed ANT to investigate e-government programmes. Stanforth's study stressed the significance of a strong mandatory passing point between international and local actors.

Similarly, [30] discovered that local actors' interests are less strongly reflected or represented than those of global actors in research on global and local actors. Holmström and Robey investigated the organisational effects of an online analytical processing tool in a municipal setting. By documenting the changes in the players' perceptions of the system, they described the politics of the successive enrollment of various group actors within the organisation. In the process of standardising electronic patient records, human and non-human actors created tensions, innovations, and significant events that Vikkelso described in great detail. She discovered that information infrastructures occasionally have unforeseen revolutionary consequences in terms of patients, professionals, and health records. These studies focus on common translation analysis and network constructions as the researchers discuss

and partially implement key ANT ideas. Although they only partially rely on it, scholars have also recommended the use of ANT in e-government scenarios. These researchers tested a particular methodological approach that involved focusing, structuring, and presenting a case study based on process events marked by the process trajectory of important activities. The rationale for this approach was the researchers' wish to avoid the risk associated with analysing broad processes with many participants where there is an excess of details. With such an approach, the application of ANT theory is strengthened. With the exception of the studies by Whitley and [41], and, to some extent, by [40], these studies are wide-ranging and empirically rich descriptions of development processes and the politics of network formation by human or organisational actors in e-government. The technical or non-human actors in processes inside various government institutions are not given enough attention despite the relevance of the symmetry principle in ANT, a gap this dissertation seeks to fill. Before considering how ANT and new institutionalism help us understand blockchain implementation in Zimbabwe, it is important to give a brief legal and policy background for mining in the country.

3 Digital Transformation of Mining Title Management in Zimbabwe: The Cadastre System

3.1 The Cadastre System or Concept

Ozah et al. [33] define a computerised mining cadastre (CMC) as a computer-based database system of mining titles, that hosts details associated with the titles and supports tools and functionalities for the execution of administrative or basic tasks. CMC aims to enhance the security of tenure, transparency in mineral licencing, and the government's regulatory capacity in the mining sector. Information or data attached to each mining title includes names of owner(s), time validity of ownership, geographical location, a track (historical) record of all transactions related to the title right from the initial application through to granting of title, all payments made, annual reports generated, lapses, re-allocations, encumbrances, or hypothecations, and title extinction. The CMC system should facilitate the execution of basic tasks related to mining titles, including simple data checks, data conversion (processing), data transfer, geometric validation of concessions, spatial queries, checking maps, simple monitoring of titles, updating of title registers, generation of statistical reports, and defining and controlling user rights, among others (Fig. 1).

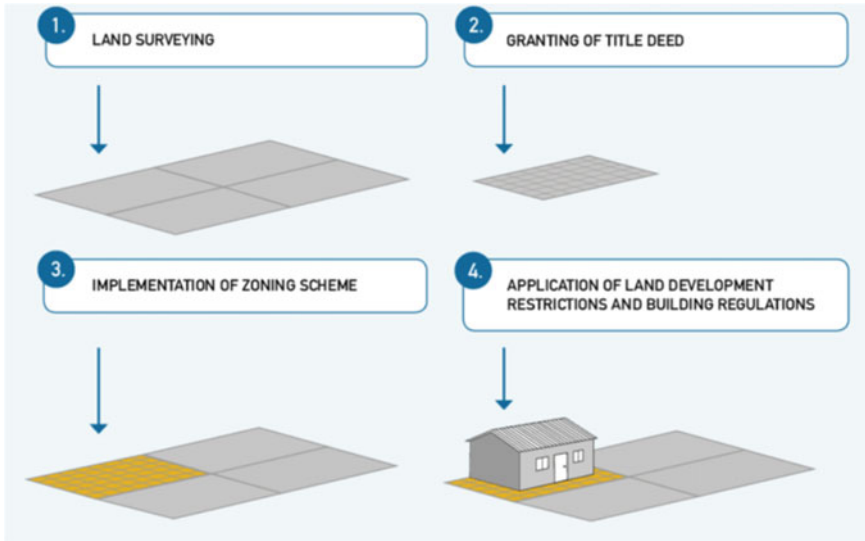


Fig. 1 Computerised mining cadastre (CMC) *Source* Charman et al. 2017

3.2 Implementation of the Computerised Mining Cadastre System: Lessons Learned from Implementing African Countries

The CMC system is being implemented in several African countries, for example, Mozambique, Zambia, Tanzania, Nigeria, the Democratic Republic of the Congo, Madagascar, and Namibia [8, 33]. The establishment of CMC is a major aspect of the Management of International Mineral Sector Reform projects run by Spatial Dimension and Swedish Geological AB, with the latter being the implementing partner [8]. The extensive work of these organisations in Africa has resulted in the development of a range of mining cadastre management information products termed FlexiCadastre, whose roots are in Africa. This uses a web portal for various mining title data processes, including data management, data reporting, task management, business logic, the use of GIS technology, and other innovative concepts.

By 2006, African countries using FlexiCadastre included Mozambique, Zambia, and Tanzania, and several countries were considering adopting it [8]. A corporate version of FlexiCadastre is used at a company level to achieve integrated organisational management and ensure overall consistency of company operations and systems, including quicker production of statutory reports. FlexiCadastre achieves efficient administration of mining titles through the creation, scheduling, and management of tasks, including tracking those that have been done by the title holder or government agencies as determined by the mining law in the concerned jurisdiction. An efficient mining cadastre fosters investor confidence and ensures

that the intentions of the mining law are fulfilled while providing important data for mining policy formulation.

According to Feast et al. [8], the implementation success of the mining cadastre depends on several factors, which include understanding the mining cadastre technology needed, the organisation's data, content, workflows, and business requirements, recognising stakeholder needs, streamlining business processes, facilitating stakeholder buy-in from the start, effectively managing expectations, and implementing adequate long-term support and maintenance. The development of the FlexiCadastre in Mozambique is funded by the World Bank, in Tanzania by the Nordic Development Bank, and in Zambia by the European Development Fund in conjunction with the World Bank. The CMC in Nigeria is funded by the World Bank and implemented by a Germany-based company, GAF AG, which has implemented similar systems in the Democratic Republic of Congo, Madagascar, and Namibia [33].

However, [33] found that the transition from a paper-based database to CMC is not guaranteed to be smooth due to institutional, administrative, and technical constraints and challenges. In the case of Nigeria, [33] identified the following constraints and challenges:

- disjointed pre-computerisation data repositories consisting of administrative data maintained in Microsoft Excel format and geometries of the applications and licences maintained in AutoCAD Drawing format;
- lack of supporting datasets, as shown through examples such as gaps in data on cartographic coverage, including spatial data on restricted areas like reserved forests, military zones, etc.
- significant mining regulatory gaps during the whole computerisation period, even though the Nigerian Minerals and Mining Act had come into effect in 2007 just before the exercise, with most of the computerised procedures being implemented temporarily, and
- the challenge of designing an optimal system that harmonises and synchronises processes across several zonal offices and the central office, given serious inadequacies in digital communication infrastructure (Fig. 2).

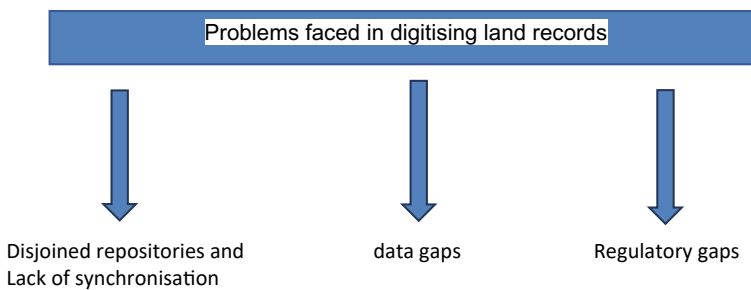


Fig. 2 Problems faced in digitising land records

The disjointed data repositories were riddled with numerous errors and inconsistencies, including administrative data errors, spelling errors, duplicate records, multiple copies of the same database with different contents, and absent data. For the geometrical or graphic data, plotting inaccuracies resulting in overlaps, multiple copies of the same concession in the same or different areas, insufficient links to administrative data, or the presence of the same database on different unconnected computers. The result was the need to go through a long and expensive process of data cleaning, consolidation, and migration to the new computerised platform. Nyaga et al. [29] identified legal hurdles, a lack of technical expertise in record keeping, and inadequate funding as serious challenges affecting the implementation of the cadastre system at the Ministry of Mining Headquarters in Nairobi.

3.3 Deployment of the Cadastre System in Zimbabwe

The Zimbabwean government has recently initiated the process of establishing a mining cadastre in line with the prevailing trend in Africa. It has created an online mining cadastre portal. The Ministry of Mines and Mining Development is currently going through a process of data capture, data cleaning and verification, and, at a later date, is expected to declare that all data have been verified. The current process apparently depends on the Ministry's own records and the input of title holders, applicants, and other stakeholders who can contribute by verifying the records in the developing cadastre and alerting the Ministry of any errors or omissions. In order to engage with the Ministry through the portal, the various stakeholders are required to register. Existing rights holders or those with pending applications need to download an MCO01 Registration Form, which they complete and submit in person, together with supporting documents. Prospective rights applicants do their self-registration completely online.

While the Mines and Minerals Amendment Bill 2015 proposes the CMC system, [17] notes that the bill does not make any significant changes to the minister's wide administrative discretionary powers over the granting, maintenance, and extinction of rights, which weaken the security of tenure. Suppose that administrative discretionary powers are not curtailed by law. In that case, the implementation of the CMC and its ability to achieve the lifeblood of the mining sector (protection of mining titles) will be seriously impeded. The seven critical success factors for CMC deployment identified by Feast et al. [8] are not easy to achieve in the context of a CMC project executed by the government, as is the case in Zimbabwe. The four critical challenges of transitioning from the manual to the CMC identified by [33] and Nyaga et al. [29] strongly apply to the Zimbabwean bureaucracy-driven process. The government has also demonstrated a lack of political will to implement initiatives that foster transparency and accountability in the mining sector.

4 Blockchain Technology and Its Implications for Land Title Management

4.1 *Blockchain in Cadastre*

A blockchain is a growing list of records or blocks that are linked using cryptography [26, 34]. A cryptographic hash is an algorithm that maps data of an arbitrary size to a bit array of a fixed size. Each block contains a hash of the previous block or a form of transactional data. The purpose of having information about the previous block is that the preceding one reinforces each additional block. Blockchains are typically managed in a peer-to-peer network for use on a distributed ledger, where nodes collectively adhere to a protocol to communicate and validate new blocks. The fundamental advantage of blockchain, which may be useful for its relevance to land title management, is that it is secure by design, as data in any block cannot be retroactively altered without altering the subsequent blocks [16]. The chain that joins the two blocks provides an iterative process and ensures the integrity of other blocks, and the integrity of the data is ensured after the block is digitally signed [18]. This means that blockchain provides a foolproof means of storing and retrieving data that is not possible with the manual storage of land titles in many developing countries.

One question arises about the digital identity of those who hold land. In the digital space, identities are constantly negotiated rather than fixed. Social scientists have raised the point that digital identities shape perceptions about physical identity. In fact, some academics have hypothesized that the embodied forms of identity that digital identity encompasses would liberate people from their bodies and blur the lines between humans and technology [21]. It has also been argued that digital identity would free individuals from distinctions of race, class, gender, etc. [39]. In this context, it would make sense to consider the notion of Self-Sovereign Identity (SSI) as an approach that gives individuals autonomous control over their digital identity [7]. According to SSI, there are entities whose identities are generated that correspond to the former. At the same time, they form the basis of attributes or identifiers. Thus, we can notice how SSI ensures anonymity at the same time as establishing trust. This is because one party often provides credentials to another party and the recipient that they can trust, and the trust of the verifier is then transferred to the credential holder [35].

Decentralised identifiers (DIDs) are often used to generate SSIs. These DIDs enable a verifiable, decentralised identity. Properties have traditionally been dealt with through cadastres, which have the capacity to provide a comprehensive recording of the real estate. Cadastral surveys document the boundaries of a property. A cadastral parcel is referred to as land that has a unique set of property rights [5]. What makes cadastral maps particularly open to politicking or manipulation is the paperwork they are generally accompanied by, compared to the seamless authentication process offered by blockchains. Specific provisions related to the jurisdictions shape how the documents are managed and authenticated. These jurisdictions shape the stakeholders who are required to sign and authorise the document. While, in

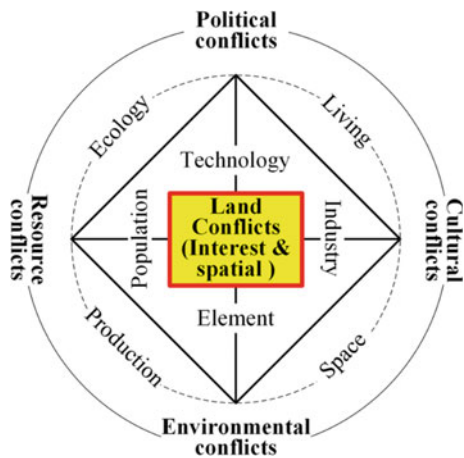
principle, the same mechanism can take place through blockchain via decentralised identities, traditional cadastral regulations often make it a tiresome process requiring multiple signatures, also creating storage problems.

4.2 Peculiar Features of the Land Mining Sector and the Applicability of Blockchain-SSI Technology to Mining Title Management: Adaptation or Adoption?

Existing research on mining in Africa has shown that private companies have far greater power over local governments in shaping mining regulations, which has resulted in the marginalisation of local owners. Due to the inability to monitor and implement regulations, it is possible that the result may be the bypassing of government regulations by private entities, which, in the past, have also co-opted armed groups to protect their installations. Land use conflicts between large-scale miners and community members have been a frequent occurrence worldwide. As a result, it has become increasingly difficult for miners who demand a significant area to co-exist with community members, especially as they rely on these lands for their livelihoods [14]. It has also been noted that government involvement for the purpose of dispute resolution is often minuscule. Thus, in the past, it has been recommended that if governments play a central role in arbitrating conflicts, it can possibly allow for improved dispute resolution. As Castro and Neilson [4] have shown, natural resource conflicts are often “severe and debilitating, resulting in violence, resource degradation, the undermining of livelihoods, and the uprooting of communities (229)” (Fig. 3).

Often, in the utilisation of natural resources like mines, parties tend to try and maximise their own interests and, in the process, end up defeating the interests of the

Fig. 3 The conceptual framework of land use spatial conflicts and risk factors
 Source Zhou et al. 2019



other [31]. Mining often has wide-ranging impacts on communities. For example, for an indigenous community that has inhabited a tract of land for several decades, a 20- to 30-year-old mining project is seen as “a disruption in the flow of time, much like the effects of an earthquake or major flood,” except that the effect can be much slower and enduring [2], 10). Yet, among developing countries in Africa, such as Zambia, mining contributes to 90% of the country’s export earnings. It provides employment to 15% of the country’s workforce, meaning that it is often in the government’s interests to incentivise mining companies over the interests of the communities [22].

In many countries, mining efforts have also meant that entire communities have effectively come to live under a country’s mining project, as in the case of the Northern Brazilian state of Para, where 400,000 individuals have come to live seeking to tap into the residents’ resources, 100,000 of whom directly live inside the gold mine [20]. Another way mining affects a local community is when demographic shifts caused by mining increase the cost of living, resulting in indigenous communities rapidly seeking to derive a livelihood in other areas [6]. Not only is there a possibility of conflict between landowners and owners of mining companies, but also between miners and, say, herders, as mining introduces new ideas of capital ownership that local communities are not familiar with.

A history of conflict between small miners and large-scale mine management characterises countries like Ghana. Often, individual mine lands are located illegally within the land occupied by large-scale mining companies, which is often the product of familial and cultural relations with the land. For instance, in Fiji, indigenous people believe they own everything above or below the ground [24], whereas, in New Guinea, 90% of the land is owned by residents [25]. This means that mining companies need to consider a range of situations based on the context, as the concern of tenure is different. Another example of this is the periodic reallocation of land among Swat Pathans, which makes it difficult for the state to institute its own definitions of property ownership. Yet, in much of the development, the government retains the right to sub-surface minerals and is eager to contract mining rights to international companies [38].

Another valuable example can be seen in the Australian mining company, which has been mining in PNG. This company began prospecting at the site in 1970 and has now established one of the world’s biggest gold and copper mines. The company has a significant positive impact on national economies, yet an enormous amount of chemical waste has been released into the surrounding environment, as [37] have shown. There have also been numerous instances of careless mining, which has resulted in wide areas adjacent to the copper slurry line becoming contaminated when the mine fractured and released slurry [11]. At the same time, as Australian companies have expanded mining activities, there has been a reaction from local populations about the violation of sacred sites [23]. Land use conflicts between local groups and the mine management, as one local landowner claimed, with the mining operation, life has changed, resulting in the loss of rainforest and the degradation of the local environment, when rural farmers depend on the local environment for their survival [13]. Such cases have been noticed globally as local groups have demanded compensation for the loss of livelihood due to mining activities. In such a case, [15]

have made a case for the need for effective communication in the mining industry and that it is crucial for the benefit of surrounding communities that public meetings be held in which their questions are effectively addressed by mining personnel.

Mining activities have thus had a direct impact on the way local communities understand their relationship to land. As [3] has shown in the New Peruvian mining industry case, trans-nationalisation and privatisation resulted in mining emerging as the key element for future development activities. The author suggests that mining deeply impacted land-tenure relations and land values and resulted in a shift in the spatial distribution of land-use patterns throughout the region. In Tanzania, “conceptual differences between local people and investors on how land and mining rights are to be perceived also contribute to considerable conflict in the country. “ However, instead of viewing local populations as passive subjects, recent anthropological scholarship has shown how in Colombia, while the government blames small-scale farmers for lacking a sense of future, “marginalised people who extract gold with small-scale techniques create an alternative sense of future by engaging with the leftovers of their gold mining practises.” Small-scale mining constantly creates simple by-products that look like waste, including gases, rubble, and mud.

According to Jaramillo, small-scale miners engage with such substances as a way to make sense of their lives and the future. Making the relationships between humans and geological substances, waste, and technology visible elucidates alternative forms of life that “get in the way” of a multinational mining company, the national government, local mafias, and financial markets hungry for gold in times of crisis (48).

Given the digitisation problems without adequate training in much of the developing world, [1] make a case for incremental digitisation or the implementation of blockchain in land title management in Bangladesh. They propose a blockchain-based solution that offers data synchronisation and transparency, ease of access, and immutable record management as a faster and cheaper solution. They state, “Considering the technological knowledge and capacity of the people and the government, we introduced a phase-by-phase Blockchain adoption model that starts with a public Blockchain ledger and later gradually incorporates two levels of Hybrid Blockchain” ([1], 1319).

The resulting decentralisation of control, in combination with the immutable representation of the transfer of possession, provides a unique opportunity to build collaborative, multi-sided, ‘trustless systems’ [9]. [32] enumerate the benefits and promises of blockchain technology in government as a means of information sharing. In general, blockchain technology allows for client-defined chains of entries, client-side validation of entries, a distributed consensus algorithm for recording entries, and a mechanism for ensuring security. The data is secured by publishing it in encrypted form in an immutable, distributed ledger [28]. The storage of critical information in the decentralised data layer protects the system from malfunction, obsolescence, and deliberate or involuntary record changes. All critical data records are secured by a “proof of existence” in the blockchain.

Every essential step in the registration process that qualifies as a high “standard of care” can be published to the blockchain. This provides data security and

long-term data stability by making the entire data set immutable via notarisation. In addition, publishing data during each step of an instrument’s recording process provides irrefutable “proof of process”, protecting the submitted records and establishing accountability. It is important to note that blockchain technology does not help address the accuracy of land titles [32]. According to [27], 80% of the work involved in adopting blockchain technology is associated with changing business processes, and only 20% is related to technology implementation. Most of this transformation happens in the background because blockchain is an invisible enabler that does not affect externally visible parts of a business or organisational entity. However, it does require commitment from the involved parties and a sound business process redesign to ensure a successful implementation.

4.3 Qualitative Analysis of Data on Mining Disputes

The study explores the following research questions based on preliminary interviews, pilot interviews, and the literature survey. The three questions explore the problems in the security of titles created by manual records, the potential of digital technologies to increase transparency and reduce the possibility of tampering with records, and finally, the role of blockchain in reducing conflicts over mines, given the ease of retrieving data and the difficulty they create in tampering with data.

1. What are the ways in which manual mining documents (including maps and titles) create possibilities for conflict among stakeholders? How does this result in a loss of economic productivity?
2. What are the attitudes among mining personnel and bureaucracies towards digitising mining titles? What are some of the hesitations, and how does digitisation ensure greater transparency and greater retrievability in mining titles?
3. What role can blockchain play in increasing transparency and reducing the manipulability of records? How much awareness do stakeholders in the mining sector have of blockchains?

The table below provides a summary of the sample size.

Group name	Description (pointer to sampling frame)	Size sample group (number of participants)
Large-scale mining company executives	Companies which are members of the chamber of mines of Zimbabwe	7
Small-scale miners	Affiliate members of the Zimbabwe miners federation	8
Government officials	Officials of the ministry of mines and mining development	8

(continued)

(continued)

Group name	Description (pointer to sampling frame)	Size sample group (number of participants)
Parliamentarians	Members of the parliamentary portfolio committee on mines and mining development	3
Civil society organisations (CSO) officials	CSOs, which are affiliate members of the national association of non-governmental organisations	2
Rural district council (RDC) officials	RDCs are affiliate members of the association of rural district councils (ARDC)	1
Academic (university lecturer)	University of Zimbabwe	1
<i>TOTAL</i>		<i>30</i>

4.4 Qualitative Data Organisation Technique

Thirty semi-structured interviews were conducted with participants across the Zimbabwean mining ecosystem. After transcribing these interviews, they were coded thematically using NVivo. Template analysis was used to develop a coding template in which themes were identified as important in the data set, which was then organised in a meaningful and useful manner. Based on the collection of additional data, codes were also iteratively adjusted to account for the new data that was interpreted. There are many benefits to using coding software. Coding software like NVivo enhances thematic data analysis by helping find recurring, meaningful patterns of information within a dataset. The use of coding software helps increase efficiency by increasing the speed of data analysis and helping process large volumes of data. The use of the coding software also allowed for increasing accuracy by helping apply consistent rules for data analysis, reducing the possibility of human error, and finding precise data patterns. Using coding software also helped categorise data into multiple classification levels and analyse the interrelationships between them. Coding software also helped to visualise the data to help understand the different levels of classification. The coding software also allowed for an increase in the validity of the data as it helped to compare templates of data collected over time. NVivo also helped to compare the study’s findings with those from follow-up studies. The coding software also allowed me to save, export, and store data organised into multiple files corresponding with different codes and sub-codes while ensuring the safety of the data.

The interview questions were semi-structured and elicited some binary responses (yes or no) to get a general sense from participants about their awareness of the

cadastre. A section with open-ended responses followed this section. The questions mainly focused on the following areas:

1. Receptivity to technological innovation in mining titles
2. Problems with the manual system
3. Contestations over ownership.

The diagram below visualises the main codes and sub-codes.

Upon initial analysis, I organised the qualitative data in the form of the table mentioned above. I later reorganised the data in the following manner by rearranging codes. For example, because the fundamental problems were related to contestations over ownership, I moved them as the first code, compared to the third one above. Then, I tweaked the code titled “problems with the manual system” to reflect the high incidence of double allocations, as all the sub-codes in the category introduced concepts and ideas related to the fundamental problem of double allocations. The third code was on receptivity to digital technologies, which was further organised into sub-codes about the receptivity of blockchains, cadastres, and digitisation more generally. The iterative adjustment of the nodes is provided in the diagram below.

Below, I provide descriptions of each of the three major codes and sub-codes within each broader code and how each classification or sub-classification was used to analyse data. Each code organised qualitative data into themes. Each sub-code was used for further classification within a code. Codes and sub-codes also corresponded with major “themes”, which were generated based on the survey of existing literature and the analysis of preliminary qualitative material collected during pilot interviews. During analysis, these themes were used to organise the data into codes and further into sub-codes. Themes are used for abstract ideas generated based on the literature analysis, whereas codes are tailored to the specific empirical findings of the interviews.

Below, I provide information about how each code and sub-code was used for analysis. First, I will discuss the importance of each code. The three major codes included:

1. Contestation over ownership
2. High incidence of double allocations
3. Receptivity to digital technologies.

These were the parent nodes. Creating the most general category, often referred to as a “parent node,” in NVivo involves identifying overarching themes or concepts encompassing a wide range of subtopics. This parent node serves as a high-level category under which you can organise more specific subcodes.

4.4.1 Code: Contestation Over Ownership

Description: The “Contestation of Ownership in Mining Titles” code is used to identify instances within the qualitative data where there are disputes, conflicts, or challenges related to the ownership of mining titles or rights. This code is applied

when the data contains discussions, narratives, or descriptions of disagreements and controversies that revolve around the legal ownership, control, or rights associated with mining claims, leases, concessions, permits, or licences.

Purpose: This code aims to systematically capture and analyse the various forms of contestation that arise in the context of mining title ownership. It helped me understand the complexities surrounding ownership disputes and the factors contributing to these conflicts. By coding instances of contested ownership, I also examine the underlying legal, economic, environmental, and social issues that shape the mining industry's dynamics.

The code was applied to the following set of qualitative data:

1. Instances where multiple parties claim ownership over the same mining title, leading to disputes and legal battles.
2. Descriptions of conflicts between indigenous or local communities and mining companies regarding ownership and land rights
3. Cases where changes in government policies or regulations lead to disputes over the ownership of mining licences.
4. Narratives of disagreements between stakeholders, such as government agencies, private companies, local communities, and environmental groups, regarding ownership rights.

Subcodes

Contestation over ownership includes different types of disputes and claims-making. The following are some of the sub-notes representing the types of disputes participants discussed in the interviews:

- i. Farmer versus miner conflict
- ii. Big versus small miners
- iii. Problems with mapping
- iv. Security of title or tenure.

Under the sub-code farmer vs. miner conflict, all the data about conflicts between farmers and miners was included. Farmers had rights over land, but miners could violate these rights. Another type of conflict would take place between big and small miners. Participants shared that big miners could influence the bureaucracy, get licences, and resolve disputes in their favour. These conflicts were also a product of problems with mapping land and mines. Participants shared that demarcations were stored and are still kept in papers. These papers could always be manipulated, stolen, faded in ink, or torn. All this led to miners sharing that they had no security of title or tenure and felt that someone with more influence and power could take over their mines and lands.

This analysis of the Zimbabwean mining industry presents an illuminating perspective on three main codes: receptivity to technological innovation, problems with the current manual system, and contestations over ownership. Based on the semi-structured interviews, the study identified key sub-codes within each broader theme, contributing to a nuanced understanding of the industry's current state and potential

future directions. The exploration of these themes sheds light on the complexity of issues inherent in the industry while highlighting the potential benefits of digitisation and emerging technologies like blockchain in overcoming these challenges. However, it also underscores the need for a comprehensive and multi-faceted approach where technology addresses the root cause of problems in the mining sector.

Contestations over ownership emerged as a complex theme with far-reaching implications. Conflicts were reported between various stakeholder groups, particularly farmers and miners and large and small miners. However, the analysis reveals that these conflicts cannot be disassociated from the broader power dynamics, historical context, or legal and regulatory framework. Furthermore, the security concerns over title and tenure reflect the dire need for transparent and enforceable regulations to safeguard mining rights.

While technological innovations such as digitising mining titles can potentially address some of these issues, they do not offer a comprehensive solution to the deeper, embedded conflicts. Equitable access to resources, recognition of land rights, and inclusive decision-making processes must be incorporated into any proposed solution to mitigate ownership contestations.

4.4.2 Code: High Incidence of Double Allocations in the Manual System

The “High Incidence of Double Allocations in the Manual System” code is applied to instances in the qualitative data highlighting the frequent occurrence of dual allocations of mining rights or licences within Zimbabwe’s manual mining administration system. This code is used to capture discussions, narratives, or accounts of situations where the same mining area, claim, or resource has been allocated to multiple parties due to errors, inefficiencies, or a lack of coordination within the manual allocation process.

Purpose: This code aims to systematically identify and analyse the challenges associated with the manual allocation system in Zimbabwe’s mining sector. By coding instances of double allocations, I sought to understand the underlying causes of these errors, their implications for various stakeholders, and potential recommendations for improving the allocation process to ensure fairness, transparency, and sustainable resource management.

Examples of the application of the code:

- Descriptions of cases where multiple mining companies or individuals are granted rights to exploit the same mineral resource, leading to disputes
- Accounts of inefficiencies and lack of cross-checking in the manual allocation process, resulting in overlapping claims.
- Narratives of conflicts between parties claiming ownership of the same mining area due to double allocations
- Discussions on double allocations’ economic, social, and environmental consequences on local communities and the mining industry.

When applying the “High Incidence of Double Allocations in the Manual System” code, I adhered to the following parameters:

- **Clear Cases:** Code instances that explicitly describe instances of double allocations, where mining rights have been granted to more than one entity for the same area.
- **Allocation Process:** Capture details about the manual allocation system’s shortcomings, such as lack of coordination, inadequate verification procedures, or insufficient data management.
- **Impact:** Document the consequences of double allocations, including disputes, legal battles, environmental degradation, loss of revenue, and negative effects on local communities.
- **Stakeholders:** Note the involvement of various stakeholders, such as government agencies, mining companies, local communities, and civil society organisations, in responding to or being affected by double allocations.

Subcodes

In this code, participants talked about a range of issues, which predominantly included the following areas, which were also used as sub-codes:

- i. Tampering
- ii. Double pegging
- iii. External influence or patronage
- iv. Ethics
- v. Increase in corruption
- vi. Bottlenecks
- vii. Loss in productivity.

The problems associated with the manual system are many. Participants mainly talked about the ability to manipulate title and ownership information. The sub-code for tampering referred to the presence of external influence to alter title documents. Participants also talked about the common occurrence of double pegging. This was usually the case when existing land allotments were either unknown or their proofs were tampered with. There were different types of external influences caused by the presence of powerful entities with connections in the government bureaucracies who could be influenced to alter title documents. All participants shared that it was inherently the manual nature of the system and the paperwork associated with it that made the system prone to these influences. Participants were also asked what they thought about the presence of ethical standards.

The data added in the sub-code included participants’ perceptions of rules, the degree to which they were followed, and whether there were punishments or penalties for violations or rent-seeking behaviours. Other problems associated with the manual system included the possibility of corruption. The data included in the sub-code of corruption mainly revolved around the lack of adequate compensation for government employees and the lucrative nature of mining, which resulted in many accepting bribes. These bribes also led to expediting applications for titles, and

transfer of ownership, and dispute resolution, without which, participants claimed, files and applications would remain stuck in bureaucracy. The manual system was also attributed as the cause of unnecessary bottlenecks. This meant the mining documents took very long to change hands, and miners were left to travel long distances and were made to wait long periods to get mining licences. Some participants stated that the only way to overcome this problem was by bribing state officials. This delay also led to a loss of productivity as lands with useful resources would remain vacant. Participants discussed the loss of productivity due to the time and money initially invested to start mining.

The investigation highlighted a plethora of issues associated with the current manual system. Concerns regarding title tampering, double-pegging, and the influence of power entities were recurrent themes. Interestingly, while these problems were predominantly attributed to the manual nature of the system, it appears they are deeply rooted in broader socio-political dynamics that transcend technological issues. For instance, corruption was highlighted as a significant concern, leading to expedited applications and dispute resolutions at the expense of fair procedures. This problem points towards a systemic issue related to governance, power structures, and economic inequality, warranting further exploration beyond mere technological solutions.

Furthermore, the manual system's inefficiencies lead to unnecessary bottlenecks and significant productivity losses, indicating the urgent need for a streamlined and efficient process. While digitisation could potentially alleviate these issues, exploring alternative solutions, such as policy reforms, capacity building, and improved governance, is equally crucial.

4.4.3 Code: Receptivity for Innovation in Mining Titles

Description: The “Receptivity for Innovation in Mining Titles” code is applied to instances within the qualitative data that highlight the attitudes, perceptions, and responses of miners in Zimbabwe towards innovative technologies and approaches related to mining titles and administration. This code is used to capture discussions, narratives, or accounts that reveal the degree of openness, interest, and willingness among miners to adopt innovations such as digital cadastres, blockchain technology, and broader digitisation efforts within the mining sector.

Purpose: This code aims to systematically explore and analyse miners' views on embracing innovative methods for managing mining titles. By coding instances of receptivity to innovation, I aimed to uncover factors influencing miners' willingness to engage with technological advancements, improve administrative efficiency, reduce inefficiencies, enhance transparency, and address challenges associated with traditional manual systems.

Examples of the application of the code:

- Descriptions of miners' opinions on using digital cadastres to streamline and simplify the process of acquiring and managing mining titles.
- Accounts of miners' attitudes towards integrating blockchain technology to enhance the security and authenticity of mining title records.
- Narratives of experiences with digital platforms for submitting applications, managing licences, and conducting transactions in the mining sector.
- Discussions on the benefits and concerns miners associate with adopting digitisation measures to improve overall operations and transparency.

When applying the "Receptivity for Innovation in Mining Titles" code, I followed the following parameters:

- **Attitudes:** Code instances that provide insight into miners' receptivity towards adopting innovative technologies, such as digital cadastres, blockchain, and digitisation.
- **Reasons:** Capture miners' reasons for being open to or hesitant about embracing technological innovations in mining title administration.
- **Perceived Benefits:** Document the potential advantages miners associate with innovative solutions, such as reduced paperwork, faster processing, reduced fraud, enhanced data accuracy, and improved access to information.
- **Barriers:** Note any concerns, barriers, or challenges miners express regarding adopting new technologies and their impact on their operations or livelihoods.

Subcodes

During the analysis of transcripts, it was found that there were sub-topics within "receptivity to technological innovation", which included:

i. Existing awareness of the cadastre

This sub-code referred to whether participants were aware of existing efforts to implement the cadastre for mining titles in Zimbabwe. A lot of participants were aware of ongoing efforts, but some did not know much about the cadastre and the benefits it could accrue.

ii. Awareness of blockchain

This sub-code referred to knowledge about blockchains generally and their application in managing mining titles. Those unaware of blockchain were provided with hypothetical examples to illustrate the security of records provided by blockchain and to elicit whether they perceived blockchain favourably or not. Many shared that blockchain could reduce the possibility of manipulating dates, ownership information, and geographical boundaries.

iii. Benefits of digitisation

This sub-code considered perceptions about the digitisation of mining titles even if participants did not have complete knowledge about the importance of blockchains.

Many suggested benefits related to transparency but also suggested that digitisation could help in communication, mapping practises, etc.

iv. Need for transparency

The majority of participants agreed that digitisation could result in greater transparency. This meant that people could access publicly available data, which is currently not accessible because of information asymmetries. On the one hand, participants raised the issue of privacy over ownership information. On the other hand, they talked about difficulties with accessing information that is meant to be publicly available.

The data reveals a mixed level of awareness and receptivity towards the potential role of technological innovation in mining titles. While a significant number of participants demonstrated awareness of ongoing efforts to implement the cadastre system, a knowledge gap concerning blockchain technology was evident. This finding is crucial, as it points towards the need for a more expansive educational initiative about potentially transformative technologies within the industry. Despite this gap, the perceived benefits of digitisation, such as enhanced transparency, improved communication, and efficient mapping practises, were consistently recognised by participants. These benefits and the need for improved transparency present a compelling argument for digital transformation in the industry. However, the interviews also hinted at a complex interplay between the need for public data accessibility and concerns over privacy in ownership information, signifying the need for a balanced approach in digital implementation.

Based on the organisation of data above, I rearranged the data I rearranged the data in the following order, which can also be noticed in the difference between Figs. 4 and 5 above. The diagram below temporally organises the data in that it arranges data based on the fact that there is a high rate of contestation over mines, which is made possible by the problem of double allocations caused by the manual system. The problem of double allocation then results in increasing receptivity towards technological innovation among miners, including cadastres, blockchains, and digitisation more generally (Fig. 6).

5 Conclusion

In conclusion, digitising mining titles in Zimbabwe represents a transformative step towards a more efficient, transparent, and sustainable mining sector. As this paper has explored, the traditional manual systems for managing mining titles have been marred by challenges such as inefficiencies, corruption risks, and disputes. The introduction of digital cadastres, blockchain technology, and broader digitisation efforts has the potential to address these issues and unlock numerous benefits for all stakeholders involved.

The analysis presented in this paper underscores the positive impact that digitisation can have on streamlining administrative processes, reducing paperwork,

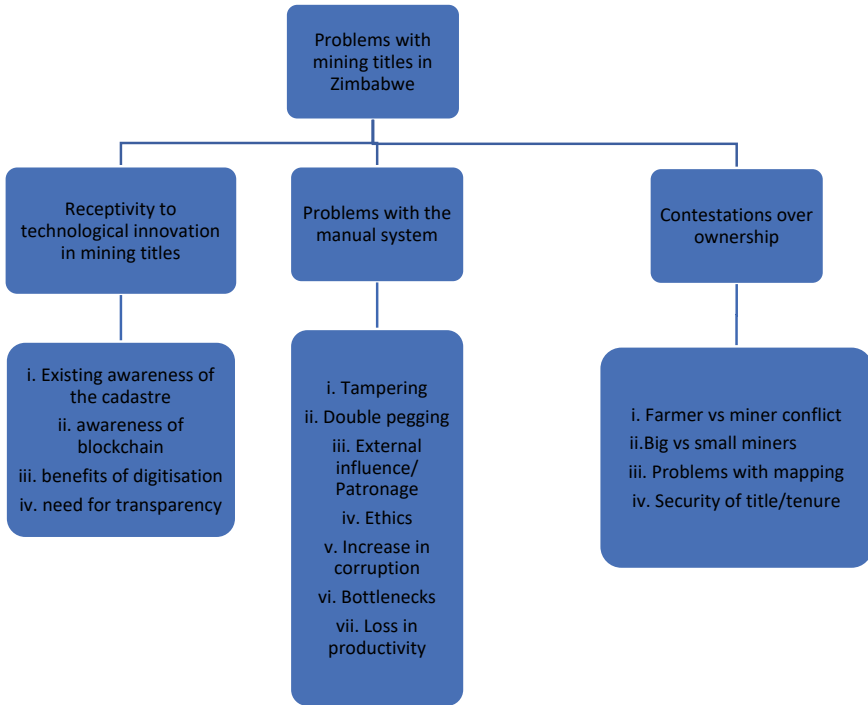


Fig. 4 Codes and sub-codes

enhancing data accuracy, and promoting accountability. Additionally, blockchain technology’s transparency and tamper-resistant nature offer promising solutions for tackling corruption and fraud in the allocation and management of mining rights.

However, it is important to acknowledge that the successful implementation of digitisation initiatives requires more than technological advancements alone. Stakeholder engagement, capacity building, and the development of supportive policies are vital components of this transition. Ensuring miners, government agencies, and local communities are adequately trained and equipped to navigate digital systems is crucial for the sustainable adoption of these innovations.

Furthermore, the potential challenges and barriers to digitisation, such as digital divide issues, cybersecurity concerns, and the need for robust infrastructure, must be comprehensively addressed. Collaborative efforts between government bodies, industry players, technology providers, and other relevant stakeholders will be instrumental in navigating these challenges.

In conclusion, the digitisation of mining titles in Zimbabwe holds great promise for modernising the sector, increasing efficiency, and fostering a more transparent and accountable environment. By embracing these technological advancements while considering the unique context of Zimbabwe, stakeholders can pave the way for a mining industry that benefits from its rich mineral reserves and contributes to the

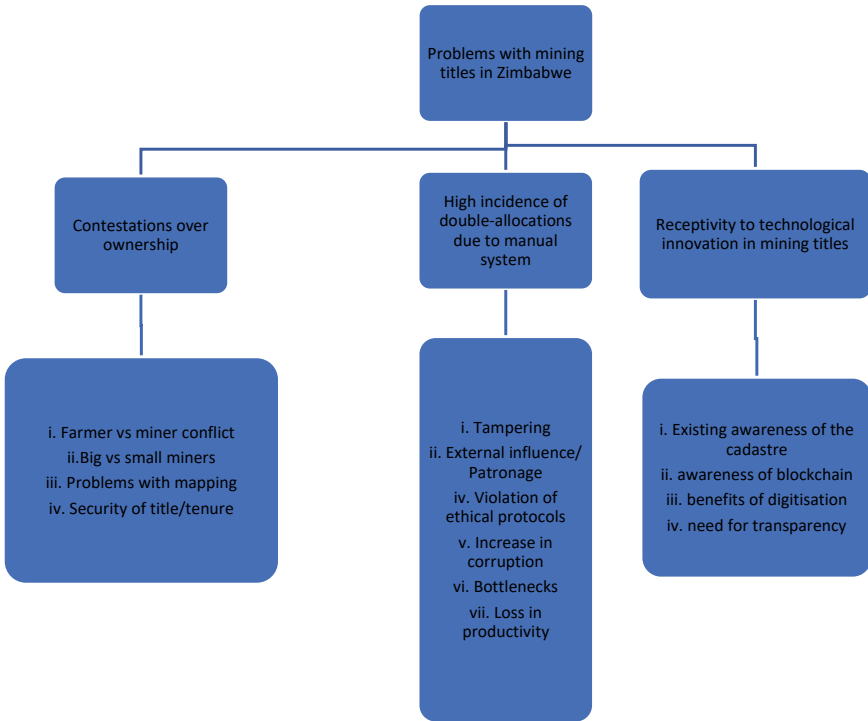
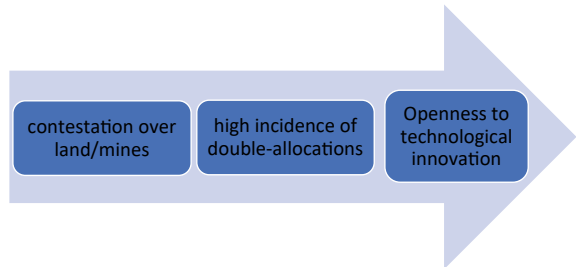


Fig. 5 Qualitative data analysis

Fig. 6 The temporal rearrangement of qualitative data



country’s overall economic growth and development. As Zimbabwe moves towards this digital transformation, careful planning, collaboration, and adaptability will be key to realising the full potential of digitised mining titles.

References

1. Alam KM, Rahman JMA, Tasnim A, Akther A (2020) A blockchain-based land title management system for Bangladesh. *J King Saud Univ—Comput Inf Sci*
2. Asp PJ (2000) Mining and indigenous peoples (special issue). *UNEP's Industry and Environment* 63
3. Bury J (2005) Mining mountains: neoliberalism, land tenure, livelihoods, and the new Peruvian mining industry in Cajamarca. *Environ Planning A: Econ Space* 37(2):221–239. <https://doi.org/10.1068/a371>
4. Castro A, Nielsen E (2001) Indigenous people and co-management: implications for conflict management. *Environ Sci Policy* 4:229–239. [https://doi.org/10.1016/S1462-9011\(01\)00022-3](https://doi.org/10.1016/S1462-9011(01)00022-3)
5. Dale P, McLaughlin J (2011) Land administration. https://doi.org/10.1007/978-94-007-1667-4_15
6. Dorian JP, Humphreys HB (1994) Economic impacts of mining. *Nat Res Forum* 18:17–29. <https://doi.org/10.1111/j.1477-8947.1994.tb00869.x>
7. Ferdous MS, Chowdhury F, Alassafi MO (2019) In search of self-sovereign identity leveraging blockchain technology. *IEEE Access* 7:103059–103079. <https://doi.org/10.1109/access.2019.2931173>
8. Feast B, Davids J, Mills T, DeVries P (2006) Mining cadastre in Africa—lessons learnt. In: 5th FIG regional conference: promoting land administration and good governance
9. Glaser F (2017) Pervasive decentralisation of digital infrastructures: a framework for blockchain enabled system and use case analysis. *HICSS*
10. Goosen M (2023) Top 5 minerals produced in Zimbabwe. *Energy Capital Power*
11. Harper A, Israel A (1999) The killing of the fly: state-corporate victimisation in Papua New Guinea. *Asia Pacific Working Paper 1999/22*. ANU, Australia
12. Heeks R, Stanforth C (2014) Understanding development project implementation: an actor-network perspective. *Public Administration and Development*. 34.<https://doi.org/10.1002/pad.1671>
13. Hettler J, Lehmann B (1995) Environmental impact of large-scale mining in Papua new guinea: mining residue disposal by the Ok Tedi copper–gold mine. *UNEP Publications*, Germany
14. Hilson G (2002) The environmental impact of small-scale gold mining in Ghana: identifying problems and possible solutions. *Geogr J* 168(1):57–72
15. Hilson G, Murck B (2000) Sustainable development in the mining industry: clarifying the corporate perspective. *Resour Policy* 26(4):227–238
16. Iansiti M, Lakhani K (2017) The truth about blockchain. *Harvard Business Rev* 1–11. https://enterpriseproject.com/sites/default/files/the_truth_about_blockchain.pdf
17. Jonhera P (2018) A comparative analysis of the security of mineral tenure under the mines and minerals act. *Dissertation: University of Pretoria*.
18. Knirsch F, Unterweger A, Engel D (2019) Implementing a blockchain from scratch: why, how, and what we learned. *EURASIP J Inf Secur* 2019. <https://doi.org/10.1186/s13635-019-0085-3>
19. Latour B (1987) *Science in action*. Open University Press
20. Lemieux A (1997) Canada's global mining presence. In: *Canadian minerals yearbook*. Natural Resources Canada, Ottawa
21. Marwick AE (2013) Status update: celebrity, publicity, and branding in the social media age. *Celebrity, Publicity, and Branding in the Social Media Age, Status Update*, pp 1–36
22. Mbendi (1999) Information for AfricaFAfrican Mining Statistics. Africa's leading business website. <http://www.mbendi.co.za/indy/ming/mingaf.htm#Overview>
23. McKillop J, Brown AL (1999) Linking project appraisal and development: the performance of EIA in large-scale mining projects. *J Environ Assess Policy Manage* 1(4): 407–428
24. McLeod H (2000) Compensation for landowners affected by mineral development: the Fijian experience. *Resour Policy* 26:115–125
25. MiningWatch (2000) On the ground research: a workshop to identify the research needs of communities affected by largescale mining. *MiningWatch Canada and the Canadian consortium for international social development (CCISD)*, Ottawa

26. Morris DZ (2016) Leaderless, blockchain-based venture capital fund raises \$100 million, and counting, *fortune* (magazine). <http://fortune.com/2016/05/15/leaderlessblockchain-vc-fund/>
27. Mougayar W (2015) Blockchain 2015: strategic analysis in financial services
28. Nakamoto S (2008) Bitcoin: a peer-to-peer electronic cash system. <https://bitcoin.org/bitcoin.pdf>
29. Nyaga RM (2020) Access and use of records through the cadastre system at the ministry of mining headquarters, Nairobi. *Inf Knowl Manag* 10(2)
30. Ochara NM (2010) Assessing irreversibility of an e-government project in Kenya: implication for governance. *Gov Inf Q* 27(1):89–97. <https://doi.org/10.1016/J.GIQ.2009.04.005>
31. Ochieng Odhiambo M (2000) Karamoja conflict study: a report. Oxfam, Kampala
32. Ølnes S, Ubacht J, Janssen M (2017) Blockchain in government: benefits and implications of distributed ledger technology for information sharing. *Gov Inf Q* 34:3
33. Ozah AP, Wever T, Weissmann T, Ghys L (2011) Prospects, challenges and strategies in the implementation of the Nigerian computerised mining information system. *J Earth Sci Eng* 1(3)
34. Patil P, Narayankar P, Narayan DG, Meena SM (2016) A comprehensive evaluation of cryptographic algorithms: DES, 3DES, AES, RSA and blowfish. *Procedia Comput Sci* 78: 617–624. <https://doi.org/10.1016/j.procs.2016.02.108>
35. Sedlmeir J, Smethurst R, Rieger A et al (2021) Digital identities and verifiable credentials. *Bus Inf Syst Eng* 63:603–613. <https://doi.org/10.1007/s12599-021-00722-y>
36. Stanforth C (2006) Analysing eGovernment implementation in developing countries using actor-network theory. iGovernment. Manchester, UK
37. Swales S, Storey AW, Bakowa KA (2000) Temporal and spatial variations in fish catches in the fly river system in Papua New Guinea and the possible effects of the Ok Tedi copper mine. *Environ Biol Fishes* 57:75–95
38. Sweeting AR, Clark AP (2000) *Lightening the lode: a guide to responsible large-scale mining*. Business and Policy Group, Conservation International, Washington
39. Turkle S (1995) *Life on the screen: identity in the age of the internet*. MIT Press, Cambridge, MA
40. Vikkelsø S (2007) Description as intervention: engagement and resistance in actor-network analyses. *Sci Culture* 16(3):297–309. <https://doi.org/10.1080/09505430701568701>
41. Whitley E, Rukanova B (2008) A symmetric analysis of the border control information systems for people and trade. In: 16th European conference on information systems, ECIS 2008

Cybersecurity at the Core: A Study on IT Experts' Policy Adherence



Rao Faizan Ali , Hamid Jahankhani , and Bilal Hassan 

Abstract Information security remains a significant concern for virtually every well-established organization globally. Extensive research indicates that a significant proportion of information security breaches can be attributed to internal employees' disregard for information security policies. Non-compliance with these policies is a complex issue that necessitates both administrative and behavioural solutions. While numerous studies have delved into behavioural aspects of information security, most of this research has centered around non-IT or non-specialized users. This research paper represents a pioneering pilot study aimed at assessing the information security policy compliance of IT professionals. Formulated hypotheses based on a comprehensive literature review, along with the development of a framework, underpin the study's methodology. The framework incorporates organizational management constructs and draws from two prominent behavioural theories—Protection Motivation Theory and the Theory of Planned Behaviour. The findings from this pilot study underscore the role of organizational management in augmenting employees' protection motivation, ultimately fostering a culture of responsible information security behaviour aligned with information security policy compliance.

Keywords Protection motivation · Information security policy · Information security behaviour

R. F. Ali (✉) · B. Hassan
University of Management and Technology, Lahore, Pakistan
e-mail: rao_16001107@utp.edu.my

R. F. Ali
Universiti Teknologi PETRONAS, Perak, Malaysia

H. Jahankhani · B. Hassan
Faculty of Engineering and Environment, Northumbria University, London, United Kingdom

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_20

1 Introduction

Information security is considered as the most critical aspect of information technology. Organizations related to all businesses are investing in information security for their data and operations safety. Most organizations consider and invest in technical controls, but eighty percent of data and security breaches occur because of human negligence. According to the Verizon cybersecurity investigation report 2018, almost 53,000 security attacks were reported from which 2216 were confirmed data breaches. The report only published the reported attacks; many organizations never confirm data breaches because of fear of losing reputation in the market [1]. According to Verizon insiders, behavioral misuse is one of the prominent reasons for security breaches. In 2017 human negligence was considered the second top reason for security breaches. Insider misuse of privileges, lack of information security awareness, and deviant behavior from the information security policies are considered crucial factors behind any data breach [2].

Information security in organizations incorporates a very complicated process that involves many factors, such as education, the human element, and technology, which are necessary to be managed under one security model. Structuring this type of framework with all these processes according to information security standards and policies such as enterprise information security studies progressed rapidly worldwide [3]. Security policies are the set of written rules which define security levels in organizations, and all employees must comply with these guidelines [4]. Enterprise information security policies consist of rules and regulations; it is just like a rule book that every legitimate user of the information system should follow. Information security policies are different for every organization; most policies include aims and goals, protection of security controls, instruction of use, protection, and disruption of enterprise's information entities. An organization needs to elaborate its security requirements, and defining requirements will help to determine necessary management actions and prerequisites to manage information security controls [3].

The purpose of this research is to evaluate and examine information security policy compliance (ISPC) of IT-based employees. A research framework has been designed with the help of behavioral theories to determine compliance. It is a pilot study to evaluate the research model and instrument analysis. The next section consists of the back-ground, research framework, and hypothesis described in section three. The research method and results are demonstrated in sections four and five, respectively.

2 Background and Motivations

The understanding between information technology and employee behavior is the critical aspect to establish adequate information security [5–7]. Information security culture is another vital factor; reasonable information security culture cultivates better ISPC. Employees' better understanding of information security and knowledge

sharing among each other creates good security culture. Organizational management can enhance information security culture, security education and training, effective ways of implementing security policy and procedure, and good security governance.

To protect information systems, the standards ISO 27000, ISO 27001, and ISO 27002 provide requirements and guidelines, specific controls, and control objectives with which the organization can accomplish enough information security [8]. The International Organization for Standardization publishes these standards in conjunction with the International Electrotechnical Commission. It is challenging for each enterprise to select the most suitable set of standards that fulfill their needs [9]. The ISO 27000, 27,001, and 27,002 standards form a Model to plan and control ISMS. These organizations are offered the chance to align their IT processes and methods for assuring an adequate degree of information security with an international standard [8].

In previous information security behavior studies, the research sample consists of business, management, or other field participants. Most of the research presented in behavioral information security has been performed upon non-IT professionals. Simultaneously, some of the research performed on IS (information systems) specialist's dataset. One of the research studies conducted on a different group of employees to assess role values resulting in IT employees had more complacent behavior than non-IT employees [10]. Similarly, [11] included IS specialists and management employees, but they never distinguish between the two groups. [12] also studied protection motivation on information security personals with insiders. However, except for very few studies, there is significantly less research focused on IT staff and specialized users.

3 Research Methodology

ISPC is a multidimensional problem, and for solving such a problem, a comprehensive framework is needed. So, we proposed a research framework based on organizational management and two behavioral theories. Organizational management is responsible for the enforcement of information security policies and procedures. Management provides all the festive activities which can cultivate good security culture in an organization. For our research framework, the researcher has selected three management constructs (Security education and training programs (SETA), security policy and procedures (SPP), and workplace capabilities (WPC). Two behavior theories were selected for this research first protection motivation theory (PMT), and the second is the theory of planned behavior (TPB).

The protection motivation theory initially presented by Rogers R in 1975; the protection motivation theory was created for fear appeals [13]. PMT consisted of two appraisals threat appraisal and coping appraisal. Threat appraisal addresses how people react when they get a threatening situation, and coping appraisal describes how people cope with the occurred threat. PMT describes people protect themselves from

any threatening situation in four ways; the perceived probability of the threat occurrence, perceived severity of a threatening event, recommended preventive behavior efficacy, and perceived self-efficacy. With the help of a literature review in this paper, perceived vulnerability (PV), perceived severity (PS), response efficacy (RE), and self-efficacy (SE), has been selected from PMT.

Icek Ajzen initially presented the theory of planned behavior (TPB) to improve the theory of reasoned action in 1991 [14]. TPB is used in various behavioral information security studies. It is the most used theory for behavior prediction in the last decade. The theory of planned behavior describes how an individual develops his intention towards actual behavior. TPB also describes the norms and perceived behavioral control effect in the actual behavior of an individual. TPB argues that an individual's actual behavior is shaped by four components attitude (AT), perceived behavioral control (PBC), subjective norms, and intention.

This research paper is an effort to construct and test a theoretical framework. This suggests that organizational management can enhance protection motivation, and adequate protection motivation can improve employees' security behavior. For organizational management researcher has taken SETA programs, SETA programs proved to be very influential in enhancing user information security behaviors [15]. The second construct in organizational management is SPP. The research argues that in-line security policy and procedures can enhance protection motivation among employees [16]. Third, work-place capabilities (job security, job satisfaction, etc.) also significantly affect employees' behaviors towards ISPC [17–19]. Hence this research proposes:

Hypothesis 1

H1a: SETA programs positively impact employees' perception of vulnerability (PV)

H1b: SETA programs positively influence employees' perceived severity (PS).

H1c: SETA programs positively impact employees' response efficacy (RE). H1d: SETA programs have a positive influence on employees' self-efficacy (SE)

Hypothesis 2

H2a: Security Policies and Procedures (SPP) have a positive impact on employees' perception of vulnerability (PV).

H2b: Security Policies and Procedures (SPP) positively influence employees' perception of severity (PS).

H2c: Security Policies and Procedures (SPP) positively impact employees' response efficacy (RE).

H2d: Security Policies and Procedures (SPP) have a positive influence on employees' self-efficacy (SE).

Hypothesis 3

H3a: Workplace Capabilities (WPC) has a positive impact on employees' perception of vulnerability (PV).

H3b: Workplace Capabilities (WPC) positively influences employees' perception of severity (PS).

H3c: Workplace Capabilities (WPC) positively impacts employees' response efficacy (RE).

H3d: Workplace Capabilities (WPC) has a positive influence on employees' self-efficacy (SE).

Hypothesis 4

H4: Perceived vulnerability (PV) influences Intention towards the security policy compliance.

Hypothesis 5

H5: Perceived severity (PS) influences Intention towards the security policy compliance.

Hypothesis 6

H6: Response efficacy (RE) influences Intention towards the security policy compliance.

Hypothesis 7

H7: Self-efficacy (SE) influences Intention towards the security policy compliance
Hypothesis 8.

Hypothesis 8

H8: Attitude (AT) influences Intention towards the security policy compliance.

Hypothesis 9

H9: Perceived Behavioral Control (PBC) influences Intention towards security policy compliance.

This research employed a survey-based methodology, and for such studies, a self-administered questionnaire proves to be the most suitable approach [20]. The questionnaire served as the primary tool for data collection, and constructs within the instruments were adapted from various reputable studies. Specifically, questions related to SETA, SPP, and WPC were sourced from [21–23], while queries regarding perceived vulnerability (PV), perceived severity (PS), response efficacy (RE), and self-efficacy (SE) were drawn from [16]. Additionally, questions concerning attitudes, perceived behavioral control, and intentions towards security policy were adapted from [21–23]. Responses were collected using a 1–5 Likert scale, where 1 denoted “strongly disagree” and 5 signified “strongly agree.” The survey was administered within an IT firm located in Perak, Malaysia, through Google Forms.

All the participants were informed about this research, and all the questionnaires were filled by the participants own will and choice. There were no controlled environment or predetermined rules for filling the questionnaire. Total of 53 people participated in this research. After data screening, 4 out of 53 responses were not eligible to include; therefore, 49 usable responses were collected. The convenience

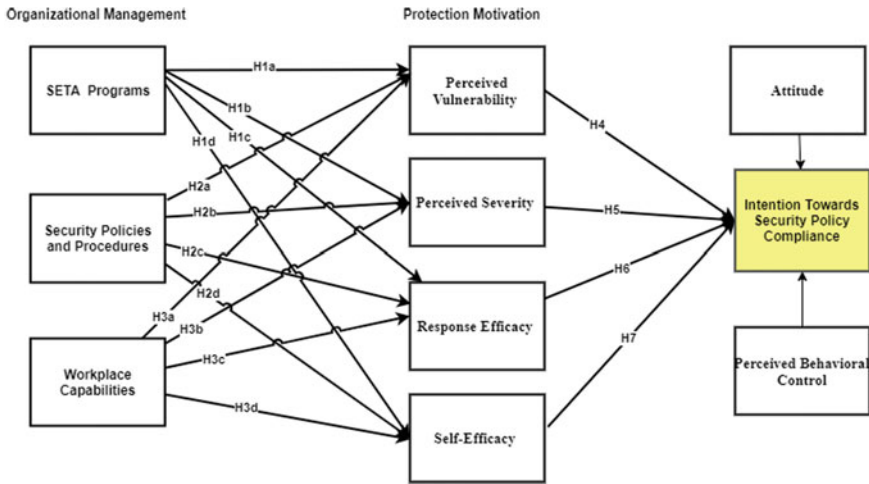


Fig. 1 Proposed research model

sampling technique was used for data collection, as this is the most suitable sampling technique for pilot studies [24]. SPSS V23 was used for descriptive data analysis; four data analysis tests were performed for this paper normality, reliability, correlation, and factor analysis (Fig. 1).

4 Results and Discussion

The normality test determines the variance and outliers in the data. If data is not in good shape, then outliers should be removed from the data. Data normality test is determined by two fundamental values Kurtosis and Skewness. If the values of Kurtosis and Skewness are between ± 2.58 , then the data is considered normal, but data is considered abnormal if the values are above or less than this value. The data for this research was found to be normal.

Factor analysis is used to reduce the larger dimensions into smaller dimensions of factors [25]. For this research, we have used EFA and principal axis factor technique (PFA) with Promax rotation. Factor loadings of factors determine whether this factor is reliable for future use or not if a factor's value should be greater than 0.5. A detailed factor analysis was performed for this study, and factors with low values removed 43 items, and 10 elements were finalized for the primary research.

After the factor analysis reliability test is performed upon data, the reliability test determines the internal consistency between the elements, and reliability is denoted by Cronbach's Alpha [26]. The threshold values for reliability suggested by Ho et al. showed in Table 1; if the Cronbach's Alpha value is less than 0.7, it is considered the weak measurement tool, and the researcher should not rely on the questionnaire [20].

Table 1 Ho et al. Threshold Values

Cronbach's alpha	Values
Value > 0.7	Acceptable
Value > 0.8	Good
Value > 0.9	Excellent

Items with low Cronbach's Alpha values are considered bad for future research and keeping those values risk the questionnaire's reliability [27]. So, we removed low-value items for the primary research; all the items shown good reliability values except 2 items; the reliability results are shown in Table 2.

Finally, the researcher has conducted a correlation test upon the data; Table 3 presented the correlation results. The correlation test determines how much one factor is correlated with another. If two values are significantly correlated, it means both the factors have some commonness in them. For this paper, the researcher conducted a Pearson correlation test, the Pearson coefficient denoted by 'r.' The range of r is between -1 and +1; the value near +1 is considered a significant relationship, and the value near -1 is considered a negative or weak relationship [20].

Based on correlation results, SETA Programs have significant effects on perceived severity and self-efficacy; SETA programs can enhance an employee's perceived severity and self-efficacy. Security policy and procedures affect positively perceived vulnerability and perceived severity about threats; moreover, SPP has shown significance towards self-efficacy. Workplace capabilities have shown a positive impact on perceived severity, response efficacy, and self-efficacy. Moreover, all the PMT constructs have demonstrated a good correlation with ISPC. The perceived vulnerability has shown a correlation of 9.10; it depicts that organizational management can

Table 2 Reliability analysis

Construct	No of items	Cronbach's alpha	Revised no items	Revised Cronbach's alpha
SETA	5	0.905	5	0.905
SPP	5	0.835	4	0.825
WPC	5	0.905	4	0.905
PV	5	0.923	4	0.923
PS	4	0.930	4	0.930
RE	4	0.812	3	0.805
SE	4	0.845	4	0.845
ATT	4	0.920	4	0.920
PBC	4	0.882	4	0.882
ISPC	5	0.915	4	0.915
Overall reliability	45	0.952	43	0.943

Table 3 Correlation analysis

	SETA	SPP	WPC	PV	PS	RE	SE	ATT	PBC	ISPC
SETA	1									
SPP	0.815	1								
WPC	0.765**	0.787**	1							
PV	0.625**	0.748**	0.715**	1						
PS	0.742**	0.731**	0.843**	0.706**	1					
RE	0.653**	0.662**	0.769**	0.723**	0.512**	1				
SE	0.877**	0.717**	0.712**	0.584**	0.757**	0.578**	1			
ATT	0.711**	0.678**	0.673**	0.835**	0.735**	0.665**	0.722**	1		
PBC	0.746**	0.724**	0.701**	0.906**	0.741**	0.716**	0.748**	0.933**	1	
ISPC	0.654**	0.713**	0.744**	0.910**	0.770**	0.598**	0.608**	0.789**	0.840**	1

N = 49 correlation is significant at the 0.01 level (2-tailed)

Note Bold values demonstrate hypothesized associations in the research framework

Table 4 Summary of hypothesis

No	Hypotheses	Decision
H1a	SETA programs positively impact employees' perception of vulnerability (PV)	Not supported
H1b	SETA programs positively influence employees' perceived severity (PS)	Supported
H1c	SETA programs positively impact employees' response efficacy (RE)	Not supported
H1d	SETA programs have a positive influence on employees' self-efficacy (SE)	Supported
H2a	Security policies and procedures (SPP) have a positive impact on employees' perception of vulnerability (PV)	Supported
H2b	Security policies and procedures (SPP) positively influence employees' perception of severity (PS)	Supported
H2c	Security policies and procedures (SPP) positively impact employees' response efficacy (RE)	Not supported
H2d	Security policies and procedures (SPP) have a positive influence on employees' self-efficacy (SE)	Supported
H3a	Workplace capabilities (WPC) has a positive impact on employees' perception of vulnerability (PV)	Not supported
H3b	Workplace capabilities (WPC) positively influences employees' perception of severity (PS)	Supported
H3c	Workplace capabilities (WPC) positively impacts employees' response efficacy (RE)	Supported

(continued)

Table 4 (continued)

No	Hypotheses	Decision
H3d	Workplace capabilities (WPC) has a positive influence on employees' self-efficacy (SE)	Supported
H4	Perceived vulnerability (PV) influences Intention towards the security policy compliance	Supported
H5	Perceived severity (PS) influences Intention towards the security policy compliance	Supported
H6	Response efficacy (RE) influences Intention towards the security policy compliance	Supported
H7	Self-efficacy (SE) influences Intention towards the security policy compliance	Supported
H8	Attitude (AT) influences Intention towards the security policy compliance	Supported
H9	Perceived behavioral control (PBC) influences Intention towards security policy compliance	Supported

enhance perceived vulnerability, and later on, it can help cultivate good ISPC attitude, and perceived behavioral control also shown a positive correlation with ISPC. Correlation results determine all the hypotheses had strong associations except H1a, H1c, H2c, and H3a. So the researcher intended to remove these hypotheses from the main study. A brief summary of hypothesis has been shown in Table 4.

5 Conclusion

This study was conducted for evaluating ISPC from IT professionals; we have determined that organizational management can enhance protection motivation among individuals, and enhanced protection motivation is very effective towards ISPC. The researcher conducted the study in an IT firm; all the respondents almost familiar with information security and its policies. The study results showed that organizational management can improve the IT user's protection motivation, and enhanced protection motivation can reduce employees' deviant Information security behaviors.

In the future, the researcher intended to experiment with a bigger dataset of IT professionals of Malaysia. For getting more insights into the problem researcher aim to collect data above 250 IT professionals. Structural equation modeling will be used for results and data analysis of the main study.

References

1. Widup S, Spittler M, Hylender D, Bassett G (2018) Verizon data breach investigations report. Retrieved from Verizon data breach investigations report
2. Willison R, Warkentin M (2013) Beyond deterrence: an expanded view of employee computer abuse. *MIS Q* 37(1):1–20
3. Yildirim EY, Akalp G, Aytac S, Bayram N (2011) Factors influencing information security management in small-and medium-sized enterprises: a case study from Turkey. *Int J Inf Manage* 31(4):360–365
4. Cresson Wood C (2005) *Information security policies made easy*. Information Shield Publisher, US
5. Wood CC (1997) Policies alone do not constitute a sufficient awareness effort. *Comput Fraud Secur* 14–19
6. Willison RA (2002) *Opportunities for computer abuse: assessing a crime-specific approach in the case of Barings Bank*. London School of Economics and Political Science, United Kingdom
7. Shostack A, Stewart A (2008) *The new school of information security*. Pearson Education, NY, US
8. Disterer G (2013) ISO/IEC 27000, 27001 and 27002 for information security management. *J Inf Secur* 4(1):92–100
9. Năstase P, Năstase F, Ionescu C (2009) Challenges generated by the implementation of the IT standards CobiT 4.1, ITIL v3 and ISO/IEC 27002 in enterprises. *Econ Comput Econ Cybern Stud Res* 43(1): 1–16
10. Moody GD, Siponen M, Pahnla S (2018) Toward a unified model of information security policy compliance. *MIS Q* 42(1):285–312
11. Ifinedo P (2014) Information systems security policy compliance: an empirical study of the effects of socialization, influence, and cognition. *Inf Manage* 51(1):69–79
12. Posey C, Roberts TL, Lowry PB, Hightower RT (2014) Bridging the divide: a qualitative comparison of information security thought patterns between information security professionals and ordinary organizational insiders. *Inf Manage* 51(5):551–567
13. Rogers RW (1975) A protection motivation theory of fear appeals and attitude change. *J Psychol* 91(1):93–114
14. Beck L, Ajzen I (1991) Predicting dishonest actions using the theory of planned behavior. *J Res Pers* 25(3):285–301
15. Yaokumah W, Walker DO, Kumah P (2019) SETA and security behavior: mediating role of employee relations, monitoring, and accountability. *J Glob Inf Manage* 27(2):102–121
16. Hina S, Selvam DDDP, Lowry PB (2019) Institutional governance and protection motivation: theoretical insights into shaping employees' security compliance behavior in higher education institutions in the developing world. *Comput Secur* 87(November): 101594
17. Da Veiga A, Martins N (2017) Defining and identifying dominant information security cultures and subcultures. *Comput Secur* 70(September):72–94
18. Furnell S, Rajendran A (2012) Understanding the influences on information security behavior. *Comput Fraud Secur* 2012(3, March): 12–15
19. Höne K, Eloff JHP (2002) Information security policy—what do international information security standards say? *Comput Secur* 21(5, October): 402–409
20. Madsen HO, Krenk S, Lind NC (2006) *Methods of structural safety*. Dover Publications, NY, US
21. D'Arcy J, Hovav A, Galletta D (2009) User awareness of security countermeasures and its impact on information systems misuse: a deterrence approach. *Inf Syst Res* 20(1):79–98
22. Hamid HA, Dali NRSM (2019) Curbing misbehavior with information security measures: an empirical evidence from a case study. *AL-'ABQARI: J Islamic Soc Sci Human* 17(1): 28–38
23. Hina S, Dominic PDD (2018) Information security policies' compliance: a perspective for higher education institutions. *J Comput Inf Syst* 1(March):201–211
24. Etikan I, Musa SA, Alkassim RS (2016) Comparison of convenience sampling and purposive sampling. *Am J Theor Appl Stat* 5(1):1–4

25. Williams B, Onsman A, Brown T (2010) Exploratory factor analysis: a five-step guide for novices. *Australasian J Paramedicine* 8(3):1–13
26. Field A (2017) *Discovering statistics using IBM SPSS statistics: North American edition*. Sage, 5 edn, US
27. Ali RF, Dominic PDD (2022) Investigation of information security policy violations among oil and gas employees: a security-related stress and avoidance coping perspective. *J Inf Sci*

Metaverse Application Forensics: Unravelling the Virtual Truth



M. A. Hannan Bin Azhar and Oliver Rush-Gadsby

Abstract Virtual reality (VR) has gained significant popularity and recognition in recent years due to its immersive experiences and widespread adoption. However, along with the positive impact it brings, VR technology also attracts individuals with malicious intent who may engage in illegal activities. Consequently, there is a growing need for enhanced forensic methodologies and practices to address digital evidence within VR environments. This paper aims to investigate the production of digital evidence during typical usage of VR, employing a forensically sound methodology. The primary findings of this investigation encompass files and system memory data that provide insights into the hardware utilised, a chronological timeline of usage, and, in certain instances, indications of specific actions performed within the VR environment.

Keywords Virtual reality · Metaverse · Memory forensics · Disk forensics

1 Introduction

Virtual reality is a rapidly expanding and extensive technological domain, evident from its increasing market share and size. Projections indicate that the combined market size of augmented reality and virtual reality is expected to reach 250 billion U.S. dollars by 2028, representing an average annual increase of approximately 41.6 billion U.S. dollars, considering its value was 28 billion U.S. dollars in 2021 [1]. These figures suggest the potential for virtual reality to become even more widely adopted in the future. Additionally, it is anticipated that shipments of augmented

M. A. Hannan Bin Azhar (✉) · O. Rush-Gadsby
School of Engineering, Technology and Design Canterbury, Christ Church University Canterbury,
Canterbury, UK
e-mail: hannan.azhar@canterbury.ac.uk

O. Rush-Gadsby
e-mail: o.rushgadsby56@canterbury.ac.uk

reality and virtual reality equipment will experience a significant surge between 2023 and 2026, with an expected annual growth rate of 31.5% in sales.

Within the realm of virtual reality (VR), various environments and social spaces mimic real-world activities such as socialising, work, and gaming. Among these applications, “VRChat” stands out as a particularly popular instance, boasting an estimated 66.3 thousand daily players in November 2022 and a cumulative player base of 7.4 million [2]. According to 99Firms [3], individuals aged 16 to 34 exhibit a higher propensity to engage with virtual reality compared to other age groups. Specifically, 34% of 16 to 24-year-olds and 35% of 25 to 34-year-olds have reported using VR. The widespread adoption of VR applications can be attributed to their accessibility to a broad range of users. For instance, in the case of VRChat, users can utilise various VR equipment from companies such as HTC, Oculus, or Valve, or simply rely on a desktop computer with an internet connection [4]. The versatility of compatible devices contributes to the increasing user base and popularity of VR applications.

Despite the initial purpose of virtual reality social spaces to provide a secure environment for work and entertainment, it is essential to acknowledge the potential for malicious activities and crimes within these platforms. This issue parallels similar challenges encountered in other domains of the internet and technology. As virtual reality technology continues to gain widespread adoption, the potential for misuse becomes a significant concern. To date, no public arrests have been made specifically for offenses committed within virtual environments. However, there are reports [5] highlighting incidents where individuals have experienced unwelcome physical contact, illustrating the manifestation of real-world problems and criminal behaviors within the realm of virtual reality. This underscores the importance of addressing and mitigating the potential risks and harms associated with the misuse of virtual reality platforms, as the technology becomes increasingly prevalent.

Limited studies have been identified that specifically address the forensic analysis of VR social applications. Reference [6] presents their investigation as the inaugural publication focusing on the forensic artefacts derived from end-users’ devices and network-related artefacts specifically for the HTC Vive and Oculus Rift VR systems. The study acknowledges the lack of comparable examples and the limited exploration of these forensic practices in the existing literature. In contrast, another study [7] takes a different approach by aiming to acquire artefacts residing in volatile memory, such as RAM, circumventing any potential system malware that could hinder the methods employed in the initial study. In [6] numerous artefacts pertaining to user data were discovered, including user IDs, information about the hardware devices employed, user activity within the application, and network activity during the relevant period. As a result, this study successfully obtained information regarding gesture tracking devices and their respective locations, establishing pathways to the associated data structures.

The research presented in this paper was motivated by the prevailing uncertainties surrounding virtual reality and the metaverse, making it an intriguing area for investigation. The primary focus of this paper revolves around virtual reality applications

in general, with a specific emphasis on investigating a subset known as the “Metaverse.” The aim is to explore and present ideas and findings regarding the forensic analysis of these virtual environments. Metaverse applications are virtual, decentralised environments that can be accessed through virtual reality technology. They strive to enable global communication and social interaction among individuals from various locations, serving as a platform for effective and efficient collaboration in diverse fields such as work, entertainment, and socialisation. The Metaverse reportedly integrates augmented reality, virtual reality, and blockchain technology to create an immersive and productive environment for users [8]. By examining the forensic aspects of Metaverse applications, this study seeks to contribute to the understanding of this emerging area and shed light on the potential challenges and opportunities it presents.

2 Methodology

In order to ensure the forensic integrity of the investigation, the methodology employed in this study adheres to rigorous forensic standards, aiming to produce evidence that would be admissible in a court of law if the need were to arise. This is achieved through the utilisation of accepted forensic tools, standard documentation practices, and the application of established forensic methods, all aimed at validating any identified artefacts and potential findings.

In selecting the virtual reality social applications, careful consideration was given to the multitude of options available in the market. The chosen applications and any accompanying software were primarily selected based on their popularity and intended purpose. Consequently, three distinct applications emerged as prominent choices due to their high player count and emphasis on social interaction within varied virtual experiences. These applications, namely “VRChat” [4], “Rec Room” [9], and “Bigscreen Beta” [10] were accessed through the widely-used digital distribution service known as “Steam” [11]. The rationale behind selecting these applications was rooted in the assumption that they would resonate with the average consumer of VR devices, as they consistently featured prominently on the front page when searching for VR metaverse games or social applications.

Table 1 provides a list of the software platforms used in the study, along with their respective versions. The software platform known as “Steam,” developed by Valve Incorporation [11], is widely recognised for its popularity and serves as a prominent distribution platform for various software, particularly video games. Users of Steam have the capability to purchase, install, and enjoy a wide range of video games, including those in the VR genre. Additionally, Steam offers personalised user accounts where installations and other relevant user information are stored [11]. In the context of this study, Steam was chosen as the software platform for several reasons. Firstly, it boasts a substantial user base, with an estimated 132 million monthly active users in 2021 [12]. This suggests that a significant proportion of desktop-based VR users are likely to encounter Steam compared to other software

Table 1 Software used in the experiments

Software name	Build version/ID
Steam	v020
Steam VR	10,835,175
Oculus	v3.1.13
VR chat	11,008,497
Bigscreen beta	11,005,838
Rec room	10,951,302

platforms. Furthermore, Steam was selected due to its provision of a built-in tool that enables users to experience VR applications installed on the platform, at no additional cost. Steam VR, which utilises the Open VR.

API developed by Valve, ensures compatibility between the software and VR equipment offered by prominent manufacturers.

Three distinct metaverse social applications—Rec Room, VRChat, and Bigscreen Beta—were chosen to explore potential variations in outcomes based on the application employed. Rec Room is a VR application that offers users the opportunity to engage in various experiences, whether individually or with others. These experiences encompass activities such as building items, playing games, and socialising within the virtual environment, all facilitated through the use of personalised virtual avatars [9]. Depending on their preferences, users are presented with multiple game rooms to join. At the time of this study, Rec Room reportedly averaged approximately 3,800 concurrent players on Steam [13]. However, it is important to note that the application is available on multiple platforms, including consoles and mobile devices, and supports cross-platform play, indicating that the user count may be higher than the aforementioned figure.

Similarly, VRChat adopts a comparable concept, enabling users to interact and communicate with one another during various experiences within the application. VRChat offers users the ability to present themselves to others in the form of unique full-body avatars. These avatars can either be pre-made or customised using the Unity SDK tools provided by the platform. VRChat provides a diverse range of experiences within the application, including games and social events, and is accessible on multiple platforms such as Steam, Quest, Rift, and Viveport [14]. The application has gained significant popularity as one of the leading metaverse virtual reality experiences, with approximately 20,000 concurrent players on Steam [15]. It is worth noting that VRChat supports cross-play, indicating that the actual number of active players is likely to be higher than the reported figure. Figure 1 provides an illustration of how users may engage in virtual world communication using the integrated text chat feature.

The third social application included in this study is “Bigscreen Beta”. Similar to other applications, Bigscreen Beta offers users an immersive environment for social interaction, industry collaboration, gaming, and movie watching. Its advertising primarily emphasizes the virtual theater setting where users can watch a variety

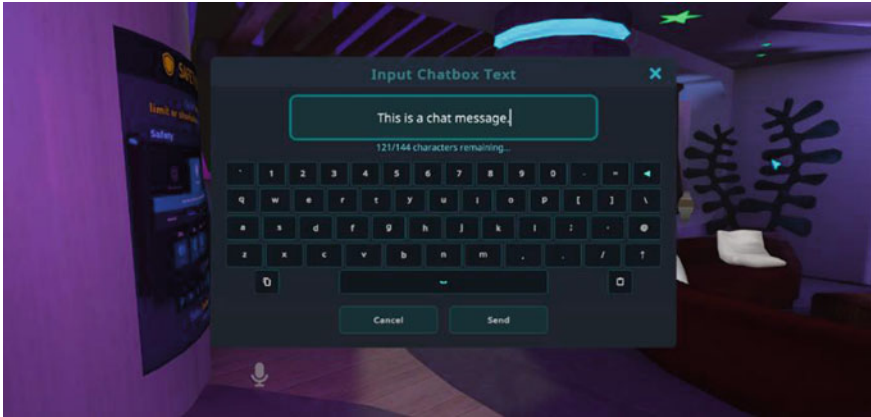


Fig. 1 Example usage of the text chat ability within VRChat

of films with others, as well as enhancing productivity through co-worker collaboration. To provide a comprehensive media viewing experience, Bigscreen Beta enables users to access streaming services, use a remote desktop client, rent movies through partnerships, and utilise a built-in video player. These features can be enjoyed in a social environment, further enhancing the overall experience [10].

Table 2 presents a comprehensive list of the technical hardware equipment utilised in the methodology of this study. The hardware includes a desktop computer with its corresponding specifications, as well as the virtual reality equipment employed to interact with the chosen social applications. The virtual reality equipment employed in this study is the Oculus Rift CV1 [16]. In addition to the headset, the fully functional setup includes two controllers and two sensors. The handheld controllers, powered by single batteries, serve as the primary means through which the user engages with the virtual environment, offering tracking capabilities, buttons, and triggers for interaction.

Table 2 Hardware configurations

Hardware	Description
Manufacturer	ASUS
Form factor	Desktop
CPU	AMD Ryzen 5 5600X 6-Core 4.2 GHz
GPU	NVIDIA GeForce GTX 1070 8 GB
OS	Windows 10 Pro, Version: 22H2
RAM	16 GB
Storage	256 GB SSD—Samsung
VR headset	Oculus Rift CV1

Fig. 2 Oculus Rift CV1 headset and its controllers



Figure 2 provides an illustrative example of the virtual reality equipment employed. The two sensors are positioned on the desk in front of the user, approximately two meters apart. These sensors enable precise tracking of the user's position by detecting the infrared LEDs embedded in the virtual reality headset and controllers. Power connections are established using a USB 3.0 and HDMI port for the headset, while each sensor requires a USB 3.0 or 2.0 port. The desktop computer utilised in this study can be classified as a mid-range device, featuring specifications that fulfill the requirements for the virtual reality equipment to function properly. It was determined that lower specifications might lead to side effects such as motion sickness, particularly during extended usage periods.

In relation to the software and hardware utilised in the methodology of this study, the investigation phase incorporated various forensic tools. The investigation process encompassed three key steps: disk image creation, memory dump creation, and subsequent analysis, all of which required software compatible with the Windows operating system. For the purpose of imaging and assessing evidence, FTK (Forensic Toolkit) Imager, an open-source forensic tool developed by AccessData [17] was used. This tool facilitated the creation of comprehensive disk images, memory dumps, and allowed for data previewing within these images. FTK Imager was selected for this study based on its wide acceptance within the forensic community for disk imaging and memory dump creation. The tool operates in a manner that ensures the preservation of data integrity and forensic soundness throughout the imaging process. By utilising FTK Imager, ensured that no alterations or modifications were made to the storage device, and the software implemented additional measures to verify the integrity of the hard drive data. These precautions safeguarded the integrity and reliability of the investigation and study. Figure 3 provides a visual representation of the process involved in using FTK Imager to create a disk image and memory dump.

The second forensic tool employed in this methodology is Autopsy, an open-source graphical interface designed for The Sleuth Kit developed by Brian Carrier. The Sleuth Kit comprises a collection of command line tools utilised for investigating disk images through the analysis of file system data [18]. Autopsy, acting as a

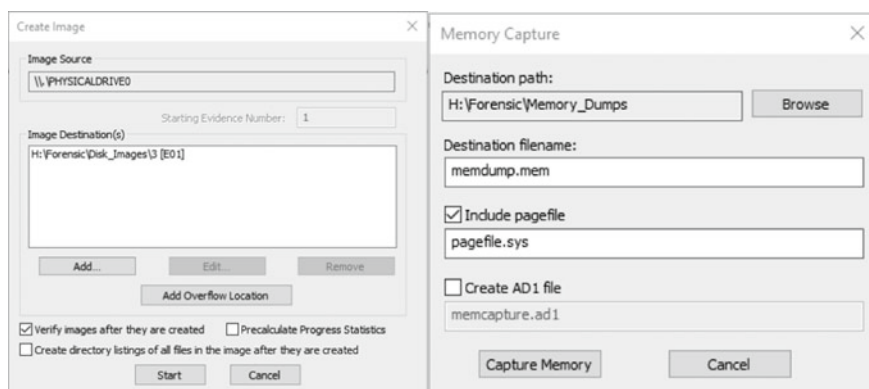


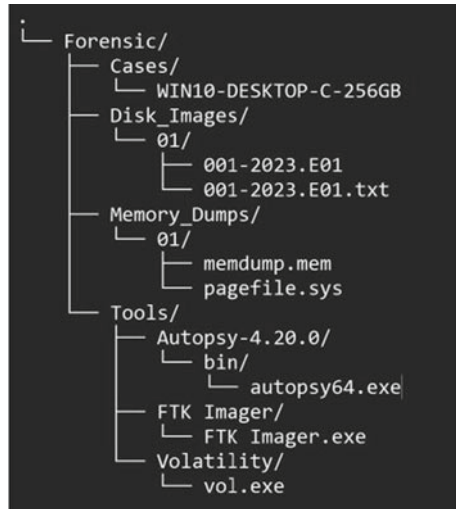
Fig. 3 Using FTK to create disk images and memory dumps

graphical interface for The Sleuth Kit, offers an alternative to terminal-based interaction, allowing for a more user-friendly experience while also providing additional forensic tools. Autopsy serves a wide range of investigation needs, including disk image analysis, recovery of deleted files, keyword search functionality, graphical timeline analysis, and multimedia file viewing [19]. Although this study may not require the utilisation of the full range of available features, it was deemed crucial to employ software that is well-regarded for its forensic utility and its ability to enhance the efficiency and effectiveness of the investigation process.

The third forensic tool utilised in this methodology is Volatility 3, developed by Aaron Walters. Volatility 3 is an open-source tool written in Python, primarily driven by contributions from its community. It specifically focuses on the analysis and investigation of the contents stored in a device's random access memory (RAM) through the examination of memory capture samples, rather than analysing the contents of the hard drives [20]. It is worth mentioning that although Volatility does offer a graphical interface called Volatility Workbench, it was observed that this graphical version lacks some commands and features available in the command line variant. As a result, the command line variant of Volatility was chosen.

In terms of file storage and organisation, this methodology involves the utilisation of two distinct storage devices. The initial storage device, considered the suspect device, is a 256 GB solid-state drive (SSD). This drive contains the operating system files, various installed programs, Steam files, and serves as the location for installing the virtual reality applications. On the other hand, the second storage device is a 2 TB external hard disk drive (HDD), which functions as a forensic device. This external HDD contains the forensic tools, memory dumps, and disk images required for the investigation. The deliberate separation of these storage devices aims to replicate a real-life crime and forensic scenario as closely as possible. For a visual representation of the directory tree illustrating the storage and organisation of the forensic hard disk, refer to Fig. 4.

Fig. 4 Directory tree of forensic hard disk



The following section provides a detailed step-by-step description of the methodology employed in this study. Initially, the selected social applications were launched, and various actions were undertaken within them to generate digital artefacts/evidence that would be produced by someone with malicious intent. These applications shared similar functions, including an introductory interface allowing users to join different experiences within the application. In each application, the following sequence of actions was performed: entering a socialising-focused room or in-game event, engaging in activities such as typing messages in the text chat, using the microphone, interacting with in-game objects, and moving around. This series of events was repeated in a separate room or experience. After approximately 45 min of engagement in each application, they were closed.

Regarding the investigation, a memory dump of the system's 16 GB RAM was created and stored on the external disk. FTK Imager, with its built-in memory capture feature, was utilised to accurately capture the memory and save it to a specific file path on the external hard drive. The creation of a memory capture was prioritised due to the volatile nature of RAM. If the system were powered off before this step, all data in the RAM would be lost and overwritten upon powering on, making data retrieval nearly impossible. It is assumed in this study that the suspect's device was found powered on and accessible, hence the execution of this step.

Subsequently, a physical disk image of the 256 GB solid-state drive was created and saved on the separate external hard drive. Disk imaging, considered non-volatile storage, was conducted as the final step in evidence collection. FTK Imager was employed again to create a byte-level copy of the storage device. A physical image of the drive, rather than a logical image, was produced, copying all partitions, including unused or empty ones, which might contain deleted files and unused data. The disk image was saved in the E01 format, recognised as the standard forensic image file format, providing checksums for data integrity, compression, and a bit-by-bit copy

of the disk. Although Autopsy supports both E01 and dd disk image formats, the E01 format was chosen due to its optimised forensic capabilities, despite the longer time required for copying compared to the dd format. The disk image was then saved on the external hard drive, and the software generated hashes for both the original disk and the disk image, which were compared to ensure they were identical copies for forensic soundness.

Subsequent to the memory capture and disk imaging, the analysis of the memory capture was conducted using the Volatility software. This involved employing various commands to extract multiple artefacts related to the actions performed during the memory capture, such as the system process list, network connections, and current malware processes. Following the memory analysis, the hard drive analysis was carried out using the Autopsy forensic tool. This software was employed to analyze the previously created disk image and search for relevant artefacts related to the use of virtual reality equipment and the metaverse applications. This included examining web browser data, such as browsing history, deleted files, log files, and other files present on the disk.

It should be noted that this methodology does not encompass certain practices employed in other forensic applications. These practices may involve capturing images of the device and its screen, documenting surrounding artefacts (e.g., notepads), and checking for the presence of malicious anti-forensic programs. Additionally, precautions such as directly powering down the device from the wall plug were not taken into account. The chosen methodology assumes that the analysed device is entirely benign in terms of any software or related activities, thereby negating the need for these additional precautions or actions.

3 Experimental Findings

This section is divided into two distinct subsections: RAM and HDD. The RAM subsection presents the potential data and evidence resulting from the analysis of the RAM capture, while the HDD subsection presents the data and evidence obtained from the analysis of the created disk image.

3.1 Results of RAM Analysis

The memory capture sample was initially analysed using Volatility 3, utilising the available commands within the PowerShell terminal on Windows devices. To execute the Volatility tool, the following text (Fig. 5) was entered into the terminal, assuming the user is in the directory where Volatility is installed.

The analysis began by using the 'windows.info' command, which provided essential data regarding the system/device from which the memory capture originated. This


```
python vol.py -f H:\Forensic\Memory_Dumps\01\memdump.mem
```

Fig. 5 Enumerated open Wi-Fi screenshot using ‘aircrack.ng’

data includes crucial information for confirming that the memory capture under investigation is indeed from the suspect’s device. The output of this command includes details such as the system’s date and time when the memory capture was created, the operating system, and the machine type. The captured information, depicted in Fig. 6, serves to establish the integrity of the analysis by confirming the match between the suspect’s machine and the device from which the memory capture originated. Furthermore, the ‘windows.pslist’ command was employed within Volatility to retrieve the list of processes running on the device at the time of capture. In some cases, this command also provides additional details, such as the creation and exit times of the processes, offering timestamps that indicate when the user initiated and terminated specific programs. Figure 7 illustrates the utilisation of this command to display exclusively the processes related to the three social applications investigated in this study: VRChat, Rec Room, and Bigscreen Beta. The output of these commands showcases the date and time stamps corresponding to the execution of the application executables, as observed in the memory, along with the timestamps indicating when the application executables were terminated. Depending on the context, this information could be crucial in establishing the timeline of the user’s interaction with the applications and their duration.

Subsequently, the ‘windows.cmdline’ command was executed within the Volatility framework, as demonstrated in Fig. 8, which depicts the extensive output generated by this command. This command provided valuable data concerning the commands executed on the system throughout the current lifespan of the memory, starting from when the system was powered on. The information presented by this command is useful in revealing the sequence of processes, including their initiation and termination, as well as any manual commands executed by the user.

The purpose of analysing this data is twofold: first, to identify any potentially malicious applications that may have been running on the device, such as those involved in file deletion or similar activities, and second, to determine if the user engaged in any suspicious activities prior to capturing the memory in an attempt to conceal potential evidence through the terminal.

In the context of this study, the ‘windows.cmdline’ command did not indicate any malicious actions performed through the terminal. However, it did reveal that since the system was powered on, various processes related to virtual reality, primarily executed through the Steam platform, had been initiated. This observation suggests that Steam served as the primary distribution channel used by the user to install and run the virtual reality applications.

Lastly, in the memory analysis phase of this study, the ‘windows.hashdump’ command was utilised. This command is employed to generate a “dump” of hashed passwords associated with user accounts on the system, which are stored in the local

```

Variable      Value
-----
Kernel Base   0xf80043800000
DTB           0x11aa000
Symbols file: //H:/Forensic/Tools/volatility3/volatility3
1.json.xz
Is64Bit       True
IsPAE         False
layer_name    0 WindowsIntel32e
memory_layer  1 FileLayer
KdVersionBlock 0xf8004440f398
Major/Minor   15.19041
MachineType   34404
KeNumberProcessors 12
SystemTime    2023-04-16 13:27:50
NtSystemRoot  C:\WINDOWS
NtProductType NtProductWinnt
NtMajorVersion 10
NtMinorVersion 0
PE MajorOperatingSystemVersion 10
PE MinorOperatingSystemVersion 0
PE Machine    34404
PE TimeDateStamp Wed Jul 22 06:49:42 2093
PS H:\Forensic\Tools\volatility3>

```

Fig. 6 System data of memory capture

```

PS H:\Forensic\Tools\volatility3> python vol.py -f H:\Forensic\Memory_Dumps\01\memdump.mem windows.plist | Select-String vrchat
Progress: 100.00 PDB scanning finished
11116 8868 VRChat.exe 0xc0023ffc000 0 - 4 False 2023-04-16 11:49:30.000000 2023-04-16 12:09:54.000000 Disabled
10020 7860 VRChat.exe 0xc002461003000 0 - 4 False 2023-04-16 12:12:05.000000 2023-04-16 12:19:28.000000 Disabled

PS H:\Forensic\Tools\volatility3> python vol.py -f H:\Forensic\Memory_Dumps\01\memdump.mem windows.plist | Select-String recroom
Progress: 100.00 PDB scanning finished
2328 34796 RecRoom_Release 0xc0023abbb000 0 - 4 True 2023-04-16 12:44:56.000000 2023-04-16 12:45:20.000000 Disabled
10740 2328 RecRoom.exe 0xc0024570f300 0 - 4 False 2023-04-16 12:45:11.000000 2023-04-16 13:16:39.000000 Disabled

PS H:\Forensic\Tools\volatility3> python vol.py -f H:\Forensic\Memory_Dumps\01\memdump.mem windows.plist | Select-String bigscreen
Progress: 100.00 PDB scanning finished
34524 34796 BigScreen.exe 0xc0024702d000 0 - 4 False 2023-04-16 13:17:20.000000 2023-04-16 13:25:50.000000 Disabled

```

Fig. 7 Information on processes of VR social applications

```

5488 vrstartup.exe Required memory at 0x11c5f00 is not valid (process exited)
12312 vrserver.exe "C:\Program Files (x86)\Steam\steamapps\common\SteamVR\bin\win64\vrserver.exe" -waitformonitor
10220 vrcompositor.exe "C:\Program Files (x86)\Steam\steamapps\common\SteamVR\bin\win64\vrcompositor.exe"
12956 vrdisplayboard.exe "C:\Program Files (x86)\Steam\steamapps\common\SteamVR\bin\win64\vrdisplayboard.exe"
12956 vrwebhelper.exe "C:\Program Files (x86)\Steam\steamapps\common\SteamVR\bin\vrwebhelper\win64\vrwebhelper.exe"
10410 vrmonitor.exe "C:\Program Files (x86)\Steam\steamapps\common\SteamVR\bin\win64\vrmonitor.exe" -nob3llprocess -startupappid 256810 -startupreason steam_vrstartup -startupapptype 7
14700 vrdisplayboard.exe Required memory at 0x11c5f00 is not valid (process exited)
35160 steamtools.exe "C:\Program Files (x86)\Steam\steamapps\common\SteamVR\tools\steamr_env\environments\game\bin\win64\steamtools.exe" -vr -retail -useappid SteamR40070 -nowindow -vccport 23009
6960 vrwebhelper.exe "C:\Program Files (x86)\Steam\steamapps\common\SteamVR\bin\vrwebhelper\win64\vrwebhelper.exe" --typegpu-process --filed-trial-handle=1824,1219546853595952169,1013185626374992000,131871

```

Fig. 8 System command line entries

machine’s SAM registry section. The resulting hashes, as exemplified in Fig. 9, consist of two types: LM hashes and NT/NTLM hashes. These hashes serve as encrypted representations of the passwords associated with user accounts on the device.

```
Volatility 3 Framework 2.4.2
Progress: 100.00 PDB scanning finished
User rid lmhash nthash
Administrator 500 aad3b435b51404eeaad3b435b51404eeaad3b435b 31d6cfe0d16ae931
Guest 501 aad3b435b51404eeaad3b435b51404eeaad3b435b 31d6cfe0d16ae931
DefaultAccount 503 aad3b435b51404eeaad3b435b51404eeaad3b435b 31d6cfe0d16ae931
WDAGUtilityAccount 504 aad3b435b51404eeaad3b435b51404eeaad3b435b 76b877254583e
Oliver Rush-Gadsby 1001 aad3b435b51404eeaad3b435b51404eeaad3b435b 83c714cc03c0c
```

Fig. 9 System SAM password hashes (partially blurred)

Decrypting these hashes using specialised software could potentially unveil the passwords of the users, thereby providing access to their accounts, such as Steam, and revealing other relevant artefacts. This information can be valuable in understanding the password patterns or gaining access to the suspect’s other accounts on the system.

Although other commands within Volatility were employed to investigate network connections and other aspects, they did not yield any noteworthy findings directly relevant to the use of virtual reality devices and software.

In summary, the memory analysis predominantly yielded artefacts pertaining to the active processes on the system, including their creation and termination timestamps, the command history of processes and users, and the potential to uncover user passwords. These findings offer an overall understanding of the activities occurring on the device at the time of the memory capture, specifically regarding the use of virtual reality software.

3.2 Results of HDD Analysis

The disk image, generated by FTK Imager, underwent analysis using the Autopsy software, which facilitated the examination of digital artefacts by displaying the files and other data contained within the disk image. The software further enabled the exploration of various types of information that would not be readily accessible without its capabilities, including users’ browser history, deleted files, encrypted files, and more. For the purpose of this study, the primary focus was on files generated through the utilisation of virtual reality devices and software. These files held the potential to offer insights into the user’s actions and usage patterns when engaging with these virtual reality devices. The focus of the investigation was to find information regarding user’s equipment usage, account, virtual reality device usage, and the timing of various events such as application launches and microphone usage.

At first the “AppData” folder was explored. This folder houses program configurations and other pertinent information. Within this folder, a directory named “VRChat” was discovered at the following path: C:\Users\User\AppData\Local-Low\VRChat\VRChat. This directory contained several output logs, such as “output_log.txt,” which contained diverse information about the utilised equipment within the VRChat application (Fig. 10). These logs encompassed details about the employed sensors, head-mounted displays, system specifications, controllers, as well

```
2023.04.16 13:04:05 Log - [VRInputManager] Logging Enabled set to: True
2023.04.16 13:04:06 Log - [UserInfoLogger] Environment Info:
VRChat Build: 2023.2.1-1296-3f15e69762-Release
Store: Steam
Platform: standalonewindows
Device Model: System Product Name (ASUS)
Processor Type: AMD Ryzen 5 5600X 6-Core Processor
System Memory Size: 16295MB
Operating System: Windows 10 (10.0.19045) 64bit
Graphics Device Name: NVIDIA GeForce GTX 1070
Graphics Device Version: Direct3D 11.0 [level 11.1]
Graphics Memory Size: 8081MB
Supports Audio: True
XR Device: Oculus Rift CV1
API Status: Available: ("pong")
CloudFront Status: Available: (pong)
Connected controller: (OpenVR Controller(Oculus Rift CV1 (Right Controller))) - Right
Connected controller: (OpenVR Controller(Oculus Rift CV1 (Left Controller))) - Left
```

Fig. 10 “Output_log.txt” file

as specifics regarding the microphone in use and the authentication time for the user on the VRChat servers. Figure 10 presents an example of querying whether the user was banned from the application. While files related to the other two applications in the study, namely Rec Room and Bigscreen Beta, were also present within the AppData folder, they did not yield noteworthy logs similar to those found in VRChat.

Subsequently, attention turned to the “Program Files × 86” directory, which holds significance as the designated location for storing and installing most 32-bit applications. This directory proved to be a rich source of digital artefacts. The investigation commenced with the examination of the “Oculus” folder, as Oculus was the manufacturer of the virtual reality equipment employed. Within this folder, a file named “OVRLibraryService.log” emerged as the sole item of interest. This log file contained records of various actions, including timestamps indicating the initiation of the Oculus software, typically launched when the headset connects to the device. Additionally, the file documented instances of device connections to the Oculus servers, thereby indicating periods of headset usage. Moving forward, the Steam folder within the program files directory was located and scrutinised. This Steam folder emerged as the primary repository for logs pertaining to software usage, featuring two significant subfolders named “logs” and “configs.” Within the “logs” folder, multiple text files were discovered, capturing information related to virtual reality usage through Steam. Initially, the “basestation_wireless.txt” file contained timings associated with the initialisation and shutdown of the “basestations” or lighthouses for the equipment.

Several files named “connection_log*.txt” were generated, containing logs of network connections between the user and the Steam servers, including their corresponding statuses (Fig. 11). Another notable file within the logs directory was “content_log.txt,” which recorded the launched app IDs and tracked their activities, such as exits, updates, and launches. Additionally, multiple files following the format of “vrclient_” followed by the application name (e.g., “vrclient_recroom.txt”) were

```
[2021-12-04 16:30:25] CCMInterface::OnNetworkDeviceStateChange -- Saw device up (192.168.1.153)
[2021-12-04 16:30:25] CCMInterface::OnNetworkDeviceStateChange -- Start connecting to Steam
[2021-12-04 16:30:25] CCMInterface::OnNetworkDeviceStateChange -- Saw device up (192.168.1.153)
[2021-12-04 16:30:25] CCMInterface::OnNetworkDeviceStateChange -- Start connecting to Steam
[2021-12-04 16:30:25] CCMInterface::OnNetworkDeviceStateChange -- Saw device up (192.168.1.153)
[2021-12-04 16:30:25] CCMInterface::OnNetworkDeviceStateChange -- Start connecting to Steam
[2021-12-04 16:30:25] IPv6 HTTP connectivity test (ipv6check-http.steamcontent.com / 0.0.0.0:80 (0.0.0.0:80)) - TIMEOUT
[2021-12-04 16:30:25] IPv6 UDP connectivity test (ipv6check-udp.steamcontent.com) - FAILED, no addresses resolved
[2021-12-04 16:38:09] [0,0] Log session ended
[2021-12-06 15:52:40] IPv6 HTTP connectivity test (ipv6check-http.steamcontent.com / 0.0.0.0:80 (0.0.0.0:80)) - TIMEOUT
[2021-12-06 15:52:41] IPv6 UDP connectivity test (ipv6check-udp.steamcontent.com) - FAILED, no addresses resolved
[2021-12-06 16:01:24] [0,0] Log session ended
```

Fig. 11 "Connection_log.txt" file

discovered. These files were present for Steam, VRChat, Rec Room, Bigscreen Beta, and the VR server. Each of these "vrclient" files shared a consistent format, documenting the application's start time, process ID (PID), shutdown time, and the associated connections between the applications and their respective servers. Furthermore, a file named "vrwebhelper_main.txt" within the logs directory contained logs pertaining to the creation of virtual web browsers within the application, potentially shedding light on the user's interaction with the applications. Lastly, within the logs directory, a file named "vrserver.txt" provided information about the virtual reality devices in use and their statuses, primarily focusing on audio, microphone, and lighthouse devices.

In addition to the previously examined directories, the "config" folder within the Steam directory was also explored. This folder contained two VDF files named "loginusers.vdf" and "config.vdf." The "loginusers.vdf" file provided information about all users who had logged into Steam on the device, including their account names and the timestamps of their last logins (Fig. 12). The "config.vdf" file contained details of user accounts logged into Steam on the device.

Moreover, an examination was conducted to gather information about how users utilised their microphones in the VRChat application. These data comprises details such as the quantity and names of microphones installed on the system, timestamps indicating when these microphones were activated and deactivated, and identification of the active microphone in use within the application.

```
"users"
  "76561199496727110"
    "AccountName" "projectacc_steam"
    "PersonaName" "treffix123"
    "RememberPassword" "1"
    "WantsOfflineMode" "0"
    "SkipOfflineModeWarning" "0"
    "AllowAutoLogin" "1"
    "mostrecent" "1"
    "Timestamp" "1681928821"
```

Fig. 12 "Loginusers.vdf" file

```
2023.04.16 13:09:32 Log - [Behaviour] Microphone device changing to 'System Default'
2023.04.16 13:09:32 Log - [Always] uSpeak: SetInputDevice 0 (2 total) Microphone (Rift Audio)
2023.04.16 13:09:32 Log - Voice DeliveryMode: UnreliableUnsequenced
2023.04.16 13:09:32 Log - [Behaviour] Spent 0.005615234s spawning players.
2023.04.16 13:09:32 Log - [Behaviour] CacheComponents: ParticleSystems 1, AudioSources 0
2023.04.16 13:09:32 Log - Found SDK2 avatar descriptor.
2023.04.16 13:09:32 Log - [Behaviour] Requesting buffered events
2023.04.16 13:09:32 Log - [Behaviour] uSpeak [7][1601267203]: InitializeSettings (owner ID 700002): codec 3, Opus48k, BitRate_24k, FrameDuration_20ms
2023.04.16 13:09:32 Log - [Behaviour] uSpeak [7][1601267203]: UpdateSettings locally, sending to remotes: codec 3, Opus48k, BitRate_24k, FrameDuration_20ms
2023.04.16 13:09:32 Log - [Always] Microphones installed (2 total)
2023.04.16 13:09:32 Log - [Always] -- [0] device name = 'Microphone (Rift Audio)' min/max freq = 48000 / 48000
2023.04.16 13:09:32 Log - [Always] -- [1] device name = 'Headset Microphone (Oculus Virtual Audio Device)' min/max freq = 48000 / 48000
2023.04.16 13:09:32 Log - [Behaviour] uSpeak [7][1601267206]: Stop Microphone:
2023.04.16 13:09:32 Log - [Behaviour] uSpeak [7][1601267206]: Start Microphone: Microphone (Rift Audio)
2023.04.16 13:09:32 Log - Measure Human Avatar Avatar
2023.04.16 13:09:32 Log - eyeToNeck:(0.0, 0.1, 0.1)
2023.04.16 13:09:32 Log - [Beha
```

Fig. 13 Microphone usage

Figure 13 details the microphone data, revealing whether users activated their microphones while using the VRChat application and, if so, the duration of microphone usage. Additionally, it offers insights into moments when users switched their microphones on or off, contributing to a more comprehensive understanding of their interaction with the application.

4 Conclusions

This paper employed a forensic methodology to analyse data from a device, encompassing both hard drive and memory analysis using various forensic tools. When comparing findings with previous literature on the topic [6, 7], notable similarities and differences emerge. The similarities are primarily observed in the HDD investigation, where some of the same files identified here were also documented in previous studies. This can be attributed to the use of Steam software in both cases. However, in the HDD investigation, the utilisation of a different VR application resulted in the creation of distinct files. On the other hand, when comparing the RAM investigation findings, significant differences arise between this study and previous literature [7] on VR volatile memory forensics. Previous studies focused on reconstructing the virtual environment through memory artefacts, while this study prioritised the collection of user-related artefacts stored in memory, yielding distinct outcomes such as understanding user activities and identifying the device in use.

For further research on this topic, it is recommended to expand the scope by including a broader range of popular metaverse applications and distributions at the time. Overall, the findings in this paper highlight the efficacy of forensic tools and techniques in identifying and retrieving significant evidential artefacts associated with the use of virtual reality in the context of criminal investigations.

References

1. Alsop T (2020) Global augmented/virtual reality market size 2016–2022 | Statistic. [online] Statista. Available at: <https://www.statista.com/statistics/591181/global-augmented-virtual-reality-market-size/>
2. Stats MMO (2023) VRChat active player count & population. [online] MMO Stats. Available at: <https://mmostats.com/game/vrchat>
3. Blagojević I (2022) Virtual reality statistics 2022 | 99firms. [online] 99firms.com. Available at: <https://99firms.com/blog/virtual-reality-statistics/>
4. VRChat (2022) Controls. [online] Available at: <https://docs.vrchat.com/docs/controls>. Accessed 1 Jan 2023
5. Basu T (2021) The metaverse has a groping problem already. [online] MIT Technology Review. Available at: <https://www.technologyreview.com/2021/12/16/1042516/the-metaverse-has-a-groping-problem/>
6. Yarramreddy A, Gromkowski P, Baggili I (2018) Forensic analysis of immersive virtual reality social applications: a primary account. [online] IEEE Xplore. <https://doi.org/10.1109/SPW.2018.00034>
7. Casey P, Lindsay-Decusati R, Baggili I, Breitinger F (2019) Inception: virtual space in memory space in real space—memory forensics of immersive virtual reality with the HTC vive. *Digit Investig* 29:S13–S21. <https://doi.org/10.1016/j.diin.2019.04.007>
8. Laeeq K (2022) Metaverse: why, how and what. [online] researchgate. Available at: https://www.researchgate.net/publication/358505001_Metaverse_Why_How_and_What. Accessed 28 Dec 2022
9. Rec Room (2023). Rec Room. [online] Rec Room. Available at: <https://recroom.com/>
10. Bigscreen Inc. (2023) <https://www.bigscreenvr.com/software>. [online] www.bigscreenvr.com. Available at: <https://www.bigscreenvr.com/software>
11. Valve Incorporation (2023) Steam, the ultimate online game platform. [online] store.steampowered.com. Available at: <http://store.steampowered.com/about/>. Accessed 23 Apr 2023
12. Clement J (2021) Number of Steam users 2021. [online] Statista. Available at: <https://www.statista.com/statistics/308330/number-stream-users/#:~:text=Steam%20had%20approximately%20132%20million>.
13. Clement J (2023) Rec room peak players on steam 2023. [online] Statista. Available at: <https://www.statista.com/statistics/1292058/rec-room-number-players-steam/>. Accessed 24 Apr 2023
14. VRChat Inc (2023) VRChat. [online] VRChat. Available at: <https://hello.vrchat.com/>
15. SteamDB (2019) VRChat—steam charts. [online] Steamcharts.com. Available at: <https://steamcharts.com/app/438100>
16. Oculus (2023) Oculus rift and touch controllers bundle: Amazon.co.uk: PC & Video Games. [online] www.amazon.co.uk. Available at: <https://www.amazon.co.uk/Oculus-Rift-Touch-Controllers-Bundle/dp/B073X8N1YW>
17. Exterro (2022) FTK imager. [online] Exterro. Available at: <https://www.exterro.com/ftk-imager#:~:text=FTK%C2%AE%20Imager%20is%20a>
18. Carrier B (2019a) Autopsy. [online] Sleuthkit.org. Available at: <https://www.sleuthkit.org/autopsy/>
19. Carrier B (2019b) The Sleuth Kit. [online] Sleuthkit.org. Available at: <https://www.sleuth-kit.org/sleuthkit/index.p>
20. The Volatility Foundation (2023). <https://www.volatilityfoundation.org/>

Cloud-Based Secure Electronic Medical Data Sharing System Using Blockchain Technology (Simulation of a Ransomware Attack with OWASP)



Rodrigue Ngomsi and Hamid Jahankhani

Abstract Attacks using ransomware are increasing, putting current healthcare institutions and other crucial infrastructure at risk. The current healthcare administration systems could be a soft target, as evidenced by cyberattacks such as WannaCry incident in 2017, that affected just over one third of NHS trusts PCs and over 250,000 workstations in the majority of countries in the world. The finding from this study shows that sharing medical data on blockchains in the cloud is resistant to ransomware attacks because from the proof-of-concept application built and the ransomware attack carried out, it was discovered that the ransomware can harm a node and might make it unavailable, it cannot spread to other nodes since each node is physically isolated from the other. Hence the concurrent processing and distributed nature of the cloud make its security unassured. As medical data is especially susceptible to security flaws and assaults such as forgery, tampering, and privacy leaks, the blockchain has been utilized in the cloud to protect and secure it.

Keywords NHS · Blockchain · OWASP · Electronic health record systems

1 Introduction

Blockchain is a decentralized network that never removes or modifies data and continuously adds new information without asking for permission. As blockchain technology purports to do away with the requirement for business-to-business trust, it has become a crucial instrument for destroying trust in online middlemen [11]. The importance of health care has increased along with the prevalence of illnesses and patients. It is essential to keep track of a person's medical history to appropriately address any health requirements. The present ecosystem makes sure that health data is transparently preserved and exchanged between entities. Blockchain is a well-known technology with several intriguing characteristics. Blockchain technology is

R. Ngomsi · H. Jahankhani (✉)
Northumbria University, London, UK
e-mail: Hamid.jahankhani@northumbria.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
H. Jahankhani (ed.), *Cybersecurity Challenges in the Age of AI, Space Communications and Cyborgs*, Advanced Sciences and Technologies for Security Applications,
https://doi.org/10.1007/978-3-031-47594-8_22

predominantly used for data management tasks in healthcare and IoT to enhance data security, including data integrity, access control, and privacy protection. In addition, it is found that despite blockchain's many benefits for the healthcare industry, authors frequently use them for data management and medication supply chain management to stop fraud and, consequently, provide patients control over their data.

The typical configuration of traditional electronic health record systems is one of centralization and a single point of failure. Electronic medical records (EMRs) are frequently shared peer-to-peer in the healthcare industry and contain essential extremely sensitive information. Blockchain creates a collaborative, permanent, and public ledger of all transactions, enabling developers to build applications with trust, accountability, and transparency. Recent modifications to Britain's medical record systems have only partially digitised a complicated system: healthcare administration systems remain incompatible. The system has numerous issues, but only a small number of them, like storage performance, interoperability, and data availability, are critical. The increasing rise in ransomware attacks puts the current healthcare system and other vital infrastructure at risk. The system used for actual health administration is vulnerable to cyberattacks. Specifically, the WannaCry assault in 2017, impacted over 250,000 machines in 150 countries as well as computers in 80 out of the 236 NHS trusts [12]. Data access is vital in the healthcare industry because it is necessary for medical treatment and because cyberattacks impair data availability and accessibility. This paper proposes utilising Blockchain technology to allow safe cloud-based medical data sharing.

2 Literature Review

The term "electronic health record" (EHR) refers to information created and supported by a healthcare institution, namely a hospital, integrated delivery network, clinic, or doctor's office, to permit employers, physicians, and other healthcare professionals gaining access to patient's health records [2]. Medical practitioners may be surprised to find that electronic medical records were first proposed in 1960 by a physician named Lawrence L. Weed. The Mayo Clinic started developing electronic medical record systems in the 1960s. In 1967, doctors and computer technicians at the University of Vermont began collaborating on the PROMIS project, an electronic medical record system [2].

Recent efforts to reorganise Britain's medical record systems have been so complicated that it has only been partially digitised: paper records are still frequently utilised at the secondary level, in addition to a number of disconnected local electronic systems that are Trust-specific. Despite considerable progress in the use of technology in clinical practises and recent large-scale budget transfers to the National Health Service, healthcare administration systems still lack interoperability [12]. This lack of interoperability can lead to both clinical and administrative errors, as was recently the case in the United States when the National Health Service (NHS) forgot to summon 50,000 women for a cervical screening test. In addition, patients

must repeatedly retell their medical history, which is time-consuming and inefficient. Incomplete information throughout this treatment could result in misconceptions and medical errors. Even more importantly, a lack of organisation can impair therapy outcomes by causing avoidable situations in which necessary information is delayed. A growing number of systems are employing Blockchain in a variety of ways to solve interoperability by providing physicians with more detailed patient data acquired through linked but independently maintained systems [12].

Numerous industries, including banking, healthcare, manufacturing, and education, use blockchain applications to take use of the technology's unique set of attributes. Credibility, reliability, organisation, and openness are all advantages of blockchain technology [5]. Due to its unique combination of characteristics, including decentralisation, immutability, and transparency, blockchain technology is capable to assist in multiple industries. In the medical care sector, blockchain is anticipated to have a substantial impact. Health informatics practitioners and researchers constantly seek to keep up with the advancement of this field's young but rapidly expanding body of research. This project provides a comprehensive analysis of current research on the healthcare industry's adoption of blockchain technology.

A blockchain is a series of blocks placed sequentially that connects a decentralised distributed ledger of data [6]. Blockchain is typically jointly administered by a peer-to-peer network that finds consensus by following a rule for confirming new blocks into the blockchain. The blockchain network lacks a centralised core node. Each network node stores a replica of the blockchain's data. If the data of a single node is edited or erased, the blockchain's data will remain untouched. A link mechanism called the irreversible hash function is utilised to verify the integrity of previous block, and it links each block in the blockchain ledger using a cryptographic hash value, meaning that each block will contain the hash value of the precedent block's content. By virtue of its status as a decentralised, distributed database, the blockchain is immutable and traceable due to this quality. The majority of a block is comprised of the block body and block header. There are six sections in a block header; the previous block's hash, the Merkle root hash value produced by the transaction ID, the timestamp at which the block was created, the version, which specifies the validation guidelines to be followed for a specific data type, the target difficulty, and the random number chosen by the consensus node to implement the consensus mechanism are all contained in the block header. The architecture of the block guarantees that one's content cannot be altered in any way.

Some blockchain's key element architectures are in charge of keeping track of transactions during a process. Within the blockchain architecture, these elements make up the node, which is the user or computer. Data in a blockchain database is stored in chunks that are connected together. The following modules are typical of a blockchain platform (Mahendra and Thomas 2019):

- **Cryptographic Services:** This layer gives entrance to encryption protocols such as the hashing algorithm and digital signatures.

- **Smart Contract Unit:** The optional module for smart contracts is only compatible with stateful blockchains. Business logic may be implemented using any programming language, such as solidity and C++.
- **Blockchain Secondary Storage:** A highly secure, dependable, and scalable storage system is necessary for storing block data inside the Blockchain distributed ledger, and conducts a large number of transactions. This layer enables the platform to permanently store data. In addition to other distributed data storage choices, LevelDB, H2 Database, and MongoDB are often deployed to store ledger data.
- **Blockchain Memory Storage:** it retains most transaction history in memory so that data recovery and transaction execution can occur more quickly. A number of data structures, including Monocyclic Directed Graph, Connectionist Array, and Merkle Tree, are utilised as memory storage in numerous Blockchain platforms.
- **Consensus Protocol Module:** It has a mechanism for achieving consensus among nodes regarding the authenticity and validity of operations. Proof of Work (PoW), Proof of Stake (PoS), Proof of Importance (PoI), Raft, Byzantine Fault Tolerance (BFT), etc., are examples of well-known Blockchain Consensus Protocols now in use.
- **Blockchain Services Layer:** By leveraging this layer, the blockchain platforms can have access to additional capabilities, such as platform services exposed via application programming interface (API), member administration, authorization, access control, event dissemination, and notification services, among others.
- **Communication Protocol:** The Blockchain Protocol facilitates communication between nodes participating in the Blockchain network by implementing standardised criteria.

3 Electronic Medical Data Sharing System Using Blockchain

The verification, preservation, and synchronisation of digital medical information in the medical system have been difficult issues for a very long time, and random allocation of patient records poses multiple risks to the privacy of patients [10]. Consequently, the topic about assuring protected data sharing while respecting the privacy of consumers becomes vital. Few years back, blockchain has been presented as a possible method for achieving secure and private data sharing because to its immutability characteristics. The maintenance and exchange of electronic medical information is an integral component of modern medical therapy. It is difficult to share data security in typical cloud-based medical record storage solutions. Due to the tamper-proof and accountable of the blockchain technology, very sensitive medical information can be transmitted [6].

Cloud-based medical record keeping has made it easier to share patient information. It is now simple and streamlined to share data across hospitals and research institutes [14]. Benefits of collaboration include gaining new insights on currently

utilised medications, analysing new ones, and enhancing population health management. However, the approach poses significant risks and challenges that must be disregarded.

Blockchain is relatively fresh compared to other rising technologies. Novel applications were deployed throughout its successful adoption in healthcare. To discover cheap cures and cutting-edge treatments for a variety of ailments, it is necessary that all the prestigious network members and healthcare providers communicate and distribute data efficiently and effectively [3]. Several enterprise-wide benefits will result from the implementation of Blockchain technologies within the healthcare business. It provides customer records, medical technology, clinical studies, the medicine supply chain, and pharmaceutical product integration.

4 Ransomware Attack Mitigation Using Blockchain

Malicious software, such as ransomware or ransom malware, prevents users from logging onto their systems or personal documents and demands a ransom in exchange for decryption [13]. The environment of ransomware is undergoing radical change. In 2018, Sophos found that, with an average of two ransomware assaults per firm, fifty-four percent (54%) of the organisations surveyed had encountered ransomware in the previous year [9]. The industries most badly impacted were healthcare, energy, business services, and retail. India, Mexico, the US, and Canada had the highest infection rates. 77% of firms used obsolete endpoint security at the time of the assault, and 54% lacked specific anti-ransomware protection [15].

The Wanna Cry, Petya, and other ransomware attacks have damaged a number of organisations. Eternal Blue is the name of the exploit that Wanna and Petya employed. Eternal Blue enables remote code execution by conveying specifically designed messages over the network to the vulnerable SMB service on Microsoft Windows-powered workstations [9]. The inherent qualities of decentralisation and dissemination bolster the blockchain's resistance to attacks. Decentralization relates to the network's ability to distribute data across all users, also known as nodes, as opposed to hosting it entirely on a single server. Due to its ability to eliminate the need for trust between organisations, blockchain technology has emerged as a significant weapon in the fight against online intermediaries [11]. Due to the increase in the number of patients and illnesses, the significance of health care has grown. It is essential to keep track of an individual's medical history in order to appropriately address any health needs. The present ecosystem ensures the maintenance and transmission of health data among businesses in a transparent manner.

Recent attempts to upgrade Britain's health record systems have culminated in a largely computerised system: Interoperability amongst healthcare management systems is still lacking. Data accessibility, scalability, interoperability, and storage performance are just a few of the system's most pressing concerns. Existing healthcare systems and other essential infrastructure are threatened by the present increase

in ransomware attacks. Actual health administration systems are susceptible to cyberattacks. Specifically, the Wanna cry attack of 2017, which infected computers in 80 of 236 NHS trusts and over 250,000 workstations in 150 countries [12]. Access to data is crucial in the healthcare industry, as it is necessary for medical treatment, and cyberattacks impair data access and availability. Multiple publications have recommended using blockchain to reduce the impact of ransomware outbreaks such as CryptoLocker [1, 4]. Consequently, there is a demand for a secure cloud-based medical data exchange utilising Blockchain technology.

5 Research Methods

Sharma et al. [7] proposed a Secure Cloud Storage Architecture for the Cloud Environment of Digital Medical Records. In the same way, this work proposes a framework for sharing medical data utilising blockchain and cloud computing. Nonetheless, this study enhances the security layer [8]. None of the authors of the evaluated publications considered the prospect of a ransomware assault, its potential impact on medical records, or its operation in real-time. These issues undermine blockchain's safety as well as its capacity to resist and rebound from crypto-ransomware attacks. Hence This research offers a cloud-based secure electronic medical data sharing system using blockchain technology and OWASP ransomware attack simulation to demonstrate the proposed system's ransomware-resistance.

This study employs a quantitative research approach. Quantitative methods place a heavy emphasis on precision in measurement techniques, statistical, mathematical, or numerical analysis of data obtained from experiments, surveys, and other types of research, as well as the manipulation of previously obtained statistical data through the use of computational techniques. Quantitative research aims to comprehend particular events or generalise findings across groups by collecting numerical data. Consequently, the author of this study investigates the inner workings of blockchain platforms for secure medical record sharing.

The blockchain is a distributed, block-based, decentralised data ledger. In the Bitcoin white paper, Nakamoto described blockchain technology, which combines encryption and peer-to-peer communication for the first time. Blockchain is typically jointly administered by a peer-to-peer network that finds consensus by adhering to a rule for authenticating new blocks into the blockchain. The blockchain network lacks a centralised core node. Each network node stores a replica of the blockchain's data [6]. If a single node's data is altered or deleted, the data saved on the blockchain will remain unaffected. The blocks of the blockchain ledger are connected by cryptographic hash values; each block contains the content's hash value, and an irreversible hash function is used as the link mechanism to verify the previous block's integrity. This property makes the blockchain traceable as a distributed, decentralised, immutable database. Block header and block body comprise the majority of a block. The block header consists of six different components. Block header information includes the hash of the preceding block, the Merkle root hash value calculated

using the transaction ID, the timestamp at which the block was created, the version, which specifies the validation guidelines to be followed for a specific data type, and the target difficulty and random number used by the consensus node to carry out the consensus mechanism [6]. The body of the block contains all transaction orders. The preceding hash value links the blocks together to form a chain. The structure of the block makes it impossible to alter or alter the content.

This work creates a blockchain-based electronic medical record model to enable the information production of the three-level diagnostic and treatment model and to solve the need for doctors to quickly obtain patients' prior medical histories in stroke referral situations. The model's construction aims to:

- Clarify patient ownership of medical record data and ensure individuals have strict access control rights to medical record data
- Ensure the confidentiality of patient identification and medical record information throughout storage and exchange.
- Develop an efficient and secure sharing mechanism to reduce user duplication of effort, and
- Realize the safe storing of voluminous, highly-sensitive digitised medical records.

Consequently, after constructing the blockchain-based data-sharing application, the study uses the OWASP framework to analyse it for security vulnerabilities. The resulting programme is subsequently distributed across numerous nodes to facilitate file sharing. A ransomware attack is simulated on one of the nodes to demonstrate the advantage of using blockchain for medical file sharing. This will further show the proposed system is compliant with the security CIA triad (Confidentiality, Integrity, and Availability).

5.1 Conventional Software Architecture Versus Proposed Software Architecture

Software architecture is the process of developing the high-level structure of a software system. It converts software qualities such as scalability, security, reusability, extensibility, modularity, and maintainability into structured business solutions. This is what is meant when referring to the system's overall design or structure. When defining the architecture of a system, a number of high-level architectural patterns and concepts are applied. Emphasis is placed on the system's visible components and their relationships with one another.

The lack of physical security in architecture model is cause for concern. The Secure Cloud Storage Architecture for the Digital Medical Record Cloud Environment was proposed by Sharma et al. [7] and [6]. This research intends to enhance the security layer [6, 8]. The preceding chapter discussed the likelihood of a ransomware assault, how it can affect medical records, and how it operates in real-time. These issues undermine blockchain's security and its ability to withstand and recover from

crypto-ransomware attacks. Hence this paper proposes a cloud-based secure electronic medical data sharing system architecture using blockchain technology alongside a simulation of a ransomware attack with OWASP to prove the proposed system is immune to a ransomware attack. Figure 1 illustrates the proposed architecture.

Figure 1 depicts the Blockchain application design proposed for the cloud-based medical data sharing system. The architecture exemplifies a secure system with physical and cloud security. The system's users comprise a physician, an administrator, and a patient. The Blockchain technology enables the distribution rather than duplication of digital information. This distributed ledger facilitates openness, confidence, and data security. These are the fundamental Blockchain architecture building blocks. The Blockchain protocol permits advanced data to be appropriated rather than copied. The sent record provides transparency, confidence, and data security.

These are the core Blockchain technology applications:



Fig. 1 Proposed Blockchain architecture

- Node—client or computer within the Blockchain technology (every node has an autonomous duplicate of the entire Blockchain record)
- Transaction—The smallest building block of a Blockchain architecture (records, data, etc.) that serves as the foundation for Blockchain.
- Block—a data frame used for storing several trades that is distributed to all hubs in the system.
- Chain—a sequence of squares in a certain request.
- Consensus (agreement convention)—There are many standards and plans for executing Blockchain operations.
- Encryption: Encryption can be utilised to scramble data so that only authorised parties are able to decipher it. Technically, it is the transformation of plaintext, which can be read by humans, into ciphertext, which is unreadable text. Therefore, the application’s Encryption Layer is important for concealing patients’ medical records.
- Decryption: Decryption is the procedure of restoring data to its original form. Typically, encryption is performed backwards. Only a trustworthy user with access to the secret key or password can decrypt the data.
- Cloud: The Internet, or more particularly, whatever you can access remotely through the Internet, is the cloud. When anything is in the cloud, it indicates that it is kept online on servers rather than on the hard disk of your computer. The clouds serve as the host of the system. Huge healthcare data is encrypted and stored outside of the blockchain due to cost, storage capacity, and other practical constraints. The doctor uploads electronic medical records from the client to a cloud server, which is then utilized to store the ciphertext of the data to ease the blockchain’s storage strain. The cloud server communicates with the blockchain in response to a user request to search for data. The electronic medical record’s ciphertext is returned to the blockchain master node for proxy encryption once the blockchain sends the ciphertext request.

This particular work uses secondary data collection techniques. Secondary research implies reference to studies that have been previously published in research reports and similar papers. These resources could be made available via the internet, public libraries, completed questionnaires, and so on. In addition, a few governments and non-government entities maintain data that can be accessed for research purposes. Existing data is gathered and summarised to enhance the study’s overall efficacy. Secondary research is substantially less expensive than primary research because it uses data that is already in circulation rather than gathering information from organisations or enterprises directly or through a third party.

5.2 Algorithm

This research adopted the public-key encryption with keyword search (PEKS) algorithm. We investigate the search for encrypted files using a public key technique. Consider user Rodriguez sending user Amanda an encrypted email using Amanda's public key. To properly route the email, an email gateway must detect whether the email contains the word "urgent." Amanda, on the other hand, is unwilling to grant the gateway access to the encryption keys for all of her conversations. Without understanding anything else about the email, we devise and construct a mechanism that enables Amanda to provide the gateway with a key that allows the gateway to determine if "urgent" is a keyword in the email. This method is known as public key encryption with search query. In the cloud-based blockchain system, the search is utilised to search encrypted medical records.

These have been decomposed below:

Algorithm for Initialization: Accepts the safekeeping parameters 1λ as input to obtain the private key prk and public key plk

$$\text{Initialize}(1\lambda) \rightarrow (prk, plk) \quad (1)$$

Keyword Encryption Algorithm: accepts the public key plk and keywords kw of documents as input and outputs the ciphertext of document keywords c

$$\text{Encrypt}(plk, kw) \rightarrow c \quad (2)$$

Trapdoor Generation Algorithm: Accepts the private key prk and search keyword kw as input and outputs the trapdoor tp corresponding to the search keyword

$$\text{Trapdoor}(prk, kw) \rightarrow tr \quad (3)$$

Matching Algorithm: Accepts the generated public key plk , trapdoor tr , and ciphertext c as input and output Boolean variable b . When the trapdoor and ciphertext correspond to the same keyword, $b = 1$, otherwise, $b = 0$

$$\text{Testing}(tr, c, plk) \rightarrow b \quad (4)$$

Depending on the encryption mechanism, searchable encryption techniques can be divided into symmetric searchable encryption (SSE) and asymmetric searchable encryption (ASE) (ASE). SSE uses the same key for both encryption and decryption, eliminating the need for communication between the data owner and data requester. ASE, or public key searchable encryption, is used to combat untrustworthy server routing [6]. Two keys are utilised in the encryption method. Using the public key, the plaintext keyword information is encrypted with the matching target ciphertext, and using the private key, keyword trapdoors are constructed. The keyword trapdoor controls search-matching behaviour in searchable encryption methods.

Once the blockchain node has obtained the search trapdoor, it can begin matching searches. In the event of a medical care referral, the patient has the authority to use and control the data, and must encrypt the keyword trapdoor with a private key. Therefore, the proposed method employs an asymmetric searchable encryption mechanism to prevent users from indefinitely use symmetric searchable encryption to generate the requisite keyword trapdoors [6].

5.3 System Implementation/Design

The system implementation is divided into two namely: The front end and the Back end. The front end was developed using HTML5, CSS3, and JavaScript. The backend was developed using PHP and MYSQL. The Ransomware was written with Shell scripts and python. The platform used for development includes Visual Studio editor, Virtual box, and XAMPP server. System design is the process of creating a system's architecture, modules, and components, in addition to its interfaces and the data it processes. System analysis is the act of decomposing a system into its component elements in order to evaluate how well those parts function together to achieve the defined criteria. To allow implementation of architectural entities as described in models and views of the system architecture, the system design process seeks to provide sufficient data and knowledge about the system and its system components.

5.4 Doctor's Registration Page

The system recognizes doctors as admins hence they are required to be registered in the system alongside their login credentials. The page collects the biodata of doctors using the system.

This page visualizes the records of the doctor in a human-readable format from the database. Although the backend database is encrypted using Open SSL private and public key pair scheme.

5.4.1 Add Patient Page

The system Admin uses a form to collect the bio-date of patients and saved it in an encrypted database. The form collects key information as regards the state of health of a particular patient over some time for data sharing for medical use.

The add access page allows the system administrator to generate access tokens and keys for data sharing. i.e. sharing a patient's medical history with another authorized medical institution. The page allows the admin to revoke access to specific files also.

This page also allows authorized institutions to access specific patient records using access keys and access tokens generated by the system admin. The authorized parties can only see files they have access to view.

5.4.2 OWASP Penetration Testing Session

Open Web Application Security Project Zed Attack Proxy (OWASP ZAP) tool was used to test the security of the medical sharing application. First, launch the tool to select the type to scan to be done. For this session Manual exploration is used.

After selecting the scan. Some configurations need to be done. First is to configure the OWASP ZAP proxy to recognize `localhost@127.0.0.1` and port 8080.

After configuring the proxy on OWASP ZAP, then the proxy on Firefox proxy which is the chosen browser for this attack is configured.

For the attack to work effectively, a digital certificate is generated from OWASP ZAP.

The next step is to start the scan by configuring an automated scan on the web application.

Once the launch attack button is activated, The attack begins in the Firefox browser.

Further analysis shows the highest security is medium. The rest are significantly low. The reason for the medium and low-level vulnerabilities was traced to the lack of HTTPS which is because the rest was done over localhost. A vulnerable JavaScript library also.

5.4.3 Ransomware Attack Simulation

After configuring each node, ransomware is introduced into one of the nodes. The ransomware used for this demonstration is an open-source ransomware script for educational purposes. The moment the attack is launched on Node 1, it knocks it offline.

However, since multiple nodes are running the application will continue to run on other nodes. This is proof that blockchain applications are isolated on different nodes and can continue to function after an attack on a single node thereby ensuring availability through resilience.

6 Critical Discussion

Applications based on blockchain technology are utilised by several industries, including banking, healthcare, manufacturing, and education, to take advantage of the unique benefits that are made available by this type of technology. Organization, transparency, credibility, and trustworthiness are some of the advantages afforded

using blockchain technology. Blockchain technology has the potential to be beneficial to many different types of businesses due to the unique mix of features it possesses. These properties include decentralisation, immutability, and transparency. The medical industry is one that many believe will be significantly impacted by blockchain technology. Both practitioners and academics in the field of health informatics put in a lot of effort to ensure that they are always up to date with the rapidly growing body of knowledge that the field is currently developing. This project provides an in-depth analysis of recent research looking into how blockchain technology might be applied in the medical field.

In this study, a blockchain application was developed and deployed to facilitate the secure sharing of medical files across authorized parties using the cloud. The system was successfully developed and tested as a proof-of-concept system. The system addressed the problem of the constant ransomware attack on critical infrastructure such as health infrastructure by providing a simulated ransomware attack to understand how to better prevent ransomware attacks from hitting critical infrastructure. The study used standard encryption schemes such as:

- Secure Hash Algorithm (SHA-256)
- Open SSL (Public and Private Key pair)
- Password-Based Key Derivation Function 2 (PBKDF2).

To ensure data security and frustrated effort to breach the system’s security. The encrypted and decryption process was tested, and the result was given for each case tested (Fig. 2).

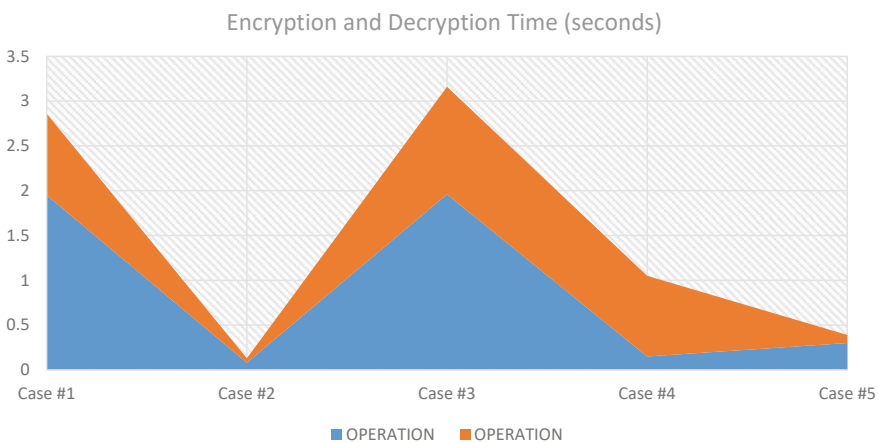


Fig. 2 Encryption and decryption time

Sharma et al. [7] suggested a Secure Cloud Storage Architecture for the Digital Medical Record Cloud Environment. Similarly, this project's System for sharing medical data using blockchain and cloud technology has improved the security layer [8]. From the list of papers reviewed none of the authors took into cognizance the possibility of a ransomware attack, how it can impact medical records and how it works in a real-time environment. These problems jeopardize the usefulness of blockchain in terms of security and its ability to withstand and recover from crypto-ransomware attacks. This research was able to prove how blockchain can build resilience against a ransomware attack.

A ransomware attack on a hospital puts more strain on its resources overall and raises death rates there. One of the reasons hackers target hospitals is probably because of these potentially disastrous effects. Healthcare institutions are a regular target for ransomware attacks because they rely so much on access to data, such as patient information, to keep their operations running smoothly. Patients may be negatively affected by even a short delay in gaining access to materials. During a ransomware attack, an adversary might encrypt many files, leaving them inaccessible along with the systems that rely on them. If a ransom is not paid, these encrypted files are typically locked permanently, requiring the organisation to rebuild the data, if feasible.

An organization's ability to function can be sceptically affected with a ransomware assault. Despite the fact the disaster recovery is implemented with backups, it could take hours to restore affected systems. Worse still, it may take days or weeks for businesses that were not adequately prepared or whose backups may have been compromised by the assault to recover completely. Consequently, sales may decrease or cease totally until they recover.

The reputation of an organisation may be negatively impacted as a result of a data breach or ransomware assault. Some consumers can interpret a successful assault as proof of lack of security procedures, or they might suffer a severe service interruption that they decide to do business somewhere else. During a ransomware attack, an adversary might encrypt many files, leaving them inaccessible along with the systems that rely on them. If a ransom is not paid, these encrypted files are typically locked permanently, forcing the organisation, if possible, to recreate the data. Even if a ransom is paid, there is no promise that a threat actor would act graciously and provide a decryption key. Furthermore, even if a key is given, it's still possible that the ransomware assault caused serious catastrophic damage, necessitating the need to restore the afflicted systems. In addition, the loss of this information might result in legal action or the loss of a competitive advantage if a threat actor steals a trade secret, intellectual information, or Protected Health Information (PHI).

Hence the need to adopt blockchain technology to build future data drive applications such as the electronic medical file sharing system.

The result is limited to the simulation attack and system implementation. Due to technical feasibility using the OWASP framework for the test was not achievable. Hence further penetration testing was not done.

Potentially, blockchain technology might be used to stop ransomware in its tracks. Frequently, the fact that just a single copy of the victim's data was ever saved on the server is the primary concern for victims. By focusing on this one point of failure, attackers might deny a victim access to their data. What if, instead of retaining a single, centralised copy of their data, the victim kept distributed copies across many servers housed by different service providers? In such a scenario, they might have isolated the infected machine from the other copies and retrieved the whole database.

Combining three popular technologies, blockchain: keys for cryptography. A network of peers that uses a shared ledger. A method of computation for storing network transactions and records. Ransomware threats are being countered with the use of emerging blockchain analytics tools' potent capabilities. Comprehensive antivirus and anti-malware products are the most common method of ransomware defence. They can search for, identify, and counteract cyber dangers. Blockchain technology can strengthen shoddy security. The entire system is powered by secure information encryption, effectively establishing a wall between hackers and personally identifiable information. Hence this study recommends the adoption of blockchain for building data-driven applications such as electronic medical file-sharing systems.

7 Conclusion

In this research, a secure electronic medical file-sharing system was developed, deployed, and tested in a virtual environment. Along with the rise of ailments and patients, the value of health care has grown. A person's medical history must be tracked to properly meet any health requirements. The current ecosystem guarantees the transparent interchange and preservation of health data between entities. A well-known technology with several intriguing features is blockchain.

The finding from this study shows that sharing medical data on blockchains in the cloud is resistant to ransomware attacks because from the proof-of-concept application built and the ransomware attack carried out, it was discovered that the ransomware can harm a node and might make it unavailable, it cannot spread to other nodes since each node is physically isolated from the other. Hence Blockchain is resilient to a ransomware attacks. The finding also shows that Privacy issues, security hazards, poor system transparency, and a lack of patient control are the key issues with popular health information Exchange systems. Blockchain technology can transform how the healthcare sector currently exchanges information. Hence the concurrent processing and distributed nature of the cloud make its security unassured. As medical data is especially susceptible to security vulnerabilities and assaults such as forgery, manipulation, and privacy leaks, the blockchain has been used in the cloud to protect and secure it. The work has contributed to the body of knowledge by improving the security layer [8]. From the list of papers reviewed none of the

authors took into cognizance the possibility of a ransomware attack, how it can impact medical records and how it works in a real-time environment. These problems jeopardize the usefulness of blockchain in terms of security and its ability to withstand and recover from crypto-ransomware attacks. This research was able to prove how blockchain can build resilience against a ransomware attack.

References

1. Alqahtani A, Sheldon FT (2022) A survey of crypto ransomware attack detection methodologies: an evolving outlook. *Sensors* 22(5):1–19. Available at: <https://doi.org/10.3390/s22051837>
2. Al Hajeri A (2011) Electronic health records in primary care: are we ready?. *Bahrain Med Bullet* 33(2)
3. Haleem A et al. (2021) Blockchain technology applications in healthcare: an overview. *Int J Intell Networks* 2:130–139. Available at: <https://doi.org/10.1016/j.ijin.2021.09.005>
4. Humayun M et al. (2021) Internet of things and ransomware: evolution, mitigation and prevention. *Egypt Inf J* 22(1):105–117. Available at: <https://doi.org/10.1016/j.eij.2020.05.003>
5. Odeh A, Keshta I, Al-Haija QA (2022) Analysis of Blockchain in the healthcare sector: application and issues. *Symmetry* 14(9). Available at: <https://doi.org/10.3390/sym14091760>
6. Qin Q, Jin B, Liu Y (2021) A secure storage and sharing scheme of stroke electronic medical records based on consortium blockchain. *BioMed Res Int*. Available at <https://doi.org/10.1155/2021/6676171>
7. Sharma S et al. (2020a) Secure cloud storage architecture for digital medical record in cloud environment using blockchain. *SSRN Electron J* [Preprint]. Available at <https://doi.org/10.2139/ssrn.3565922>
8. Sharma S et al. (2020b) Secure cloud storage architecture for digital medical record in cloud environment using blockchain. *SSRN Electron J* [Preprint]. Available at <https://doi.org/10.2139/ssrn.3565922>
9. Sophos (2017) Best practices to block ransomware with a firewall, Techgoondu. Available at: <https://www.techgoondu.com/2017/10/02/best-practices-block-ransomware-firewall/>. Accessed 31 October 2022
10. Sun J et al. (2020) A blockchain-based framework for electronic medical records sharing with fine-grained access control. *PLoS ONE* 15:1–23. Available at: <https://doi.org/10.1371/journal.pone.0239946>
11. Uddin MA et al. (2021) A survey on the adoption of blockchain in IoT: challenges and solutions. *Blockchain Res Appl* 2(2):100006. Available at: <https://doi.org/10.1016/j.bcr.2021.100006>
12. Vazirani AA et al. (2020) Blockchain vehicles for efficient medical record management. *npj Digital Med* 3(1):1–5. Available at: <https://doi.org/10.1038/s41746-019-0211-0>
13. Walker M (2019) Blockchain, a barrier against ransomware, *thefintechtimes*. Available at: <https://thefintechtimes.com/blockchain-barrier-ransomware/>. Accessed 31 October 2022
14. Xia Q et al. (2017) BBDS: Blockchain-based data sharing for electronic medical records in cloud environments. *Information (Switzerland)* 8(2). Available at <https://doi.org/10.3390/info8020044>

15. Y Connolly L, Wall DS (2019) The rise of crypto-ransomware in a changing cybercrime landscape: taxonomising countermeasures. *Comput Security*, 87. Available at:<https://doi.org/10.1016/j.cose.2019.101568>