A Decentralised Secure and Privacy-Preserving E-Government System

NOE ELISA NNKO

A thesis submitted in partial fulfilment
of the requirements of the
University of Northumbria at Newcastle
for the degree of
Doctor of Philosophy

Research undertaken in the
Department of Computer and Information Sciences
Faculty of Engineering and Environment

November 2020

Abstract

Electronic Government (e-Government) digitises and innovates public services to businesses, citizens, agencies, employees and other shareholders by utilising Information and Communication Technologies. E-government systems inevitably involves finance, personal, security and other sensitive information, and therefore become the target of cyber attacks through various means, such as malware, spyware, virus, denial of service attacks (DoS), and distributed DoS (DDoS). Despite the protection measures, such as authentication, authorisation, encryption, and firewalls, existing e-Government systems such as websites and electronic identity management systems (eIDs) often face potential privacy issues, security vulnerabilities and suffer from single point of failure due to centralised services. This is getting more challenging along with the dramatically increasing users and usage of e-Government systems due to the proliferation of technologies such as smart cities, internet of things (IoTs), cloud computing and interconnected networks. Thus, there is a need of developing a decentralised secure e-Government system equipped with anomaly detection to enforce system reliability, security and privacy.

This PhD work develops a decentralised secure and privacy-preserving e-Government system by innovatively using blockchain technology. Blockchain technology enables the implementation of highly secure and privacy-preserving decentralised applications where information is not under the control of any centralised third party. The developed secure and decentralised e-Government system is based on the consortium type of blockchain technology, which is a semi-public and decentralised blockchain system consisting of a group of pre-selected entities or organisations in charge of consensus and decisions making for the benefit of the whole network of peers. Ethereum blockchain solution was used in this project to simulate and validate the proposed system since it is open source and supports off-chain data storage such as images, PDFs, DOCs, contracts, and other files that are too large to be stored in the blockchain or that are required to be deleted or changed in the future, which are essential part of e-Government systems.

This PhD work also develops an intrusion detection system (IDS) based on the Dendritic cell algorithm (DCA) for detecting unwanted internal and external traffics to support the proposed blockchain-based e-Government system, because the blockchain database is append-only and immutable. The IDS effectively prevent unwanted transactions such as virus, malware or spyware from being added to the blockchain-based e-Government network. Briefly, the DCA is a class of artificial immune systems (AIS) which was introduce for anomaly detection in computer networks and has beneficial properties such as self-organisation, scalability, decentralised control and adaptability. Three significant improvements have been implemented for DCA-based IDS. Firstly, a new parameters optimisation approach for the DCA is implemented by using the Genetic algorithm (GA). Secondly, fuzzy inference systems approach is developed to solve nonlinear relationship that exist between features during the pre-processing stage of the DCA so as to further enhance its anomaly detection performance in e-Government systems. In addition, a multiclass DCA capable of detection multiple attacks is developed in this project, given that the original DCA is a binary classifier and many practical classification problems including computer network intrusion detection datasets are often associated with multiple classes.

The effectiveness of the proposed approaches in enforcing security and privacy in e-Government systems are demonstrated through three real-world applications: privacy and integrity protection of information in e-Government systems, internal threats detection, and external threats detection. Privacy and integrity protection of information in the proposed e-Government systems is provided by using encryption and validation mechanism offered by the blockchain technology. Experiments demonstrated the performance of the proposed system, and thus its suitability in enhancing security and privacy of information in e-Government systems. The applicability and performance of the DCA-based IDS in e-Government systems were examined by using publicly accessible insider and external threat datasets with real-world attacks. The results show that, the proposed system can mitigate insider and external threats in e-Government systems whilst simultaneously preserving information security and privacy. The proposed system also could potentially increase the trust and accountability of public sectors due to the transparency and efficiency which are offered by the blockchain applications.

Contents

List of Figures

List of Tables

Acknowledgements

I wish to express my sincere appreciation to my principal supervisor, Dr. Longzhi Yang who guided and taught me to be professional and encouraged me to do the right things even when the research study got tougher. Without his persistent help, the goal of this PhD thesis would not have been realised. Also, many thanks go to my second supervisor Dr. Honglei li for all of her support whenever needed.

Secondyl, I would like to thank Dr. Edmond Ho and Dr. Naveed Anwar, who were respectively the subject expert and chairman of the annual progression research panel throughout this PhD study. Their suggestions and comments helped me to achive the goal of this PhD study easier more than I could ever give them credit for here.

Thirdly, I wish to acknowledge the great love and moral support I received from my lovely and supportive wife, Rachel; my kids, Ronel and Camila; my father and mother, Mr. and Mrs. Elisa; and all of my siblings. The completion of this PhD thesis would have been more difficult without their unconditional love, encouragements and prayers.

I am especially indebted to my colleagues Dr. Jie Li, Dr. Zheming Zuo, Dr. Yao Tan and others from the department of Computer and Information Sciences at Northumbria University, UK, for their social and academic support during the course of this PhD study. I really appriciate their genuine support.

I would also like to express my sincere appriciation to Dr. Julie Greensmith from the University of Nottingham, UK, who introduced the first version of binary Dendritic cell algorithm of which this PhD study has benefited a lot from it. She accepted our invitation to attend my presentation on a multiclass version of the algorithm. Her comments and suggestions were really commendable.

In addition, the financial supports from the Commonwealth Scholarships Commission (CSC-TZCS-2017-717) in the UK is truly appreciated. Without their funding, this PhD thesis could not have reached its objectives.

Finally, I would like to extend my deepest gratitude to the management of the University of Dodoma, Tanzania, for granting me a study leave and permission to persue this PhD study.

Dedication

This PhD thesis is dedicated to my parents Mr. and Mrs. Elisa Yesaya Nnko. Thank you for the sacrifices you have made for me throughout my life. It is also dedicated to my lovely wife and kids, Rachel, Ronel and Camila. I am deeply grateful for your moral support and the abundance of love during this journey. Lastly, I dedicate this thesis to my principal supervisor, Dr. Longzhi Yang, thank you for believing in me.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the University Ethics Committee on 27/09/2018.

I declare that the Word Count of this Thesis is 38,912 words.

Name: Noe Elisa Nnko

Signature:

Date: November, 2020

Chapter 1

Introduction

This chapter presents an overview of e-Government systems relevant to the scope of this PhD work. It includes details of the following sections: E-Government Systems, Challenges of Implementing E-Government Systems, Motivations, Goals and Objectives, and Structure of the Thesis.

1.1 E-Government Systems

The use of ICTs such as Internet and electronic devices in different organisations across the world has significantly increased over the past few decades. Online information sharing has a great impact on individuals as it makes daily communication easier and more efficient [155]. E-Government is one such system that uses ICTs to deliver public services to individuals including citizens, agencies, businesses, employees and other Governments [116, 155]. The transformation of Government systems from traditional paper based information sharing to its electronic counterpart increases transparency, accountability, participatory, effectiveness and efficiency of services delivered by Government agencies, resulting into a better Government system [155]. Additionally, e-Government is considered as one of the attributes of good Governance since it makes public administrators and officials more democratic and responsible due to transparency and interoperability provided [122, 14, 155]. Through e-Government system, citizens can participate in e-Democracy by engaging themselves in public events and decision making, and thus they are able to hold the Government accountable [155].

Generally, Government networks can communicate to each other better than business networks, because most of them are connected for transferring information to the public without competition [155]. One of the main goals of introducing e-Government system was to keep the public sectors working seamlessly for 24 hours every day. The electronic provision of the public services is very beneficial solution particularly to people with certain disabilities as they can access services online such as tax clearance, insurance registration and etc, with no need to appear physically in public offices. Correspondingly, individuals of e-Government are able to evade long queues in public offices whilst saving time and transportation costs; and at the same time the service providers can deliver services more effectively with low cost [122, 155]. Additionally, e-Government systems aim to link various Government departments, agencies and ministries so that they can instantly deliver online services to citizens and other stakeholders on demand [155, 122]. On demand information sharing is essential for promoting equality, increasing revenue, promoting competition among agencies and firms and accelerate marketing within the public sector.

According to the United Nation e-Government survey, 2018 [158], almost every Government around the world is currently providing its citizens and other stakeholders e-services via websites and mobile applications. For instance, US's Government portal (https://www.usa.gov/), UK's Government portal (https://www.gov.uk/), Tanzania's Government portal (http://www.ega.go.tz/), china's Government portal (http://www.gov.cn/english/), and e.t.c, provide information and e-services to individuals. Thus, regardless of their physical locations around the world, the use of e-Government web-portals enables citizens and organisations in private and public sector to interact directly [122, 155].

E-Government services typically can be categorised into 4 groups [155, 116]: G2G (Government to Government), G2C (Government to Citizens), G2B (Government to Business) and G2E (Government to Employees). Briefly, G2G offers transaction between Government departments such as central/national and local councils as well as information flow between Governments. G2C disseminates information to individuals such as driving license renewal, application of birth/death/marriage certificates, online payment of income tax etc. G2B exchanges information between Government and businesses such as policies, rules and regulation, online application of business permit etc. G2E exchanges information and documents between Government and its employees.

1.2 Challenges of Implementing E-government Systems

The complexity and continuous advancement of ICTs bring several challenges to the implementation and management of e-Government systems [122, 6, 101]. The common challenges that have been identified to hinder the implementation and acceptance of e-Government systems are discussed below.

ICT Infrastructure

The implementation of e-Government systems always face technological limitations such as compatible ICT infrastructure and standards among Government agencies and departments. Thus, ICT infrastructure is considered as one of the main challenges facing e-Government initiatives [101]. For e-Government services to be accessed and delivered to citizens, internet plays a crucial role. Internet-working between e-Government systems and user devices is needed to facilitate appropriate information sharing and provide new channels for services delivery [122, 101]. Studies have indicated that, ICT barriers that face e-Government implementation include existence of different technology infrastructure among agencies and departments, lack of interoperability of application interfaces and presence of different implementation frameworks [101, 137].

For instance, the data formats used by a particular application is not readable or is incompatible with other application within the same public sector. Transforming Government from traditional paper-based to e-Government requires a uniform architecture, ICT standards and models and guiding set of principles within the public sectors [137]. More precisely, a strong technological infrastructure within the public sectors is required for e-Government systems implementation. Hence, Government must develop an effective ICT infrastructure in order to deliver e-Government service efficiently [101]. Appropriate and interoperable ICT infrastructure within the public sectors is one of the requirements that must be maintained to ensure sustainable e-Government systems.

Information Privacy

Citizen's online privacy continues to be a critical challenge in e-Government implementation and acceptance [101, 8]. Privacy refers to the assurance of a suitable level of information protection and integrity to an individual. Government has an obligation to ensure that, the privacy of citizens' data collected is maintained while being processed and shared among public departments and agencies [123]. Privacy concerns due to disclosure of sensitive information, sharing and mishandling of private information as well as web and online tracking in e-Government systems are being frequently reported around the world [101, 123, 155]. Also, there is the concern that e-Government systems may be used to track and monitor individuals and violate their privacy.

Public sectors are required to address privacy in e-Government systems as a technical and policy issue in order to make sure that sensitive information about individuals are not intercepted due to poor network security or disclosure of information by officials. Addressing privacy concerns in e-Government systems is of paramount importance in increasing citizens' confidence and winning their trust. Mutual trust is required by citizens in order to allow Government to collect and use their data transparently. Therefore, Government authorities and policies must state and communicate clearly with citizens and other shareholders about the use and how the data collected are being processed by the Government.

Information Security

Information security is the protection of systems against an intentional or accidental disclosure to an unauthorised modification, an unauthorised access or destruction [104]. Information security is considered to be among the critical success factors of e-Government implementation and adoption process [164]. It also plays a key role in citizens' willingness to use e-Government applications and services [101]. Information security accounts for the protection of e-Government architecture including network, computers and software as well as controlling access to the stored information. However, information security has been identified as a challenging factor in the implementation of e-Government systems [122, 101, 115]. This is due to applications for accessing e-Government services are designed independently rather than being part of the whole systems and thus end up requiring their own security mechanisms that become inevitably incompatible with other applications and systems [101]. Compatibility can be provided by making sure that each e-Government application and system is equipped with the functionality such as encryption, digital signatures, authorisation, authentication, non-repudiation, protection of user IDs and passwords, and proper validation [123]. Information security should include protection of network and the documents it stores by suing firewalls and limits those who have access to the system [137].

Policy and Regulation

Adoption and implementation of e-Government systems is considered as an organisational issue rather than technical issue [60]. It requires a range of new policies, laws, rules and changes in Governance to work on electronic activities such as transmission of information, storage, archiving, protection, copyright issues, intellectual property issues and cyber-crimes. Online information sharing within the public sectors involves lawful signing a digital agreement or contract for protecting and securing all activities being done by individuals. Most countries across the world have not developed e-Government laws which make it difficult to regulate information sharing and handling [123, 158]. Particularly, many government agencies in different countries over the world are in the process of developing their own specific policies. Lack of detailed e-Government policy, or their early stage of formulation, has reduced the speed of e-Government implementation in many countries [101, 155]. Establishment of e-Government policies and regulations within the public sectors is necessary to ensure that, privacy, security and legal issues are provided. Until e-Government policies and regulations are put together and full conceived is when e-Government agencies would speed up implementation and acceptance of e-Government applications and technologies.

Lack of Qualified Personnel and Training

Lack of appropriate technical skills within the Government departments and agencies has been identified as another major concerns of e-Government implementation and integration [101, 122]. There has been a lack of qualified individuals and insufficient human resources training within e-Government agencies for years [123]. In fact, the majority of e-Government projects heavily rely on external consulting firms who provide human resources and support [123, 122]. Dependence on external consulting firms provides short term benefits in helping public sectors to deliver services faster to individuals, however, it raises longer term concerns on over-dependence on external firms for continuous support and maintenance. Indeed, the significance of key ICT experts on information systems projects has been appreciated for long time [33]. Therefore, critical knowledge is lost when an external expert leave the project which can contribute to failure of e-Government implementation and integration [101]. Hence, the availability of adequate skills is important for successful implementation and integration of e-Government systems. Technical skill is required for design, implementation, maintenance and installation of ICT infrastructure. Human capital challenge can be addressed by providing training to IT staff and managers in order to develop and create the basic and fundamental skills for e-Government use and application.

Digital Divide

The digital divide is the gap between those who have the opportunities to access computers and Internet against those who do not. The opportunity and ability to use Internet and computers have become a critical success factor in e-Government implementation and integration [101]. Individuals who do not have access to Internet are not able to access online services offered by the e-Government [101]. Across the world, not all citizens have access to computers and Internet particularly due to lack of financial resources or lack of skills to operate the technology [101, 123]. For instance, in developing countries the digital divide is too huge making it difficult for citizens to use and accept e-Government systems [122]. In fact, computer education is required for citizens to be able to benefits from e-Government applications [101]. Governments are responsible for training their citizens and officials on the basic skills of using ICT in order for them to be able to participate in e-Government integration process. Larger digital divide in developing countries increases the cost of training and technical barriers in implementing and sustaining e-Government services [122].

Leaders and Management Support

Government departments and agencies across the world are not well prepared for undertaking the implementation and integration of electronic services in public sectors [101]. E-Government agencies are mostly engaged in learning about e-Government implementation and integration process rather that direct acceptance within the public sector [164]. Studies have indicated that without support and acceptance from the top management of the Government, ICT projects are likely to be abandoned [122, 123]. Thus, e-Government implementation should receive the fully support from the top management for successful integration [122]. Top management means the commitment from the highest level of Government to provide best environment that influence participation of individuals in e-Government initiatives. The support from the top management in a country plays a critical role in the implementation and adoption of e-Government systems [122].

1.3 Motivation

E-Government is one of the most complex information systems which needs to be distributed, secure and preserves the privacy of sensitive information shared within the public sector. The existing e-Government systems such as websites and electronic identity management systems (eIDs) are centralised where all storage and processing are performed at a central duplicated servers and databases. Centralised management and validation system always presents a single point of failure and make the system a target to cyber attacks such as malware, ramsonware, virus, spyware, DDoS, DoS, and etc. Additionally, to perform a software upgrade to a centralised system may require to halt the entire system which leads to unavailability; and hence, users may fail to access the stored information and services.

Additionally, e-Government systems collect, store and process a significant amount of confidential and sensitive information about citizens, employees, customers, products, researches, financial status amongst others, using electronic computers. The compromise of such information usually leads to the loss of users' trust and confidence, opportunities, and financial advantages, etc [102]. In the future, the number of devices using e-Government services will increase dramatically due to the fast evolution of smart homes, IoT, smart cities and interconnected networks [15, 170]. Commonly, as the number of devices using e-Government systems increases, the number of malicious nodes trying to access and abuse unauthorised and sensitive information grows accordingly [170, 15, 171]. In the past, a research study conducted by [115] found that, more than 80% of e-Government web sites around the world were vulnerable to cross-site scripting (XSS) and structured query (SQL) injection due to lack of proper authentication mechanism applied to input data from users.

Data breaches in e-Government systems have been significantly increasing in recent years based on various sources. For instance, according to the 2019 Cyber Security Breaches Survey by the UK Government, around 32% of businesses and 22% of charities reported facing cyber security breaches or attacks in 2019, such as phishing, viruses, malware including ransomware attacks and impersonation of emails [154]. In 2017, the United State Government suffered one of the largest e-Government attacks, causing the loss of over 145 million Government employees' confidential information, including security clearance information, social security numbers, identities, passwords, etc [159]. In 2017, 3.34 million computer were compromised by hackers in China and caused a leakage of citizens' sensitive information like user name, photos, address and identity card numbers [30]. Also, according to the report in [148], in 2016, Tanzania Government was hit by cyber-terrorists, technology spies, hackers and digital fraudsters causing it to lose around 85 millions US dollars. In addition, more than 1,500 user accounts in Singapore were hacked in the Government platform in 2014, where hackers gained access to create new businesses and apply for work permits [138]. Due to supremacy and political differences, in May 2019, the Palestinian cyber warfare division was identified to have carried out a cyber attack against Israeli's IT infrastructure connected to their e-Government systems and gain access to state's sensitive information [84].

It is therefore of ultimate importance to ensure the security, privacy, confidentiality, integrity and availability of e-Government systems. Thus, there is a need of developing a decentralised secure e-Government system equipped with cybersecurity attacks detection to enforce system reliability, security and privacy. Since its inception as a solution for secure cryptocurrencies sharing in 2008, the blockchain technology has now become one of the core technologies for secure data sharing and storage over trustless and decentralised peer-to-peer (P2P) systems [32]. Blockchain technology enables implementation of highly secure and privacy-preserving decentralised systems where transactions are not under control of any third party organisation. Old data and new data are stored in a sealed compartment of blocks (ledger) distributed across the network in a verifiable and immutable way [121]. Information security and privacy are enhanced on the way in which data is encrypted and distributed across the network.

Therefore, this PhD work develops a decentralised secure and privacy-preserving e-Government system by innovatively using blockchain technology. The developed secure and decentralised e-Government system is based on the consortium type of blockchain technology, which is a semi-public and decentralised blockchain system consisting of a group of pre-selected entities or organisations in charge of consensus and decisions making for the benefit of the whole network of peers. Ethereum blockchain solution [23] was used in this project to simulate and validate the proposed system since it is open source and supports off-chain data storage such as images, PDFs, DOCs, contracts, and other files that are too large to be stored in the blockchain or that are required to be deleted or changed in the future, which are essential part of e-Government systems.

This PhD work also develops an IDS based on the Dendritic cell algorithm for detecting cybersecurity attacks to support the proposed blockchain-based e-Government system, because the blockchain database is append-only and immutable. Note that, once the information is added to the blockchain database cannot be deleted or changed in the future [7]. The proposed DCA-based IDS effectively prevent unwanted transactions such as virus, malware or spyware from being added to the blockchain-based e-Government network. Briefly, the DCA is a class of artificial immune systems which was introduce for anomaly detection in computer networks and has beneficial properties such as self-organisation, scalability, decentralised control and adaptability. Three significant improvements have been implemented for DCA-based IDS. Firstly, a new parameters optimisation approach for the DCA is implemented by using the Genetic algorithm. Secondly, fuzzy inference systems approach is developed to solve nonlinear relationship that exist between features during the pre-processing stage of the DCA so as to further enhance its cybersecurity attacks detection performance in e-Government systems. In addition, a multiclass DCA capable of detection multiple attacks is developed in this project, given that the original DCA is a binary classifier and many practical classification problems including computer network intrusion detection datasets are often associated with multiple classes.

1.4 Goals and Objectives

This PhD project aims to 1) develop a decentralised secure and privacy-preserving e-Government system; 2) develop an IDS for detecting and mitigating cybersecurity attacks targeting e-Government systems. To achieve the main goals, the following objectives will be met.

1. To investigate security and privacy issues in the existing e-Government systems.
2. To develop a prototype of decentralised e-Government framework with privacy preservation, and cybersecurity attacks detection functionality, using blockchain and an artificial immune system.
3. To develop a decentralised e-Government system based on the consortium blockchain technology.
4. To develop an IDS based on the Dendritic cell algorithm for detecting cybersecurity attacks in the proposed blockchain-based e-Government system.
5. To evaluate and validate the performances of the proposed approaches in e-Government system by applying real-world cybersecurity datasets and other anomaly detection benchmark datasets.

1.5 Structure of the Thesis

The structure of the remainder of the thesis is outlined in this section. Briefly, the works carried out in Chapter 1 and Chapter 2 achieved the first objective. The work presented in Chapter 3 accomplished the second objective. Chapter 4 is linked to the third objective. The fourth objective is achieved by the works detailed in Chapter 5 and Chapter 6. And the performance of the proposed approaches in e-Government systems, which is the fifth objective, is evaluated in Chapter 4, Chapter 5 and Chapter 6.

Chapter 2: Background

Chapter 2 provides an in depth background of the emergence of e-Government systems, blockchain technology, intrusion detection systems and artificial immune systems. In particular, privacy and security issues in e-Government systems is reviewed. Public-key and Symmetric-key cryptography which are the essential components for the implementation of the blockchain technology are discussed in this chapter. Additionally, types of blockchain technology and their advantages and limitations are discussed. Finally, DCA algorithm is reviewed as an artificial immune system used to develop the IDS in this PhD project.

Chapter 3: The E-Government Framework

Chapter 3 proposes a decentralised e-Government framework with privacy preservation, and insider and external threat detection functionality, using blockchain technology and the DCA algorithm. The proposed e-Government framework is comprised of three main modules. Firstly, a decentralised e-Government module is comprised of a P2P network with each node representing a public department based on the blockchain technology. Secondly, an external attack detection module based on the DCA detects unexpected traffics coming from the Internet to the e-Government system for further investigation by the network administrator. Thirdly, an insider threat detection module based on the DCA identifies internal anomalies from legitimated accounts of the e-Government system for further investigation. The theoretical and qualitative analysis on security and privacy of the proposed framework shows that, encryption, immutability and the decentralised management and control offered by the blockchain technology can provide the required security and privacy in e-Government systems. Insider and external threats associated with the blockchain transactions from users are detected and reported by the DCA-based IDS to avoid any invalid operations to the blockchain database. Thus, it can be applied in Government organisations to implement a decentralised and secure e-Government systems to overcome design challenges such as interoperability, integration and complexity. Additionally, this framework has the potential to increase citizens' trust in the public sectors. The work in this chapter has been published in [48, 49].

Chapter 4: Consortium Blockchain for E-Government Decentralisation and Privacy-preservation

Chapter 4 proposes a decentralised e-Government system based on the consortium blockchain technology. The consortium blockchain is particularly chosen in this project because it has moderate computational cost which is crucial for e-Government systems. The consortium blockchain allows decentralised and flexible information access control, where the accessibility of the information stored in the consortium network can be limited to validators (e-Government departments), authorised users (registered citizens and shareholders), or not limited at all (public information). Also, unlike public blockchain where consensus process and transaction audit are carried out by all nodes with high computational cost, consortium blockchain performs the consensus process using pre-selected trusted nodes with moderate cost. The proposed decentralised system was simulated and evaluated by using Ethereum Visualisations of Interactive, Blockchain, Extended Simulations (eVIBES simulator) [36].

eVIBES simulator was selected because it is open source and supports off-chain (sideDB) data storage such as images, PDFs, DOCs, contracts, and etc; and these are essential part of e-Government systems. The performance evaluation based on the number of transactions processed per second and on the time for processing a single transaction by varying the number of nodes (validators) in the consortium blockchain network have proved that, the proposed system is suitable for decentralisation, security and privacy assurance in e-Government systems. consortium blockchain technology provides the decentralised environment and control required in the proposed e-Government system. The work in this chapter has been published in [52, 49].

Chapter 5: E-Government Security and Privacy-Preserving Using Enhanced Dendritic Cell Algorithms

Chapter 5 develops three different cybersecurity attacks detection systems based on enhanced DCA for identifying and mitigating unwanted traffics in e-Government systems. Firstly, a new parameters optimisation approach for the DCA was implemented by using GA; since the original DCA uses manual method to pre-defined the weights for its objective function. Secondly, fuzzy inference systems approach was used to developed an approach which can solve nonlinear relationship that may exist between input features and the resultant three DCA's signals during its pre-processing stage. Thirdly, a new signal categorisation method for the DCA was proposed based on Partial Shuffle Mutation of GA to automatically categorise the input features into the three DCA's signal categories; given that the original DCA uses manual categorisation technique based on domain or expert knowledge of the domain. The experimental results show that the enhanced DCA approaches are capable of detecting cybersecurity attacks in e-Government system with effective performances while simultaneously ensuring privacy to blockchain transactions and data. The work in this chapter has been published in [53, 46, 45, 50, 47].

Chapter 6: E-Government Multi-Attack Detection using Multi-Class DCA

Chapter 6 proposes a multi-attack detection system for e-Government system by transforming the binary DCA to support multi-class classification (McDCA). The McDCA was implemented by generalising the natural behaviors of DCs to allow multiple situations to be considered rather than simply normal and anomaly. To further prove the potential of the proposed McDCA, a multilayer McDCA is also proposed and simply implemented by allowing the use of DCs in a layered structure; this ultimately opens the door for its further extension to be implemented as a deep learning approach. The experimental results based on the implementation of the proposed McDCA and multilayer McDCA demonstrated the working and efficacy of the system, with overall better performance than those from the commonly used and recently proposed conventional multi-class classifiers. The results obtained by using real-world cybersecurity datasets with multiple attacks prove that, McDCA is able to simultaneously detect multiples attacks targeting e-Government system. Hence, compared to the binary DCA which groups all attacks in the same class as anomaly, the McDCA can identify and report each individual attack independently in e-Government system. The work in this chapter has been published in [44, 51].

Chapter 7: Conclusions

Chapter 7 concludes the thesis and points out the possible future works including short-term and long-term developments.

Appendices

Appendix A lists the publications arising from work presented in this thesis.
Appendix B lists the Acronyms/Abbreviations used in this thesis.

Chapter 2

Background

This chapter presents an in depth background of the emergence of e-Government systems, blockchain technology, intrusion detection systems, and artificial immune systems. In particular, privacy and security issues in e-Government systems are reviewed. Also, public-key cryptography, symmetric-key cryptography, digital signature and cryptographic hash function which are the essential components for the implementation of the blockchain technology are reviewed. Additionally, types of blockchain technology and their advantages and limitations are provided. Finally, the background of the DCA algorithm which is a class of AIS used in this thesis to develop IDS is presented.

The rest of this chapter is structured as follows. Chapter 2.1 discusses the emergence of e-Government systems as well as security and privacy issues in the existing e-Government systems. Chapter 2.2 details the background of blockchain technology and cryptography. Chapter 2.3 presents the background of intrusion detection systems. Chapter 2.4 provides the background of artificial immune systems and biological immune systems, and finally Chapter 2.5 summarises the chapter.

2.1 The Emergence of E-Government Systems

Over the past few decades, the rapid advancement of the ICTs has accelerated the development of many electronics services. Electronic services available online to users are often denoted with a prefix *"e"*, for instance e-learning for distance online learning services, e-banking for electronic banking services, e-commerce for electronic commerce services, e-business for electronic business services, and etc. E-services became possible after the internet appeared to be the main media for information exchange and the introduction of World Wide Web (WWW) in 1990s. Commonly, since 1990s, Web-based services and application started to become an integral part of e-Government for delivering services and information to the public particularly in advanced developed countries such as USA, UK, Germany, Canada, Australia e.t.c [155, 134]. Development and adoption of e-Government systems in public sector was hugely motivated by the advancement of e-commerce systems and shifting of economy from goods to services through utilisation of ICTs [66]. Additionally, it is becoming a mandatory for most countries across the world to use digital communication between citizens, businesses and Government in order to flourish and survive in the digital economy era [155, 158].

There are many definition of e-Government systems in the literature [122, 123, 155]. Throughout this thesis, e-Government system is defined as the use of ICTs within the public sectors to improve services delivery to citizens and other stakeholders; thus making the public sectors more transparent, participatory, accessible, accountable, efficient and effective. E-Government systems are designed and equipped to offer a range of services and information, including licenses registration and renewal, tax filling, voting registration, passport and visa application, useful information to the public, business and employment opportunities etc [122, 155].

According to the United Nation report on e-Government development, almost every nation around the world has developed a website for providing information to its citizens and other stakeholders [155]. Good examples are the UK's Government portal (https://www.gov.uk/) and the Singapore Government's eCitizen portal (https://www.ecitizen.gov.sg/). Websites provide Government e-services and information to individuals such as availability of public services and the procedures to follow in order to get those services. Note that, a citizen-centered, business-focused and environment-aware e-Government system can lead to greater transparency and convenience, higher revenue and efficiency, and less corruption and operational overhead [155].

Digital identity (eID) is another online e-Government service that was introduced in order to provide individual's identity for verification while accessing services from different Government departments as well as the legal validity of online transaction inquired [144]. eID provides a simple way for citizens to prove electronically that they are who they say they are, in order to access online services. eID in e-Government helps to distinguish between different citizens and business uniquely. The same eID can be used in multiple sectors (e.g. taxation, social security, education, telephony services, banking services) and while fulfilling different roles (e.g. a civil servant, a lawyer or a father) depending on the context. Also, eIDs and online identities can be used beyond national boundaries to authenticate and authorise citizens to e-services anywhere. The task of issuing and validating eID is assigned to a single organisation which become responsible for distributing to other member state [144].

Due to the sensitivity of information stored in e-Government networks, each department and agency must make sure that only authorised users can get access to the system. Information security technology must be provided in order to safeguard and maintain smooth operation of e-Government system. Protection of e-Government networks ensures confidentiality, integrity and availability of the data [122]. Any e-Government system is vulnerable to security breaches if no security policy, security mechanism or countermeasures are prepared and put in place [59, 144]. Note that, information collected from individuals are stored in centralised databases and servers in the existing e-Government systems [155, 170, 122].

2.1.1 Categories of E-government Systems

The kind of services and information shared in e-Government systems can be classified as feedback and opinions, public information, critical information, business information and personal data. Therefore, based on the interaction and inter-relationship between e-Government and individuals, they are often categorised into four groups namely Government to Citizen, Government to Business, Government to Government and Government to Employees [155, 134]. Each of these categories is describes as follows.

Government to Citizen

G2C is the interaction between Government and citizens by using online electronic applications such as Websites and mobile applications. G2C communication helps citizens to access high quality services and information from the Government in an effective way and efficiently manner [158]. G2C disseminates information to individuals, such as driving licenses renewal, application of birth/death/marriage certificates, online payment of income tax, and etc.

Government to Business

G2B is an interaction between Government and business firms through the Internet in order to provide transparent environment for businesses in a particular country. G2B interactions help to reduce the expenses to Government of buying and selling services and goods from private businesses. G2B exchanges information between Government and businesses such as policies, rules and regulation, online application of business permits, and so forth. In G2B, the Government services are provided to private firms according to the established rules and regulations but in more modernised form of economy [122].

Government to Government

G2G indicates an interaction between Government departments, authorities and agencies with one another at local, regional or national level in order to share the information and services available among public bureaucracies and with individuals [5]. It also entails the international interactions between a Government and other Governments around the world. More precisely, G2G is the computerisation of routine tasks performed by the Government while allowing automatic sharing of service and information between departments and agencies. G2G is designed to work according to the principles and rules governing the public sectors while delivering online services.

Government to Employees

G2E means the interaction between Government and its employees by using online application in order to make communication more effective and efficient. G2E interaction enhances the productivity and transparency with the public sectors by allowing Government employees to access various Government information and services while simultaneously encouraging information sharing with the public. For example, online management of the public employees' payroll, social security services and pension. Additionally, G2E is concerned with the sharing of the public documents among employees of the Government. It can also be referred to as intra-Government interaction as it coordinates the transactions which are essential to staff working within the public sectors.

2.1.2 Benefits of E-Government

Local, regional and state Governments across the world are constantly spending large amount of money on ICTs projects to develop electronic applications that increase efficiency of e-services provision to the public [123, 155]. Electronic services provision by Governments to citizens and other stakeholders has a number of benefits as it has been pointed out in different studies [122, 6, 123, 153]. These benefits are summarised below.

- Provides a better way of managing information compared to traditional paper-based method. For instance, website is a cost effective way of sharing information between the owner and users. Also, it is quicker to publish information on website than sending to individuals one by one.

- Improves efficiency of public services delivery and facilitates compliance with Government policies and regulations.

- Strengthens wider citizens participation in public sectors by involving them in the process of decision making such as e-voting. This also increase the level of trust in public sectors.
- Creates a better and cost-effective business environment in a country by simplifying interaction between businesses and Government which leads to cost reduction and revenue growth.
- Increases accountability in the public sector which enables the Government to meet its citizens' expectation through improving the quality of services delivery.
- Helps to reduce bureaucracy and corruption in the public sector through transparency. Online communication between individuals and public officials provide a painless process to navigate bureaucracy and corruption within a state. Transparency make Government departments and agencies more responsible as they know that every process and action is closely monitored and recorded.
- Makes Government data more accessible to citizens and other shareholders. Also, e-Government systems provide more insight into public data and better control of Government activities.
- Increases democracy as a result of citizens participation in decision making at all levels of Governance. Citizens can actively participate in online forums regarding development and political issues [122, 123].
- Speeds up information sharing between individuals and public sectors. Web and mobile applications enable instant transmission of high volume of data across the country and over the world at any time of the day.

Note that, the benefits of e-Government are the same in both developed and developing world [122, 155]. The ability of a Government to make public services accessible online to individuals irrespective of their locations across the country offers the biggest benefits of e-Government systems.

2.1.3 Security Issues in E-Government Systems

Although the adoption of e-Government systems in the public sectors provides efficient and effective services over the Internet, security issues remain a major concern [123, 88]. The most common cyber-attacks to e-Government systems are DoS, DDoS, unauthorised network access, theft of personal information, website defacement, application layer attacks such as cross site scripting (XSS), and penetration attacking [14, 127, 115]. Individuals of the public are not ready to engage with electronic public services due to a lack of trust which is identified as a significant barrier to the adoption of e-Government systems [127]. E-Government users perceive Government as one entity, which means security issue that affect one department (agency) may be viewed as a threat to the whole e-Government system [88]. Failing to secure public users' data has both financial and legal consequences. Financial consequences is a result of business partners losing trust while legal consequences is due to Government failing to address security of the data they collect from its users.

The existing e-Government systems have been identified to be faced with the potential security vulnerabilities and suffer from a single point of failure due to centralised databases and servers [123, 170, 153]. Generally, it is very hard to respond to flooding-based DoS and DDoS attacks in centralised systems due to a large number of malicious traffic which are sent to the network to render it inoperable. The next generation of e-Government systems will be required to integrate with services such as geospatial information, regulatory publications and public deliberation data, which enhances the experience and the innovation of applications [151]. As technologies advance, the number of attacks increase as cyber-criminals are coming up with new attack methods. Threats to e-Government will also be more complex targeting the client end points, the communications infrastructure, and the back end servers [140]. According to [5], security of e-government systems must be protected to increase users' trust while accessing the public services online. If e-Government systems are not well secured, cyber-attacks are inevitable.

As ICTs advance, new devices are getting connected to e-Government systems which increase a greater chance of a vulnerable devices being added to the network and open up an attack space [127]. Due to the nature of these devices, security issues is a major concern. “Traditional" cyber security solutions may not provide the required level of defense [56]. Cyber criminals are coming up with new sophisticated attack techniques every day making it difficult to predict the kind of attacks e-Government systems will be subjected to in the future. This is more of a concern with the introduction of devices that have the ability to establish communication links without a user's intervention, such as IoT devices. Cyber-attacks can take several forms, some may causes system damage, disruption to a communication infrastructure or extract sensitive information. In the past, a research study conducted by [115] found that, more than 80% of e-Government web sites around the global were vulnerable to cross-site scripting (XSS) and structured query (SQL) injection due to lack of proper authentication mechanism applied to input data from users. Due to cyber warfare, there are other new motivations for attacks such as political differences, extortion, cyber terrorism, and even contests for the supremacy which can occur within a nation or between different nations [128].

According to [86], security measures can be implemented at the physical, technical or management levels. Physical security includes safeguarding e-Government network equipment, data, information and other valuable assets from being destroyed by operation mistakes, natural disaster, computer crime or any other attempt that can cause physical destruction of assets, loss of information or interruption of the system operations. Technical security is achieved by using computer network products such as firewalls, IDS, Intrusion Prevention Systems (IPS), secure routers and switches as shown in Figure 2.1 to secure E-Government system against many cyber-attacks. Computer security devices are configured to filter network traffic into and out of e-Government systems (Router and Firewalls) and also to identify any signs of suspicious activity (IDS), this is an area where AI has been exploited. Management security measure focus on the setting up of policies, regulations, and legal protection for the purpose of easing the integration of e-Government management and technology while guaranteeing the security of e-Government systems.

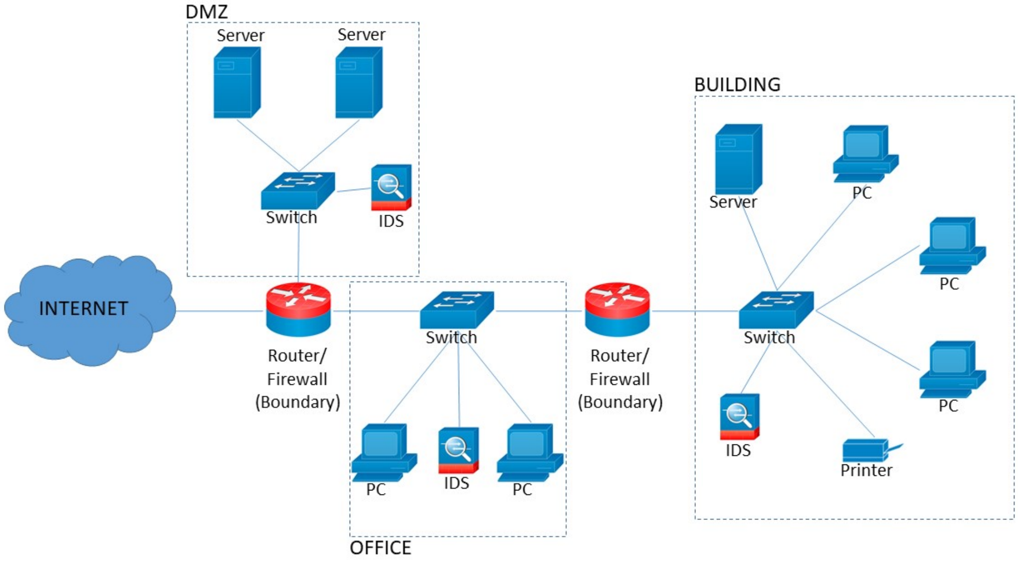


Figure 2.1 E-Government technical security measures

Different non-technical e-Government security maturity models have been proposed for guiding and bench-marking the security implementation of the e-Government system. For instance, a comprehensive e-Government information security maturity model for guiding the inclusion of security in e-Government systems was reported in the work of [90]. This model focuses only on the organisation's security mechanisms setting, security evaluation, security policy setting and information security awareness to users, but it lacks guidance for built-in security that can ensure e-service security and privacy.

2.1.4 Privacy Issues in E-Government Systems

Privacy concerns are identified as a key challenge to policy, regulatory and legislation in the 21^{st} century [130]. Generally, e-Government systems perform three basic operations: data transfer, data processing and data storage [9]. Privacy may be violated in any of these operations. User's personal information such as their identity, medical records, and other sensitive information could be disclosed through metadata analysis. If an attacker can compromise the privacy and security of e-Government system and access the information then the confidentiality, integrity and availability, will be in jeopardy due to connected devices storing large amounts of sensitive personal information i.e., photocopier hard drives, printers and scanners, mobile phones etc. It is the responsibility of a Government to ensure an individual's information is secured and the users' privacy is preserved during data collection, processing, storage, and exchange. Therefore, the e-Government infrastructure and all the devices connected to the infrastructures need to be protected with appropriate measures.

One major concern with respect to Government systems is the capturing and use of personal and sensitive information, and there are fears that the information could be used to monitor the public, which is considered by many as an invasion of their privacy [160]. There are also fears that the information could be obtained by cyber criminals. These concerns are well founded as there are a number of well documented cases where information has been leaked in e-Government systems [159, 148, 154, 30, 84]. The compromise of information privacy in e-Government systems usually leads to the loss of citizens' trust and confidence [102]. It is therefore of ultimate importance to ensure the privacy, confidentiality and integrity of e-Government systems. Thus, much need to be done to come up with a better privacy and security solution for e-Government systems because new security breaches are being reported every year especially on cyber-attacks targeting citizens' sensitive information.

The e-Government systems could improve privacy and security of information by introducing Transport Layer Security and using a Secure Socket Layer (TLS/SSL) certificates which implement a public key infrastructure (PKI) [102]. PKI requires that users maintain their software to ensure the latest TLS/SSL certificates are being used. PKI's use a trusted third party, a certification authority (CA) to offer certificates to the user's devices. If CA is compromised it may lead to privacy and security problems[55]. Setfanova et al. propose that biometric security should be incorporated in e-Government portals using either fingerprint, iris or facial recognition for authentication [144]. However, biometric technology can be expensive and difficult to implement.

An authentication framework known as Greek Authentication Framework (GAF) [131] can ensure security and privacy of E-Government users by applying different registration and authentication procedures using a single central public portal interfaced with ministerial departments (service providers). The framework consists of two parts namely Identity Provider (IdP) and the Service Provider (SP). This is the same process as used with Shibboleth authentication which is extensively employed in Higher Education (HE) and research in the UK [87]. Users have to register with an IdP at a central portal that is then used to access services from a service provider. Since all users are administered by a single central portal, it must be configured in such a way as to not present a possible single point of failure. This framework uses PKI to ensure confidentiality but having a single authentication center limits the integrity and availability of data in case it is compromised.

It is the responsibility of a Government to ensure an individual's information is secured and the users' privacy is preserved during data collection, processing, storage, and exchange. Therefore, the E-Government infrastructure and all the devices connected to the infrastructures need to be protected with appropriate measures. Incorporating new technologies such as the blockchain technology in e-Government systems will ensure privacy and security of information between devices, users and Smart Systems [15]. Note that, cybersecurity and privacy attacks and information breaches cannot be eliminated entirely despite the advancement of available technologies. As well as the technological issues that must be addressed, the public must also be educated to keep their data secure and safe through information security awareness programs such as https://www.getsafeonline.org/, http://www.safetynetkids.org.uk, https://www.infosecawareness.in/ and https://staysafeonline.org/.

2.1.5 Cybersecurity and Privacy Threats

Cybersecurity and privacy threats which are consistently targeting networked systems such as e-Government fall into two categories namely insider and external attacks. Insider threats are malicious actions performed by insiders within an organisation through their authorised account with a motivation of causing information theft, electronic fraud or system sabotage. Insider threat manifests in many forms such as disgruntled employee, consultant or officer within the organisations [67, 93]. Organisation networks are almost always secured by using

IDS and firewalls, but insider threats cannot be detected by these externally-facing security measures; this is because insider threats always originate from trusted accounts [67]. The longer an insider threat incident occurs without being detected, the more costly it gets [161]. According to a recent Insider Threat Survey Report of 2019 [93, 161], 20% of cyber security attacks and 15% of information theft originated from insiders within an organisation, with a single insider costing an organisations a loss of around $11.45 millions annually.

Due to the nature and sophistication of insider threats, the detection of such threats is considered as a very challenging task in organisations in all sizes. Nevertheless, suspicious activities from insiders are commonly used as early stage warnings for potential insider threats [67, 93], such as using the organisation network to download or access large amounts of sensitive data, and using the organisation network to copy files from sensitive folders. In addition to these traditional pre-caution measures, machine learning and artificial intelligence techniques, such as support vector machines and deep learning, have been developed as a promising solution that can be used to detect, contain and deter insider threats if designed appropriately [93].

External attacks are incoming network traffics that deviate from what is set as a normal network behavior, and usually performed by an outsider who wants to gain access to the network resources illegally [67, 94]. Such attacks and anomalies include malware, keyloggers, network scan, spying, DoS, DDoS, Ramsonware, and etc, which can cause massive damages to e-Government services and applications. External attackers can use port scanning technique to gain access to a computer system and its files by exploiting the weak points available through surveillance. Buffer overflow and rootkit attacks are also used by unauthorised users to gain super user privileges by exploiting vulnerabilities that allow normal user to gain a root privileges. Additionally, unauthorised user can use password guessing techniques to access of a computer resources from a remote machine. Similarly, malware, spyware, keylogger and Trojan horses can be used by attackers to facilitate and gain privileged access to the system. Attackers also can use IP spoofing (modify network configurations) to generate an IP address similar to the actual address known to the system so as to trap the system that it is communicating with an authorised user and, therefore, to grant access. Encryption and IDSs are the common measure to prevent and detect external attacks in a given computer network [93]. Many artificial intelligence techniques, such as fuzzy interpolation [172, 120], AIS [50], artificial neural networks [69], are employed to develop IDSs.

2.2 Privacy and Security Technologies

This section reviews cryptography and blockchain technology and highlights how the blockchain has appeared to be one of the core technologies for secure data sharing and storage over trustless and decentralised systems. Firstly, cryptography is presented since it is the underlying technology behind the working mechanisms of the blockchain technology.

2.2.1 Cryptography

Cryptography is the study of hiding and securing information from being revealed to an adversary such an eavesdropper [91, 113, 61]. It encompasses a number of techniques such as encryption, decryption, digital signature, key distribution, and etc [142]. The main goal of cryptography is to provide a protected and accurate way of transferring sensitive information over an insecure communication channel while simultaneously ensuring that the shared data is kept in a secure storage. More precisely, cryptography provides a secure communication over an insecure channel such as the internet or a cell phone. In cryptography, privacy is preserved by encoding a message through encryption, while the message is decoded by using decryption.

Note that, cryptography is the key component of network security as it ensures the safe transmission of data across the computer networks. The three fundamental goals (i.e., *Big Three*) of network security are Confidentiality, Integrity, and Availability (CIA) [142, 61]. Confidentiality ensures that the shared data is available only to the intended and authorised entities. The Integrity ensures that the shared data is reliable and is not changed by unauthorised entities. The role of Availability is to ensure that the shared data and communication channel are continuously available to the authorised users, on demand.

Public-Key cryptography

In public-key cryptography (also known as asymmetric-key), the sender and receiver use two different keys for encryption and decryption [142]. The two keys are in a pair referred to as the public key and the private key. Conventionally, in public key encryption, all parties interested in secure communications are required to publish their public keys. Thus, if Party A confidentially wants to communicate with party B, A encrypts a message by using B's publicly available key. Then, since only B has access to the corresponding private key, such a communication can only be decrypted by B.

Public-key encryption can also be used to provide authentication. For instance, if Party A wants to send an authenticated message to party B, A encrypts the message with his/her own private key. Then, since this message can only be decrypted with A's public key, that establishes the authenticity that A was indeed the sender of the message.

Mathematically, in public key cryptography, to generate private and public keys, the first step is to choose two different large prime numbers p, q (assume $p < q$, generally) and then compute $n = p * q$. Secondly, find a pair e and d for the private and public keys to the extent that for a given message M, it gives M^{ed} *mod n* = M *mod n*. Finally, (e, n) is published as the public key and (d, n) is kept as the private key.

Encryption: A message is encrypted by using a modular exponentiation with public key (e, n) as given by the following equation.

$$c = M^e modn. \tag{2.1}$$

Decryption: Is achieved by exploiting the private key (d, n). So, for a given cipher-text c (enecrypted message M), the the original message is recovered as shown in the following equation.

$$M = c^d modn. \tag{2.2}$$

One disadvantage of public-key encryption is slow encryption and decryption speed compared to symmetric-key encryption. Symmetric-key encryption uses a single key to encrypt and decrypt a message which speeds up the process. Contrarily, public-key encryption uses two different keys to encrypt and decrypt a message which are derived by using a complex prime number factorisation process which take more time and computer resources. Additionally, if the private key is lost, the received messages encrypted by using its corresponding public key could not be decrypted. Thus, managing the keys is not easier compared to symmetric key encryption.

Symmetric-Key cryptography

In symmetric-key encryption, the encryption key (also known as secret key) is known to both sender and receiver [142]. The decryption of a message is done by using the same key used to encrypt the message. Therefore, only one key is required to encrypt and decrypt a message.

For instance, if Party A wants to confidentially communicate with party B, A encrypts a message by using the shared secret key. Then, B decrypts the message by using the same shared secret key. Although the symmetric-key cryptography is faster than asymmetric-key cryptography during encryption and decryption, but it suffers from one major limitation where the key needs to be kept secret. This can be challenging since the key is required to be moved safely between two locations for encryption and decryption to take place.

Digital Signatures

A digital signature is created by using public-key cryptography. It is used to prevent an adversary from creating a message and present it as written by an authorised entity (impersonation) [61]. It is also used to authenticate the identity of the sender and to detect unauthorised modifications to information. Thus it provides authentication, integrity and non-repudiation to a message.

Digital signal operation requires a signing algorithm and a verification algorithm [142, 61]. A sender uses his/her private key with the signing algorithm to create a digital signature from a message. Then, the receiver uses the public key of the sender with the verification algorithm to verify the sender of the message. Some of its practical application include verifying websites such as Facebook.com. For instance, given a message M, a signature S of party A, a signing algorithm sigA(M) and a verification algorithm verA(M), the relationship between them is given by Equation 2.3.

$$S = sigA(M, private\_key) = verA(M, public\_key). \tag{2.3}$$

Cryptographic hash function

Cryptographic hash function is a mathematical function which is used to transform an arbitrary length message to a fixed-length m-bit output known as hash [61, 142]. The motivation behind the hash function is public-key cryptography such as elliptic curve cryptography (ECC) and the RSA (Rivest-Shamir-Adleman) [61]. It was proposed after realising that encryption of data is not sufficient to protect its authenticity. Its main goal is to enforce the integrity and authenticity of messages. Examples of cryptographic hash functions include SHA1 (Secure Hash Algorithm), MD5 (Message Digest), SHA256, and etc [61]. Note that, cryptographic hash functions are known to the public.

The following are the fundamental properties of a good cryptographic hash function, **H()**:

- It takes on input of any size.

- It produces a fixed-length output.
- It is easy to compute (efficient).
- Given any hash value h, it is computationally infeasible to find any value $\boldsymbol{y}$ such that $\mathbf{H}(\boldsymbol{y}) = h$.
- For any given value $\boldsymbol{y}$, it is computationally infeasible to find another value $\boldsymbol{x}$ such that $\mathbf{H}(\boldsymbol{x}) = \mathbf{H}(\boldsymbol{y})$ and $\boldsymbol{x} \neq \boldsymbol{y}$.
- It is computationally infeasible to find any $(\boldsymbol{y}, \boldsymbol{x})$ such that $\mathbf{H}(\boldsymbol{y}) = \mathbf{H}(\boldsymbol{x})$ and $\boldsymbol{x} \neq \boldsymbol{y}$.

Therefore, a cryptographic hash function is a one way function since it is hard to invert [142, 61]. A change of one bit of input message, causes changing of the output hash. Hashing is commonly used for password protection and message authentication in computer networks and thus protects against intentional or unintentional modifications.

2.2.2 Blockchain Technology

Blockchain is a P2P distributed database (i.e., ledger) which maintains a list of continuously growing records called blocks that are linked linearly and chronologically and secured by using public key cryptography and cryptographic hashing [121]. By the blockchain technology, new information is added to a block and becomes available to all nodes in a distributed network, rather than adding to the centralised database in the traditional centralised system. Conventionally, every time a set of new transactions is submitted to the blockchain network, a new block is created for storing the transactions and then becomes another block in the chain and hence the name *"Blockchain"*.

Although blockchain technology was initially introduced for the purposed of sharing digital currencies, nonetheless, is not limited to financial transactions but can be programmed to record any kind of information and data. Each block in a blockchain is identified by a hash value generated by using the secure hash cryptographic algorithm-256 bits (SHA256) [121]. The hash value of a current block header (parent) is linked and stored in the next block (child) as depicted in Figure 2.2 [7]; therefore, if there is an alteration in any block's content, its hash will also change accordingly and the change will be propagated throughout the network to invalidate that block [121]. Based on this mechanism, the blockchain technology does not require an intermediary or trusted third party as it is decentralised and distributed.

The blockchain participants have private keys assigned to them to digitally sign and validate the transactions they make. Additionally, the blockchain is immutable, and thus, once the data is entered in the chain, it cannot be erased or tampered with. Since all transactions are linked together and shared across the network, in order to hack the blockchain network, the attacker must hack not just only one computer, rather, every single computer on the network, which is nearly impossible.

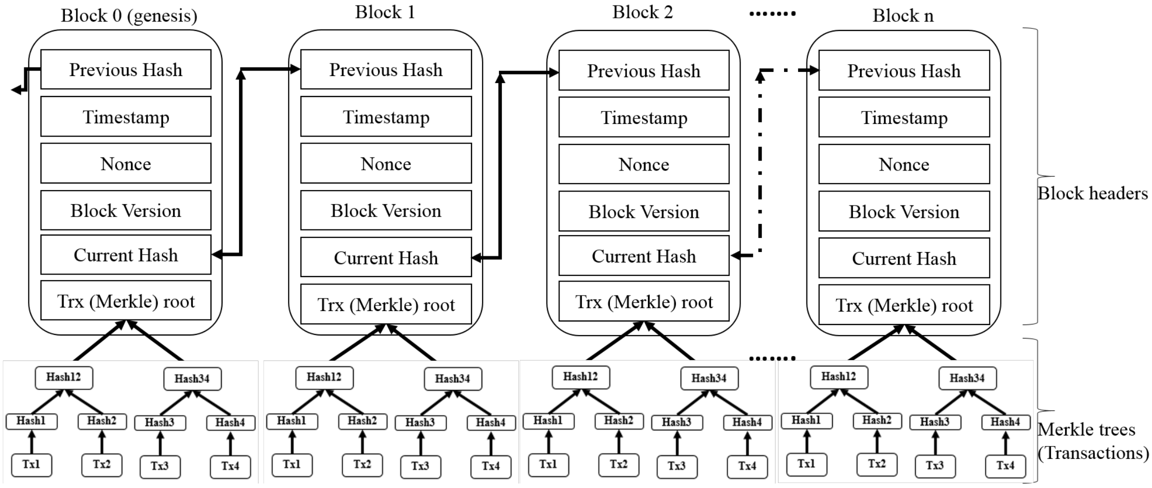


Figure 2.2 An example ledger with details of blocks

As shown in Figure 2.2, a block is composed of a header containing the meta data, and a long list of transactions performed in that block. The block header contains the timestamp, nonce, version and proof of difficulty. Timestamp indicates the time a block is created; nonce is a random number generated by the consensus algorithm to compute the hash value of a block, version indicates a version number of the blockchain, and proof of difficulty is a generated hash value which must be less than the current target hash value.

The first block, known as a genesis block, is hard-coded by embedding some random data into the blockchain application [121]. Although each block has only one parent and child, a valid block may have two or more children temporarily created when two or more nodes (network peers) are added to a block at the same time leading to two or multiple branches from the same parent [7]. This situation is called ‘fork’ and is eliminated by taking the chain whichever becomes longer than the others as a valid blockchain, and making all other shorter ones invalid (orphan), with a two-branch situation demonstrated in Figure 2.3. It is possible that the formed branches have the same length; in this situation, the process of adding new blocks continue for all the to-be-validated chains until one branch becomes longer than the others thus valid.

Within a block, all the transactions are linked together using a merkle tree [121]. A merkle tree is an upside down binary tree used by the blockchain technology to summarise

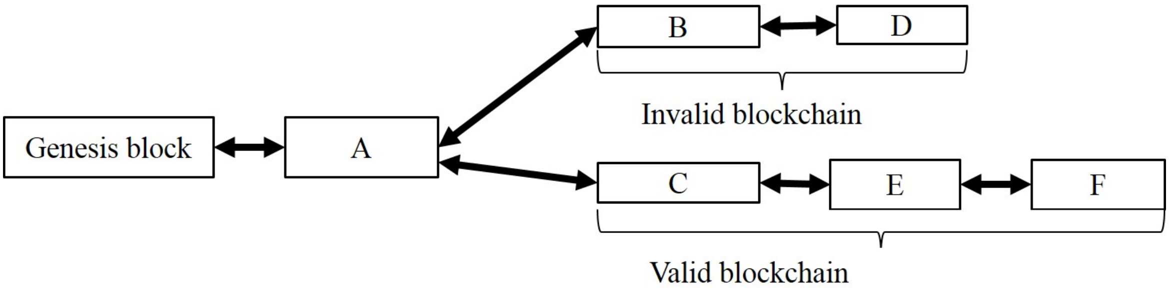


Figure 2.3 Blockchain validation for the fork situation

all the transaction in a block. To construct a merkle tree, a pair of transactions are hashed recursively until they form only one root node at the top of the tree termed as the merkle root [121], as shown in lower part of Figure 2.2. More precisely, a merkle root is the hash of all the transactions that make up a block in a blockchain network. Any tiny modification of the data will change the merkle root hash leading to an invalid record. The cryptographic hash algorithm used to construct a merkle tree is usually implemented by the secure hash algorithm 256-bits (SHA256). If there is an odd number of transactions, the last transaction hash is duplicated to create an even number of transactions thus ending up with a balanced tree.

Example applications of blockchain technologies include Bitcoin (support decentralised cryptocurrencies) [121], Ethereum (support self-executing digital smart contracts) [23], IBM Hyperledger Fabric (support development of general enterprise solution) [24], amongst others.

Note that, the elliptic curve cryptography (ECC) approach is adopted for implementing encryption and digital signature in the Bitcoin and Ethereum blockchain technologies since the ECC offers similar level of security as RSA (Rivest-Shamir-Adleman) but it consumes far less number of bits [136]. For instance, a 256-bit key in ECC offers the same level of security as that provided by the RSA using a 3072-bit key. Shorter key usually means low CPU consumption, low memory usage and fast key generation. These advantages are can beneficial to e-Government systems in facilitating fast creation of the transactions and sealing of the blocks. A summary of key length study between the RSA and ECC is provided in Table 2.1 [136]. 256-bit ECC keys are usually used in blockchain technology as they can provide the required level of security for the majority of applications.

Table 2.1 Comparison between RSA and ECC key lengths in bits

RSA key length	ECC key length	Approx. ratio (RSA:ECC)
1024	160	6:1
2048	224	9:1
3072	256	12:1
7680	348	20:1
15360	512	30:1

Consensus mechanism in blockchain technology

Nodes in the blockchain network run a consensus algorithm to validate transactions. There are several consensus algorithms (protocols) available for the blockchain technology, such as Proof of Work (PoW), Proof of Stake (PoS), Delegated Proof of Stake (DPoS), Proof of Difficulty (PoD) [178, 7], Byzantine Fault Tolerance (BFT) algorithm, and etc. For instance, Bitcoin employs the PoW while Ethereum and Bitshare implement PoS and DPoS, respectively [178]. In PoW, miner nodes which want to add (mine) a new block to the blockchain network must first solve a difficult mathematical puzzle which requires great computational power. The first miner to be able to solve the puzzle adds a new block and get rewards in terms of bitcoins [7].

Unlike the PoW, with the PoS, a node which creates a new block is chosen deterministically depending on its stake (wealth) [178]. PoS saves energy that is required in PoW to solve mathematical puzzle, and only the wealthy of a node (validator) is required to validate the new transactions and blocks. The DPoS attempts to solve the consensus problem by using delegates [178]. DPoS uses a real time voting and reputation system to create a panel of limited trusted delegates who will witness and validate the blocks. The witnesses have the rights to create blocks and add them to the blockchain network, in addition to prohibit malicious nodes from participating in adding blocks. Principally, in PoS and DPoS, stakeholders of the network shares are not expected to deliberately make bad decisions for the network.

There are three main properties of blockchain technology which ensure it robustness namely decentralisation, transparency and immutability. Studies have shown that, an attacker is required to take control of 51% of participants in the blockchain network to be able to change or access the database without consensus [168]. In fact, in order for a hacker to change a single block, he/she must change every single block after it on the blockchain which requires a huge and impracticable amount of computing power. No centralised server or database exists for hackers to corrupt in the blockchain.

2.2.3 Types of Blockchains

Typically, the blockchain technology can be either public (permissionless), private (permissioned), consortium (semi-public and semi-private) or Hybrid as detailed below.

Public blockchain

In a public blockchain, any individual can view, modify, and audit the blockchain without having a single entity in charge of the whole network. The consensus and decision making is reached through a decentralised consensus manner such as PoW in bitcoin [121]. The computation power of the participants of the blockchain network is used to select one participant powerful enough to add new transitions to the distributed ledger. The participants are incentivised every time when adding new transactions to the blockchain network, and thus this motivates everyone to use more computations to get the chance of adding transaction to the ledger. In public blockchain network, the higher the number of users, the more secure the network; as it creates a network of trusted individuals between the participants.

Private blockchain

The private blockchain is owned by an individual organisation who is responsible for granting access to the network for new users. Only few specific individuals in the organisations have rights to validate transactions and blocks, and append them to the blockchain network. It is centralised compared to the public blockchain. Thus, the computation power of the participants of the private blockchain network is not required to be higher like in the public blockchains.

Consortium blockchain

The consortium blockchain consists of a pre-selected set of nodes or computers that are responsible for controlling access to the blockchain network resources [37]. The goal of the consortium blockchain is to eliminate the individual/single autonomy of the private blockchain by having multiple entities or organisations in charge of consensus and decision making for the benefit of the whole network of peers. Since only pre-selected organisations are allowed to validate transactions and consensus, incentives are not necessary in this network. The pre-selected set of nodes make it partially private, partially public and semi-decentralised. More precisely, it provides the benefits of public blockchain in terms of efficiency and scalability while still permit some degree of central safeguarding and monitoring like in private blockchain. The consortium blockchain such as Hyperledger fabric [24] is designed to meet the needs of the enterprises where a group of collaborating agencies exploit the blockchain technology to improve service delivery. All consensus participants of the consortium blockchain are known and reputable, therefore, malicious users cannot join the network freely.

Hybrid blockchain

It combines the benefits of a private blockchain with that of public blockchain and create a secure, transparent and privacy-preserving network of peers. In hybrid blockchain, there is no a group of participants or a single individual who make decision on behalf of the whole network as opposed to consortium blockchain. Participants can freely join the blockchain network and participate in consensus process. Generally, the participants of the hybrid blockchain decide what data should be kept private and which data must be made public transparently when necessary. Thus, participants of the network control who gets access to which information stored in the blockchain. Hybrid blockchain provides transparency to business operations without affecting security and privacy of the information shared among participants. It has a private network for storing information which need to be kept private, although any information created in the network is verified by the public network to maintain transparency.

Blockchain for e-Government and other application areas

Blockchain has been widely applied for security, trustness, and privacy-preservation in many areas, such as IoTs [83], smart home [39], smart city [15], educational systems [152], land registry [133], healthcare [129], although it was originally introduced for exchanging digital currency. Many nations around the world have initiated and completed various blockchain projects to explore the potential of blockchain technology in offering efficient public services to individuals and organisations [89], as summarised in Table 2.2. Each of these projects usually focuses on a particular electronic online service, such as e-residency, e-health, e-land-registration, and each of these countries is developing their own systems. The blockchain-based e-Government systems developed by different countries may lead to difficulties to communicate between national boundaries for international information exchange and collaboration. In fact, these projects are expected to be either permissionless (public) or permissioned (private) [100]. Consortium blockchain is designed to meet the needs between collaborative organisations, which has been exploited in this PhD project for a decentralised, secure and privacy-preserving e-Government framework to support e-Government internationally.

The main goal of e-Government is to increase participation, accountability, transparency, interoperability, efficiency and effectiveness in public sectors [122, 155, 25]. Generally speaking, Government networks can communicate to each other better than business networks, because most of them are connected for transferring information to the public without competition as opposed to banking database systems [155]. Also, compared to the banking systems, in the future, the number of devices using e-Government services will increase dramatically due to the fast evolution of smart homes, IoT, smart cities, and other interconnected networks, since these are part of e-Government systems [15]. Note that, the major features that the blockchain technology offers to its users include greater transparency, interoperability, immutability, efficiency, enhanced security and privacy, decentralised control which leads to less failure, and accessibility [156], which are essential for e-Government systems. Therefore, the blockchain technology can complement the requirements of the e-Government systems.

Note that, to enforce confidentiality in blockchain networks, each user is assigned a wallet for storing his/her records and for facilitating communication with the network upon demand. A blockchain wallet is a digital storage which lets individuals manage and confidentially store their records, account credentials including ID, passwords, private and public keys, and other information associated with their accounts. The wallet has a unique ID, similar to a bank account number, such that it can allow users and organisations to safely and securely transfer and exchange information between themselves. Normally, a blockchain wallet is stored in mobile and web applications, and is accessible by using mobile phones and computers.

2.3 Intrusion Detection Systems

Over the past few years, network attacks have been increasing in number and severity, and thus, IDS have become a necessary measure in addition to the security infrastructure of most organisations [10]. IDSs are software used to intelligently monitor network traffics for suspicious activity, which could be an attack or unauthorised activity such as exploiting vulnerable services in a network, applications layer attacks such as SQL injection, privilege escalation in the system, unauthorised logins, virus attacks, malware etc. IDSs are equipped

Table 2.2 Blockchain-based e-Government projects

Country Name	**Project Description**	**Status**
Estonia [145]	Adopt blockchain technology in electronic ID (eID), E-health and E-Residency.	Running;Initiated in 2014.
Dubai [41]	Exploit blockchain technology to fully power public transactions in every sector by 2020.	Ongoing; Initiated in 2016.
Switzerland [100]	Develop Ethereum-based, uPort-powered e-residency ID in the Swiss city of Zug.	Running, Initiated in 2019.
USA [19]	Create new rules that will enable public sector to use blockchain for security and collaboration.	Ongoing; Announced in 2016.
Luxembourg [100]	Develop a public framework that will allow blockchain applications to be integrated in all sectors.	Ongoing; Announced in 2019.
Canada [17]	Develop e-Government information system based on Ethereum blockchain.	Ongoing, Initiated in 2018.
Mexico [18]	Integrate blockchain in public procurement, agriculture and in finance.	Ongoing, Announced in 2017.
China [100]	Integrate blockchain in e-health, eID and e-Voting.	Ongoing; Initiated in 2016.
France [89]	Support Banks and other firms to develop blockchain platforms to allow secure business transactions.	Ongoing, Announced in 2016.
Russia [89]	Explore blockchain in managing Government documents, E-health services, land and property registry.	Ongoing; Announced in 2017.
Africa [100]	Countries such as Ghana, Kenya, South Africa, Ethiopia, Liberia, Nigeria, are exploring the possibility of using blockchain in agricultural sector, land registry, finance, transport and e-services.	Ongoing; Announced in 2017.
Argentina [16]	Develop blockchain-based eID to improve access to public services to citizens.	Ongoing; Announced in 2019.
Singapore [100]	Exploit Ethereum blockchain in educational institutions to offer digital certificates since 2019.	Running; Initiated in 2018.
Sweden [100]	Explore the use of blockchain in the existing land registry system.	A proof-of-concept since 2016.
New Zealand [100]	Adopted blockchain on E-voting in 2018.	Running; Initiated in 2018.
India [100]	Explore the possibility of using blockchain on E-voting, land registry and e-Government.	Ongoing, Announced in 2018.

with mechanisms to alert the administrator when an anomaly behavior from the traffics is sensed, although, they cannot provide direct protection to the system against attacks. They are usually implemented as a set of trained classifiers that can automatically detect malicious network traffic. For instance, in [114], Meng et al. proposed that using blockchain technology would allow a collaborative intrusion detection architecture to be produced. This IDS framework would have the ability to detect attacks such as DoS by allowing various decentralised IDS nodes to exchange data and information with each other. Each IDS node monitors and records network events and exchanges it with the rest of the nodes to determine any sign of anomaly associated with the traffics. This distributed architecture isn't a new idea as it is already used in Honeynets [141]. The difference would be that the end point nodes in a honeynet are deployed using client/server architecture where the blockchain would exchange information in a peer-to-peer architecture.

2.3.1 Categories of IDS

The IDSs are implemented at two levels, firstly at the perimeter boundaries of each network and secondly within each individual computer (host). This is to address any network traffic and transaction based malicious attacks. There are different ways of categorising IDS based on its deployment point on a network and analysis approach. Based on the deployment, IDSs are divided into network-based (NIDS) and host-based (HIDS) [2] as discussed in the next two subsections.

Network-based IDS

NIDS are used to monitor the incoming and outgoing network traffics in a particular network environment and analyse the traffics for suspicious activities. NIDS are positioned at the entry and exit point of traffics from the private network to the internet so as to capture all the data passing through the network. Network-based IDSs are exploited to detect malicious traffics originating from outside the organisation network such as flooding attacks or port scanning. NIDS systems normally gather information about the whole computer network being monitored. They collect information from the network traffic and packet flow as data travels on the networked system [10]. Usually, the NIDS detects malicious traffic or any anomaly by analysing the contents and header associated with the incoming packets while moving across the monitored network. The NIDS normally comes with stored attack signatures that are rules to define the behaviour of common attacks, although, most NIDS allow network administrators to define their own signatures [10]. Thus, the administrator can customise the NIDS based on a particular individual network's requirements and types of application. The NIDS compare the stored signatures to the packet that it captures so that they can identify malicious traffics.

Although NIDS can only monitor traffic on a specific network segment, their operation do not depend on the operating system that they are installed and running on; rather they listen for all malicious traffics, regardless of the operating system (OS) running on the destination machine. NIDS can be installed as part of a network, then, the packet's information can be collected conveniently with minimal work. Usually, the data that is require to be collected for analysis is the configuration of a network card alone. This is crucial in case the topology of the network is changed or the network resources have been relocated then the NIDS also can easily be relocated and exploited normally as needed. The limitation of NIDS is that, since they detect malicious packets based on their signatures, in case of new attack for which no signatures have been seen before, they cannot be detected and thus potentially causing damage to the system resources. Additionally, NIDS cannot be able to detect traffic moving on other communication devices such as dial-up phone lines [2]. But, NIDS are capable of identifying evidence of certain type of packets such as packet storms, DoS and DDoS, that are not detectable or visible to HIDS.

Host-based IDS

HIDS is installed on a host (single system) to monitor traffics that are originating and coming to that particular hosts for suspicious activity [10]. Apart from monitoring incoming traffics to the host, it can also analyse the file system of a host, users' logon activities and running processes. The host-based IDSs are employed to detect internal threats such as a virus or malware downloaded by users accidentally or deliberately, before they spread inside the whole system. Generally, they are typically installed on a computer that is known to be susceptible to cybersecurity attacks. They work by collecting information about processes accessing the host being monitored. This information and data about each process and event are recorded by operating system logs known as audit trails [162, 10]. The logs are simple text files which are recorded as the processes and events access the operating system. Normally, audit trails and system logs contain data about subject initiating an event or activity, as well as any objects related to that activity. Thus, the HIDS can use the information about the subject to detect which process initiated an event as well as the original identification of user associated with that process.

HIDSs are limited to system logs and audit trails as they mostly rely on them, although the manufacturer of the HIDS do not provide logs so interoperability with the hosting operating system become a necessity. As a result, developers of HIDS are required to alter the operating system kernel design to create access to processes and events data [10]. This may affect performance of the host which is not acceptable by the used running the system. Despite this limitation, logs and audit trails are still exploited to develop host-based IDS due to the lack of access to operating system code by the manufacturer of IDS [3]. In fact, the aim of any operating system is protect its audit layer and the associated detail of the system being recorded. Since HIDS is heavily dependent on hosting operating system logs and audit trails, if there exist any vulnerabilities with the OS, this will also weaken its integrity and performance. Note that, HIDSs are very useful since they can be used to keep track of the behavior of individual users within an organisation. HIDS are able to detect an attacker attempting to bypass the host through a dial-up connection while the NIDS cannot. HIDS can examine the command being executed on the host system whether it is malicious or violate any security policy while NIDS cannot do the same.

2.3.2 Classification of IDS

IDSs are usually implemented as a set of trained classifiers that automatically detect malicious network traffic. Based on their analysis approach, IDSs broadly fall into two classes, misuse-based (MIDS) and anomaly-based (AIDS) [10] as discussed in the next two subsections.

Misuse-based IDS

In MIDS, the signature of the known attacks are captured and stored in a database and the attacks are detected by matching their behaviours with the stored signatures [162]. Thus, MIDS searches for specific attacks that are already documented in a large database of attack signatures usually using some pattern matching techniques [143]. As only known attack signatures are stored, the problem with misuse detection resides on being static that it cannot detect novel attacks, and therefore, it has very high false positives. A problem can arise when system software on devices is not kept up to date and a vulnerability or exploit is exposed. A exploitable device in a network could create an entry point to an entire system. If the device is part of a system then the whole network is vulnerable.

One solution to maintain the effectiveness of MIDS is to patch security vulnerabilities. More precisely, the only solution to this type of IDS is for regular or automatic software updates to be carried out. As vulnerabilities and exploits are identified by security specialists they are cataloged in a database. They are then ranked based upon threat severity on a scale of 0 - 10. 0 is the lowest threat level, 10 is the highest [162]. However there is a caveat on automatic patching. The patches should always tested in an offline development environment before applying them to a live system. A patch may make changes that require intervention or may change the operation of a device e.g. a protocol version change may interfere with device interoperability. Another solution to support the MIDS is to employ strong authentication and encryption schemes to make it difficult for unauthorised applications to access the data stored on computer system using passwords and encryption methods. Before any application is allowed onto an the system it is checked by the computer operating system for malicious attempt. So, all processes and sessions initiated to the computer system must be authenticated before transmitting data to ensure security. To reduce the possibility of the encryption being compromised public/private key encryption are be used. The keys are be provided by a third party to provide non-repudiation and an assurance that keys are genuine. These types of services can be provided by companies such as Verisign^{TM} or Thwate^{TM}.

Anomaly-based IDS

Contrarily, AIDS uses a dynamic approach by setting up a set of rules while considering the abnormal activity on the network. More precisely, AIDS defines the behaviours of normal activities, then any traffic's behaviors which deviate from the pre-defined behaviours are treated as anomalies [162]. AIDS can react to new attacks like zero-day if they have abnormal behaviors compared to the normal traffic although not all abnormal traffic are malicious, so this approach usually leads to false negatives if the model is not well trained [2]. Intelligent and classic anomaly-based IDS can be designed using artificial immune systems and other artificial intelligence techniques [20, 95, 172, 29].

Machine learning algorithms have been used to develop IDSs for the identification of abnormal network traffic, for example intrusion/threat detection [120, 107, 172, 34], DoS [172], and email phishing [62]. These detection methods are usually developed by training machine learning algorithms using data sets that include both normal and adversarial activity that describe the patterns of behavior for both normal users and attackers. The algorithms are then applied to test data sets, that are not part of the original training data, to verify the ability of the algorithm in identify normal and adversarial activity. This is similar to the identification of phishing or spam emails using the bag of words (BoG) approach [114]. The BoG relies on a set of keywords (or in ML terms “features") which are used to distinguish between normal and spam emails. The BoG is usually constructed from historic emails (normal and spam). According to Huang et al. existing intrusion detection solutions using machine learning cannot eliminate adversaries completely, therefore the development and use of hybrid systems may be more appropriate [81].

Another technique for developing AIDS is deep packet inspection which is used by AIDS to analyse network traffic payloads. Deep packet inspection analyses the payload component of a packet to identify what type of data is being transferred on a network. This is usually carried out in real time by applying “signatures" or regular expressions to the payload for well known protocols [114, 166]. The IDSs can be evaluated in the field using open source tools. Many of the most common tools can be found in the Kali Linux [125] which is specifically designed for Cyber Security testing e.g. nmap for network reconnaissance via port scanning. Port scanning is used for malicious purpose to search for potentially vulnerable machine on a network. The nmap tool can be used to carry out several types of scans such as ping scans and can be tailored through parameters to run in different modes/speed to try and circumvent detection. These tests should all be carried out as part of a penetration test. Generally, IDS must be complimented by other intrusion prevention technologies such as firewalls, vulnerability assessment, and an organisation security policy, to have a more secure system.

2.3.3 Metrics for Evaluating IDS

The core challenge to IDSs is to detect anomalous behavior and take actions before any adverse effects are caused to the network, information systems, or any other hardware and digital assets which forming or in the cyberspace. The performances of IDSs are measured in order to determine their positive detection rate (correctly detected attacks divided by total attacks attempted) and false alarm rate (falsely detected attacks divided by total traffics). The common widely used measurement metrics for evaluating IDSs in the research community include sensitivity (true positive rate), specificity (false negative rate), accuracy, precision (positive predictive value), recall (hit rate) and F-Score [166, 162]. These metrics are computed as follows:

$$Sensitivity = \frac{TP}{TP+FN} \tag{2.4}$$

$$Specificity = \frac{TN}{TN+FP} \tag{2.5}$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \tag{2.6}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.7}$$

$$Recall = \frac{TP}{TP+FN} \tag{2.8}$$

$$F-score = \frac{2*Recall*Precision}{Recall+Precision} \tag{2.9}$$

where TP, FP, TN, and FN refer respectively to: true positive, false positive, true negative and false negative.

High sensitivity means that the IDS is generating few false negatives and high trues positives, whereas, high specificity means the model is generating few false positives and high true negatives. Precision, recall and F-Score metrics are used to facilitate the understanding of how good is the IDS when there is an uneven class distribution in the dataset (i.e.; class imbalance). Note that, high accuracy indicate that the IDS is doing better only when the datasets are symmetric (i.e.; false negatives and false positives samples are almost balanced). F-score is efficient than accuracy when the datasets have an uneven class distributions. High precision is a confident indicator that the model is producing low false positives. High recall indicates that, the model is able to predict the positive samples with low false negatives.

2.4 Artificial Immune Systems

Artificial Immune System (AIS) is the class of computational intelligence systems inspired by the HIS, which is designed to solve engineering problems related to anomaly detection, classifications and optimisations. The AIS employs mathematical and computational techniques to model the immune system behavior as a metaphor to anomaly-based NIDS. Since 1990s, AIS researches on intrusion detection system have been carried out by different researchers [34, 35].

2.4.1 Human Immune System

The HIS is made up of tissue (e.g, skin and lung), cells (e.g, blood cells), and organs (e.g, heart and liver) that are networked together to protect and defend the body against foreign invaders that are trying harm it (e.g. bacteria and viruses) [112, 22]. When foreign invaders are encountered for the first time by the HIS, it retains their memory to be able to identify them when encountered again in the later stage [150]. Natural HIS's cells have the capabilities to recognise the presence of infectious or harmful foreign substances and take action by eliminating them or generating immune tolerance. Thus, the main purpose of the HIS is to act as the body's own army.

Self-Non-self discrimination theory of HIS

The self-non-self theory articulates that, HIS is able to discriminate between what is foreign and potentially harmful called non-self, and its own cells called self. Self-non-self theory is realised through two natural processes known as negative selection and the positive selection. During the negative selection process, T-cells that react against self-cells are eliminated; thus, only those that do not bind to self-cells are allowed to survive and mature in the adaptive immune system (thymus). After negative selection process is done, the matured T-cells are then circulated throughout the body to protect the body against foreign antigens. Thus, the negative selection process is used to make the HIS sensitive to foreign substances while simultaneously providing tolerance and adaptation for self cells. Note that, the positive selection process works as the opposite of the negative selection process [22].

An inspiration from the negative selection and positive selection processes were used to develop numerous AIS algorithm such as Negative selection algorithm (NSA) [63] and Positive selection algorithm (PSA) [35] mainly used for classification tasks. The NSA is derived from the fact that all new born immature T-cells in HIS must undergo a process of negative selection in the thymus where the self-reactive T-cells binding with self-proteins are eliminated. Therefore the mature T-cells are released to the blood circle can only bind to nonself antigens. In AIS, negative selection algorithm collects a set of self string that define the normal state of the monitored system and then generated a set of detectors that only recognize nonself strings [63, 35]. This detector set is used to monitor the anomaly changes of the data in the system in order to classify them as being self or non-self. Positive selection algorithm is an alternative to negative selection in which the detectors for self strings are evolved rather than for non-self.

Despite of the advantages of error tolerance, adaptation and self-monitoring [94], the self-nonself AIS algorithm have been criticised and found to have weaknesses such as scalability, require initial learning phase, high false positives, etc [2]. To overcome these limitations, a new family of AIS algorithm based on the danger theory and characteristics behaviors of DCs was introduce [73, 2] as detailed in the next subsection 2.4.2.

Danger Theory of HIS

Danger theory (DT) articulates that, HIS relies not only on making a discrimination between self-cells and foreign cells but rather react to what might cause damage and things that might not [112]. The recognition of danger is based on the type of antigen detected. Antigen (e.g. virus) is a foreign molecule that is capable of causing HIS to generate immune response (i.e.; tolerance or elimination). Danger signals are produced when DCs are exposed to new environment associated with distress signals. The following three signals are used by HIS to discriminate between normal and abnormal antigens.

- **PAMP** are abnormal proteins produced by viruses and bacteria which can easily activate immune response.
- **DS** are released from the disrupted or stressed cells in the tissue which indicates an anomalous situation but with lower confidence than PAMP.
- **SS** are produced by normal cell death process in the tissue, which is an indicator of normal cell behavior.

Biological Dendritic cells

In biology, the DCs coordinate antigens (e.g. virus) presentation from the external tissues (e.g. skin and lung) and immune system [13]. They produce co-stimulatory molecules (csm) on their cell surface which limit the time they spend sampling the antigens in the tissue. DCs play a crucial role during initiation and regulation of immune response. Usually, DCs exist in one of the following three states depending on the concentrations of SS, PAMP or DS signals in the tissue as also illustrated in Fig. 2.4.

1. **Immature DCs (iDCs):** are found in tissues in their pure state where they still collect antigens (i.e.; normal proteins or something foreign). The concentration of the signals of the collected antigens causes iDC to move to a full-mature or semi-mature state.
2. **Mature DCs (mDCs):** iDCs are transformed to mDCs when they are exposed to a greater quantity of either PAMP or DS than SS which causes immune reaction.
3. **Semi-mature DCs (smDCs):** iDCS are transformed to smDCs when they expose to more SS than PAMP and DS which causes immune tolerance.

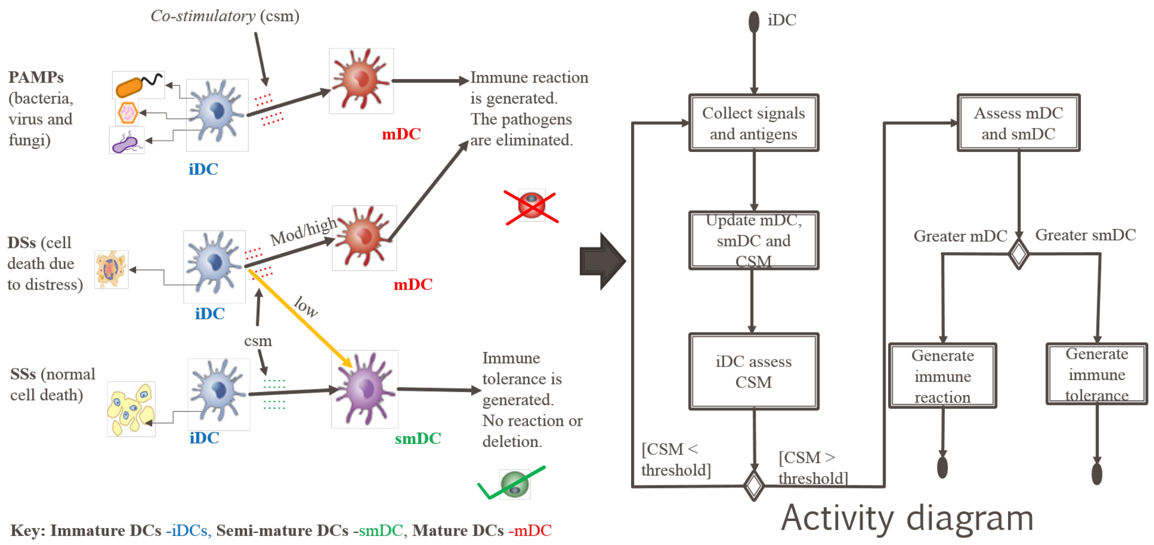

Co-stimulatory (csm)
PAMPs (bacteria, virus and fungi)
iDC
mDC
Immune reaction is generated. The pathogens are eliminated.
DSs (cell death due to distress)
Mod/high
low
csm
iDC
mDC
SSs (normal cell death)
iDC
smDC
Immune tolerance is generated. No reaction or deletion.
Key: Immature DCs -iDCs, Semi-mature DCs -smDC, Mature DCs -mDC
iDC
Collect signals and antigens
Update mDC, smDC and CSM
iDC assess CSM
[CSM < threshold]
[CSM > threshold]
Assess mDC and smDC
Greater mDC
Greater smDC
Generate immune reaction
Generate immune tolerance
Activity diagram

Figure 2.4 Signals sampling by natural DCs

2.4.2 Dendritic Cell Algorithm

Inspired by the biological danger theory and functioning of natural DCs, DCA is a classification algorithm developed for the purpose of anomaly detection in computer networks [73]. The DCA is a population based system where a population of artificial DCs is created to form a pool from which a number of DCs are selected to perform data items sampling, signals categorisation and classification as described in Algorithm 1. The DCs in the pool are exposed to current signal values and the corresponding data items included in the data source. Each DC has an ability to sample multiple data items. During classification, an aggregated sampling value from different DCs for a particular data item is computed which is used to classify a data item as normal or anomalous. Note that, the DCA goes through five stages in its life cycle to perform anomaly detection or classification tasks as describes below.

Feature Selection and Signal Categorisation

The inspiration from the DT and the behaviour of DCs leads to the development of the DCA which is a population based binary classification system. The data pre-processing is performed to select the most important features from the input training dataset. Then, the selected features are categorised into *PAMP*, *DS* and *SS*. The features which have higher values in the anomalous class compared to the normal class in the dataset are categorised as *PAMP*, confidently indicating the sign of abnormality [70]. The features which indicate low variation in both normal and anomalous classes are categorised as *DS*, indicating abnormality

Algorithm 1 DCA

```
input: the dataset $D$, the DC pool size $n$, sampling rate $s$, migration threshold $\theta$, anomaly-threshold $th$
output: Anomaly or Normal for data items;
/** Pre-processing & Initialization phase**/
create development DC pool $P_d$ with $n$ DC cells;
create migrated DC pool $P_m$ with unlimited size;
signal categorisation;
/** Detection phase**/
for each $d$ in $D$ do
    calculate the concentrations of $c_{csm}$, $c_{mDC}$ and $c_{smDC}$;
    calculate the cumulative values of $CSM$, $smDC$, and $mDC$ from their concentrations;
    for 1 to $s$ do
        randomly select a $DC$ from $P_d$;
        associate $d$ with $DC$;
        if cumulative $CSM > \theta$ then
            migrate DC to $P_m$ ;
            create new DC;
        end if
    end for
end for
/*Context Assessment phase */
for each $DC$ in $P_m$ do
    if cumulative $smDC \leq$ cumulative $mDC$ then
        DC-context=1;
    else
        DC-context=0;
    end if
end for
/* Classification phase */
for each $d$ do
    if DC-context == 1 then
        mature++;
    end if
end for
for each $d$ do
    MCAV = mature $\div$ total-presentation-by-DCs;
    if MCAV $> th$ then
        anomalous;
    end if
end for
return Normal or Anomaly for data item.
```

when their values increase in either class. The features which have higher values in normal class than those in anomalous class are categorised as *SS*, indicating an increase which is associated with the legitimate traffic's behavior.

The assignment of features to the three categories have been investigated in the literature. For instance, expert knowledge on the problem domain was used to select the most interesting features and map them into their appropriate signal categories in the work of [73]. Also, any other feature selection approaches such as fuzzy-rough feature selection [27, 26] can be used. Once the features are categorised, the value of each attribute from each signal category is then normalized within a range [0,1], usually using the min-max linear normalization technique. From this, the average of the normalised multiple attribute values in each signal category is taken as the signal concentration value [73, 29].

DC Initialisation and Sampling and Migration

A population of iDCS is initialised in a sampling pool and the DCA moves to the sampling stage. A population of 100 iDCs has been commonly used [73], but more investigation is required to determine the optimal size. Every iDC in the pool randomly samples data items. A *csm* migration threshold is applied to each iDC to limit the amount of data instances it can sample during the cycle, and the iDC becomes either mDC or smDC once once threshold is reached. In DCA, the concentration of *csm* is calculated as:

$$c_{csm} = \frac{(w_{P,csm} * x_P) + (w_{SS,csm} * x_{SS}) + (w_{DS,csm} * x_{DS})}{w_{P,csm} + w_{SS,csm} + w_{DS,csm}}, \tag{2.10}$$

where $w_{P,csm}$, $w_{SS,csm}$,$w_{DS,csm}$, represent the *PAMP*, *SS* and *DS* weights regarding the concentration of the *csm* value; x_P, x_{SS}, and x_{DS} are the *PAMP*, *SS*, and *DS* signal values respectively. The weights are pre-defined in the original version, but it can also be determined using other approaches proposed in this PhD thesis, such as generic algorithms. The migration threshold is often determined based on the characteristic of the dataset and the amount of DISTINCT data items the iDCs can sample.

Context Detection

A number of migrated DCs, often 10, are used for context detection to obtain the context values by calculating the concentration of *mDC* or *smDC*, using:

$$c_{mDC} = \frac{(w_{P,mDC} * x_P) + (w_{SS,mDC} * x_{SS}) + (w_{DS,mDC} * x_{DS})}{w_{P,mDC} + w_{SS,mDC} + w_{DS,mDC}}, \tag{2.11}$$

where $w_{P,mDC}$, $w_{SS,mDC}$,$w_{DS,mDC}$, represent the *PAMP*, *SS* and *DS* weights regarding the concentration of the *mDC* value; x_P, x_{SS}, and x_{DS} are the *PAMP*, *SS*, and *DS* signal values respectively. The concentration of *smDC* can be calculated using the same way, but with a different set of weights. Note that, the weights, again, are assigned by experts in the original version, but this PhD thesis has suggested a number of other ways for weight calculation. Once the concentrations of *smDCs* or *mDCs* are calculated, the context values are assigned to each sampled data items for context assessment; in the meantime, the DCs are reset to iDCs and returned to the sampling pool to maintain the population size.

Context Assessment

The cumulative values of *smDC* and *mDC* obtained from the context detection phase are used to perform context assessment. If the data items collected by a DC has a greater *mDC* value than its *smDC* value, the context is assigned as abnormality (or one particular class for other binary classification tasks); otherwise normality (or the other class). This information is then used in the classification phase to compute the number of anomalous data items present in the data set i.e.; those with a binary value of 1 are potentially anomalous.

Label Assignment

All collected data items are analysed by their Mature Context Antigen Values (*MCAV*). The *MCAV* value is used to assess the degree of anomaly (or the degree of belonging one particular class for other binary classification tasks) of a given data item, and based on this the label of a data item can be assigned. In DCA, the *MCAV* value is calculated by dividing the number of times a data item is presented in the *mDC* context to the total number of presentation by in DCs (either smDCs or mDCs). Data items with their *MCAV*s greater than a pre-determined anomaly threshold are classified into the anomalous class (or one particular class for other binary classification tasks) whilst the others are classified into the normal one (or the other class). The anomaly threshold is derived from the training dataset by dividing the total number of anomaly class data instances and the total data instances in the training data set.

Indeed, the DCA has been successfully applied to a wide range of anomaly detection applications with significant performances, such as intrusion detection, robotic, fault detection in wind turbines, computer virus detection, e.t.c [29]. Due to the robustness of the DCA, this thesis aim to use it to develop an intrusion detection system for e-Government systems.

In this PhD thesis, an improved DCA system is developed in oder to be used to detect intrusions in netwoorked systems in particular e-Government systems.

Further Developments of DCA

Since its invention in 2005, the DCA has been further developed for better performance and less expert-dependence, such as stochastic algorithm [74] and a deterministic version [72]. Notably fuzzy systems have been employed to support DCA in a variety of ways, initially to classify the input signals [26, 27], and later in the context assessment process [28]. Parameter optimisation is another way for performance enhancement, such as [103]. Recent investigations show the need for a dynamic migration threshold parameter and variable size populations [71]. Theoretical research helps understand the algorithmic dynamics [75], assisted by the development of a deterministic variant [72]. Further theoretical analysis of the DCA is reported in [124], which analysed the DCA as a set of linear classifiers, without analysing the impact of the data item stream.

2.5 Summary

This chapter has presented a literature review of e-Government systems, blockchain technology, intrusion detection systems and artificial immune systems. The first part of the chapter mainly reviewed the overview of e-Government systems as well as privacy and security issues in e-Government. Briefly, e-Government uses ICTs to deliver public services to individuals and organisations effectively, efficiently and transparently. E-Government is amongst the systems that stores sensitive information about citizens, businesses and other affiliates, and therefore becomes the target of cyber attackers. The existing e-Government systems have been identified to be faced with the potential privacy issues, security vulnerabilities and suffer from a single point of failure due to centralised databases and servers.

The second part of the chapter reviewed the privacy and security technologies as well as common cybersecurity threats to information. Firstly, insider and external cybersecurity and threats were reviewed. Secondly, public-key cryptography, symmetric-key cryptography, digital signature and cryptographic hash functions which are essential components for the implementation of the blockchain technology are discussed. Then, the blockchain technology was reviewed in details. There is a number of consensus algorithms used in blockchain technology including PoW, PoS, DPoS, Byzantine Agreement, PoD, and etc. Typically, the blockchain technology can be either public (permissionless), private (permissioned) or consortium (semi-public and semi-private).

The third part of the chapter reviewed the intrusion detection systems. IDS is defined as a software that is used to detect anomalies and attacks in a given computer network. With respect to the detection technique, IDSs are categorised into two main groups, namely MIDS and AIDS. The critical limitation of MIDS is the over dependence on up-to-date signatures, so cannot detect zero day attacks. AIDS can be able to detect novel patterns such as zero day attacks.

The fourth part of the chapter has reviewed the artificial immune systems. Computer networks resemble HIS as it is easy to associate cyber-attacks with foreign molecules (pathogens) and the computer network to the mammalian body. The AIS employs mathematical and computational techniques to model the HIS behavior as a metaphor to a NIDS. This part has reviewed different types of AIS algorithms including the NSA, PSA and DCA. Particularly, inspired by the HIS, DCA algorithm is a classification algorithm developed for the purpose of anomaly detection in computer networks based on the danger theory and the functioning of natural DCs. The DCA algorithm has been identified to be more suitable for developing modern NIDS due to its beneficial properties such as error tolerance, adaptation and self-monitoring.

Chapter 3

The E-Government Framework

As discussed in chapter 1 and 2, the existing e-Government systems, such as e-Government websites and eIDs management systems, are centralised where one or duplicated central servers and databases store and provide information to users. The centralised management and validation system is likely to suffer from a single point of failure and makes the system a target to cyber attacks such as DDoS, DoS, malware, and etc. Any e-Government system will remain vulnerable to privacy and security breaches if there are not any better security technology and countermeasures designed to combat these threats in the future.

Therefore, this chapter proposes a prototype of a decentralised e-Government framework with privacy preservation, and insider and external threat detection functionality, using blockchain and an artificial immune system. Blockchain technology has recently appeared to be one of the core technologies for secure data sharing and storage over trustless and decentralised systems. It enables the implementation of highly secure and privacy-preserving decentralised applications and systems where information is not under the control of any centralised host or third parties. Existing data and new transactions are stored in linked blocks (i.e. ledgers) distributed across the network in a verifiable and immutable way. Information security and privacy are enhanced on the way in which transactions are encrypted and distributed to all nodes. Note that, data in blockchain is append-only and immutable; in other words, once a piece of information is added to the chain, it cannot be deleted or altered in the future [7]. Adding unwanted transactions to the blockchain network is a critical concern due to its immutability nature [156]. Unwanted traffics such as spyware, worms, ransomware, spam can be very economically costly and business-wise catastrophic [161]. Thus, unwanted traffics must be detected and prevented from getting into the proposed e-Government blockchain network. Note that, the main characteristics of the blockchain technology include transparency, interoperability, decentralisation, immutability, efficiency, better security and privacy, and faster settlement of transactions within the network. Whereas, the mains goal of e-Government is to increase greater transparency and convenience, higher revenue and efficiency, better interoperability and effectiveness, and less corruption and operational overhead. Thus, the exploitation of the blockchain technology in e-Government systems will help to easily realise its aforementioned goals.

This chapter therefore also proposes an IDS based on AIS for identifying and mitigating unwanted insider and external traffics in the proposed e-Government framework. One particular implementation of AIS, i.e. DCA algorithm, has been successfully applied for anomaly detection in computer networks with competitive performances [73, 29]. Therefore, DCA is also employed in this PhD project due to its intrusion detection capability in computer networks [73] and beneficial properties such as self-organisation, scalability, decentralised control, and adaptability [29].

The theoretical and qualitative analysis on security and privacy of the framework shows that, encryption, immutability and the decentralised management and control offered by the blockchain technology can provide the required security and privacy in e-Government systems. In addition, the DCA-based IDS can detect and mitigate attacks before getting into the e-Government system; and if these attacks are missed or manage to bypass the IDS, there is no centralised server that can be a direct target. The proposed framework can be applied in any Government organisations to implement secure e-Government systems to ensure consistency and completeness.

The rest of this chapter is structured as follows. Chapter 3.1 details the overview of the proposed e-Government framework and prototype. Chapter 3.2 describes the decentralised control in the proposed framework. Chapter 3.3 presents privacy-preservation mechanism in the proposed framework. Chapter 3.4 details cybersecurity provisioning in the proposed framework. Chapter 3.5 presents the discussion about security and privacy analysis of the proposed framework, and finally Chapter 3.6 summarises the chapter.

3.1 Overview of the E-Government Framework

This section presents a framework and prototype of a decentralised e-Government system, which can be adopted by any Government for the purpose of ensuring both security and privacy while simultaneously increasing trust in the public sector. The proposed e-Government framework consists of three modules as illustrated in Figure 3.1. Firstly, a decentralised e-Government module is comprised of a P2P network with each node representing a public department based on the blockchain technology. Secondly, an external attack detection module based on the DCA detects unexpected traffics coming from the Internet to the e-Government system for further investigation by the network administrator. Thirdly, an insider threat detection module based on the DCA identifies internal anomalies from legitimated accounts of the e-Government system for further investigation. The anomaly detection modules ensure the legitimacy of the transactions or blocks before they are appended to the blockchain, as the blockchain database is append-only and immutable. Note that, both the external attack detection module and the insider threat detection module will be implemented by using the DCA, which are jointly discussed in Section 3.4 below.

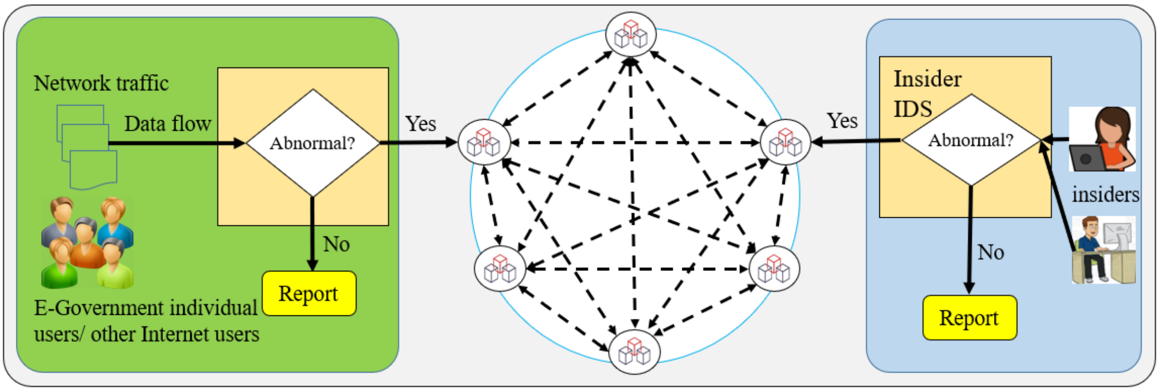


Figure 3.1 The decentralised secure and privacy-preserving e-Government framework

3.2 Decentralisation

The decentralised e-Government system is made of a P2P network of e-Government devices (nodes) and user's devices. Blockchain is used in the proposed e-Government framework to eliminate the centralised control of data, and protect sensitive information against unauthorised access. It is designed operated between government departments and users to address the privacy vulnerabilities in the e-Government system. To implement and test the distributed ledger in an e-government system using the blockchain technology, Ethereum platform can be adopted [167]. Ethereum is an open software platform based on blockchain technology with the tools to build decentralised applications for instance smart contracts. A smart contract protocol running on the Ethereum platform can be used to simulate real contracts such as tax and insurance payment, employment contracts, utility bill payment etc [98]. Its ability to facilitate contract negotiation, simplify contract terms, implement contract execution, and verify contract fulfillment state is a best alternative in an e-Government system to store citizens' sensitive records. It reduces third party costs in traditional transaction and guarantee the security and reliability. In a blockchain technology, a transaction represent an atomic change of the record state in a system. Each transaction is validated by all nodes and agreed by most of the nodes before it is added to the chain.

Briefly, any new e-Government device or individual device joining the system will be reviewed by the existing devices of the network and one of the peer will be selected to set up a network node and blockchain address of a new device. When a new user register with the system through his/her device or one of the Government department, (s)he is assigned with a user ID and blockchain wallet for collecting and storing his/her transaction. Using their IDs and the blockchain addresses, e-Government users can submit and access their records from anywhere and everywhere.

3.2.1 Types of Decentralised Nodes

There are two types of nodes in blockchain terminology, which are full nodes and light nodes [7]. A full node downloads a full copy of the blockchain when it joins the blockchain network, which allows it to fully validate transactions and blocks. A light node does not download a complete copy of the blockchain when it joins the network, but it downloads only the block headers to validate the authenticity of transactions. To be able to transfer their transactions to the network and receive notifications when transactions affect their blockchain wallets, light nodes usually refer to a copy of a trusted full node of the blockchain. Therefore, in the proposed framework, e-Government department nodes serve as full nodes while user's devices serve as light nodes, although any business node is allowed to download a complete copy of the blockchain. The p2p network connectivity in the proposed network can be provided by using wireless broadband, thanks to the fact that many countries across the world are trying to incorporate a city-wide wireless broadband networks across the city using Wi-Fi technology [165].

3.2.2 Consensus Algorithm

The validation of a transaction or ledger requires a consensus from the majority of the network nodes. Since all department nodes have the same record, they validate by ensuring the information matches their blocks for a particular blockchain address. There are several consensus algorithms available for the blockchain technology, including PoW, PoS, DPoS, and etc as discussed in Chapter 2.2. In PoW, miner nodes which want to add (mine) a new block to the blockchain network must first solve a difficult mathematical puzzle which requires great computational power. Unlike the PoW, with the PoS, a node which creates a new block is chosen deterministically depending on its stake (wealth) [178]. PoS saves energy that is required in PoW to solve mathematical puzzle, and only the wealthy of a node (validator) is required to validate the new transactions and blocks. The DPoS attempts to solve the consensus problem by using delegates [178]. DPoS uses a real time voting and reputation system to create a panel of limited trusted delegates who will witness and validate the blocks. The witnesses have the rights to create blocks and add them to the blockchain network, in addition to prohibit malicious nodes from participating in adding blocks. Principally, in PoS and DPoS, stakeholders of the network shares are not expected to deliberately make bad decisions for the network.

Therefore, the choice of the underlying blockchain technology for the prototype implementation mainly relies on the availability and efficiency of the consensus algorithm implementation. In the meantime, the computational energy required for validating transaction must be affordable while the security of the established network must be guaranteed. For instance, Ethereum platform version 2.0 [85] implements PoS and smart contract protocol used to simulate real contracts such as tax and insurance contracts, employment contracts, and land registry [163]. It provides an important alternative in e-Government to store citizens' sensitive records, due to its ability to facilitate contract negotiation, simplify contract terms, implement contract execution, and verify contract fulfillment state. Thus, in the proposed framework, the PoS will be adopted as the consensus algorithm due to its computational efficiency in adding transactions and sealing a block.

3.2.3 Registering New Nodes

The process of registering a new node to the proposed e-Government blockchain network is summarised in Algorithm 2. Any e-Government department can join the blockchain network by setting up a full network node while a user node can only set up a light client. Once a new node join the network, a functional node will generate its blockchain wallet and address containing public and private keys as shown in lines 7 and 8 in Algorithm 2. The private key is used by each node to sign and validate transactions therefore it must be stored safely (line 9). After generating the address, a node will contact delegates in the blockchain network to send its registration request where one of the delegates will verify its registration and transfer some e-Government tokens (registration record) in its blockchain address.

Thereafter, a new node is added to the network, and its registration is broadcasted to the network peers by the assigned delegate (line 12 to 14), allowing other network peers to receive its wallet information for sending transactions in the next cycle. Additionally, a new node receives instructions for a network node setup so that it can be elected as a miner to validate transactions in the next cycle. Subsequently, the new node sets up the network node according to the instruction provided. In particular, the instruction consists of the size of initial token, the blockchain address of a node, public and private keys for signing and validating transactions before adding a block. The process of adding a new node is completed when a full network node is successfully set up and broadcasted to the network by a validator. Note that, information security is enhanced through the way that data is encrypted and distributed across the network. Therefore, even if a malicious node is registered as a department node, it cannot alter the data as every participating peer in the network is able to detect any alteration and invalidate the change.

Algorithm 2 Adding a new node to the e-Government network

```
input: Node registration request,
    Nodes $N$ in the current blockchain network, tokens
output: A newly created node $d$ in the e-Government blockchain network
if (the request is from Government) then
    create a full node $d$;
else
    create a light node $d$;
end if
generate public and private keys through
    $(K_{pub}, K_{pr}) \leftarrow generatePublicandPrivateKeys(d)$;
create a blockchain address for $d$ through $Addr \leftarrow createBlockchainAddress(d)$;
create blockchain wallet for $d$ through $Walt \leftarrow createBlockchainWallet(d, K_{pub}, K_{pr})$;
store the keys and wallet for $d$ through $storeKeysandWallet(K_{pub}, K_{pr}, Wal)$;
add input tokens to $d$'s address through $Addr \leftarrow Addr +$tokens;
select a validator $\beta$ among $N$ nodes to distribute $d$'s registration details to other nodes
    through $\beta \leftarrow selectValidator(N)$;
for each $n \in \{N - \beta\}$ do
    distribute $d$'s details through $distributeRegistration(n, d)$;
end for
$d \leftarrow verifiedNewNode()$;
return $d$.
```

3.3 Privacy-preservation

The proposed framework protects privacy and integrity of information in e-Government systems by using encryption and validation mechanism offered by the blockchain technology.

3.3.1 User Registration and Authentication

Users make their registrations using their devices with the process summarised in Algorithm 3. As described in lines 2 and 3 in this algorithm, an ID for a user is issued and a new blockchain address will be generated for the user containing public and private keys, allowing the identification of the owner. A blockchain wallet for this new user is created (line 4) and broadcasted with other user's details to other network peers in lines 6 to 8, so that each node can store it in its blockchain address. The created blockchain wallet is also used to send and receive transactions related to user's account. User IDs and private keys are stored safely (line 5) in the wallet file or the database of the user's device. Users can conveniently view their records and the new transaction available in their blockchain addresses through the wallet interface.

Algorithm 3 Registering a new user

```
input: User registration request
       Nodes $N$ in the current network
output: A newly registered user $u$
 1: generate public and private keys for $u$ through $(K_{pub}, K_{pr}) = generateKeys(u)$;
 2: create user ID through $uID \leftarrow createUserID(u)$;
 3: create a blockchain address for $u$ through $Addr \leftarrow createUAddress(u, K_{pub}, K_{pr})$;
 4: create blockchain wallet for $u$ through $Walt \leftarrow createBlockchainWallet(u, K_{pub}, K_{pr})$;
 5: store the ID and keys for $u$ through $storeIDandKeys(uID, K_{pub}, K_{pr})$;
 6: for each $n \in N$ do
 7:     distribute $u$'s registration details to all nodes through $distributeDetails(n, u)$;
 8: end for
 9: $u \leftarrow verifiedNewUser()$;
10: return $u$.
```

When a user submits a record to the e-Government network, the transaction is authenticated and initialised. From this, the block is updated to a new version which is broadcasted across the network for validation and then transferred to his/her blockchain address in all network peers. The transferred record is stored in the blockchain address of the user with the following data content: (1) the ID of the user, (2) the record value such as property registration, and (3) the record identification such as tax registration number. Each data instance in a blockchain represents an asset.

When a third party organisation (e.g., business) requests to access to a user's information for any official issues, the user needs to provide his/her blockchain address for verification. The organisation can then use the blockchain web API to access the blockchain data stored in user's address. All e-Government users are required to backup their private keys and keep them safe. If any user lost his/her private key, (s)he will be required to create a new blockchain address and request one of the e-Government department node to transfer his/her information from the old blockchain address to a newly created blockchain address.

Generally, when a registered user wants to access the network, his/her device and identity will be validated and authenticated. This helps to minimise human error which has always been considered as a cause of failure and a weak link to access information stored in information systems [110]. As a results, Government information will flow securely and seemingly to the right individuals at right time, right place and everywhere. Human errors in cyber security include sending sensitive data to the wrong recipient and exposing login credentials such as username and passwords. According to [59], human error remain a greater cause of cyber security breach in the public and private organisations.

3.3.2 Creating and Validating Blocks

Every block will be created by one selected e-Government node (validator) who is selected at random by the majority of peers from the list of active nodes. If the validator misses a block, another node will be tasked to create and validate the block to join the blockchain network. In PoS, a fixed period of time is set for block creation, often five seconds [178]. Algorithm 4 highlights the fundamental steps involved in the process of adding a block to the blockchain using the PoS consensus algorithm.

A block is added in a regular interval of time, T_c. Within this interval, the block undergoes the following phases of activities. First, an empty set of transactions R is initialised. Then, one node from the e-Government network is selected to create and validate transactions to the blockchain. Secondly, all the transactions are sent to the selected node. This process continues until the adding node stops accepting any new transaction for the block. Thirdly, the validator assembles the new block and sends it to the network delegates for review and verification. This allows nodes that selected the validator to digitally sign the block to prove its correctness. The signed block is returned to the validator and added to its local blockchain while simultaneously distributing the new block to the network. The validator cannot mine its own transaction and hence in lines 4 and 9, β (validator) is excluded from the set of nodes N. At the end of the algorithm, the blockchain is distributed to all Government nodes in the network.

Algorithm 4 Creating and adding a new block to the blockchain

```
input: A set of $N$ nodes in the current network,
    A blockchain $B$ including blocks from $b_0$ to $b_n$,
    Consensus time $T_c$ required to create a new block
output: A newly created block $b_{n+1}$
Initialise an empty set of transaction $R = \{\}$;
select a validator $\beta$ among $N$ nodes through $\beta \leftarrow validatorSelection(N)$;
while transaction_time $< T_c$ do
    for each $n \in \{N - \beta\}$ do
        $R \leftarrow R + GetTransactionsfromNode(n)$;
    end for
end while
$b_{n+1} \leftarrow createBlock(b_n, R)$;
for each $n \in \{N - \beta\}$ do
    $signBlock(b_{n+1}, n)$;
end for
$B' \leftarrow B + b_{n+1}$;
for each $n \in N$ do
    $distributeBlockchain(B', n)$;
end for
return $b_{n+1}$.
```

3.4 Cybersecurity

The IDS is used to function as an entrance to the decentralised e-Government P2P system, to address any security issues due to malicious traffics. When a user submits a transaction to the e-Government network, the IDS validates the traffics, and the validated traffic is forwarded to a validating department node to create a record and update the rest of the network with the newly created block using blockchain technology. All un-validated traffic will be dropped. The IDS is usually implemented as a trained classifier that can automatically discriminate malicious from normal traffics. Blockchain is not developed to discriminate between anomaly and normal traffics, but the intrusion detection can help detect anomalies during blockchain transactions.

IDSs are developed purposely to identity intruders who attempt to access network resources without authority or permission. Networked systems such as e-Government resemble natural immune system as it is easy to associate cyber-attacks with harmful bacteria (foreign) and the computer network to the human body. Over centuries, HIS have appeared to be strong and robust in protecting the human body against foreign molecules such as virus and bacteria by employing its adaptability, lightweight and autonomy. AIS algorithms are developed by mimicking the characteristics of HIS. DCA is one implementation of the AIS algorithms which is developed by mimicking the immune danger theory and the functioning of biological DCs.

Due to the robustness and effectiveness of the DCA, this section details how an IDS based on the DCA will be developed for detecting unwanted insider and external traffics to support the proposed e-Government framework, because the blockchain database is append-only and immutable. Of course, separate datasets are require for developing the modules, one set for external attacks detection and one set for insider threat detection. The IDS effectively prevent unwanted transactions such as virus, malware or spyware from being added to the blockchain-based e-Government network. The existing conventional DCA will be significantly enhanced in three ways before being applied to the framework. Firstly, a new parameters optimisation approach for the DCA will be implemented by using the GA or any other optimisation methods since the original DCA uses manual method to pre-defined the weighted parameters. Secondly, fuzzy inference systems approach will be used to developed an approach which can solve nonlinear relationship that exist between features during the pre-processing stage of the DCA so as to further enhance its intrusion detection performance in e-Government systems. Thirdly, a multiclass DCA capable of detection multiple attacks will be developed, given that the original DCA is a binary classifier and many practical classification problems including computer network intrusion detection datasets are often associated with multiple classes.

DCA-based IDS Parameters Optimisation

During the context detection phase, the DCA requires users to manually pre-define the parameters used by its weighted function to process the signals and data items. Manual derivation of parameters of the DCA has received criticisms as it cannot guarantee that the optimal set of weights is obtained [29, 28]. Therefore, the main goal is to develop a weight optimisation approach for the original DCA by using optimisation methods such GA in order to enhance its detection performance in the proposed e-Government framework. The optimised DCA-based IDS will be validated and evaluated using publicly available datasets such as CERT (for insider threat study) [67], and the KDD99 [92] and the UNSW_NB15 [117] (for external threat study).

Enhanced DCA-based IDS by Using Fuzzy Inference Systems

A TSK-based fuzzy inference systems approach for aggregating the features from the dataset, either linearly or non-linearly, to generate DCA input signals will be developed so as to further improve the DCA-based IDS detection performance in the proposed framework. In order to implement the proposed TSK approach, a data-driven rule base generation method will be employed to generate three sub-TSK fuzzy rule bases, corresponding to the three input signals of the DCA. Then, the TSK+ fuzzy inference approach will be applied to compute the value of each input signal from the assigned features for each data instance, before the application of the DCA. TSK-DCA will be validated and evaluated by using the same cybersecurity datasets as the optimised DCA.

Multiclass DCA for Multiple attacks Detection

A generalisation of the conventional DCA to facilitate multiple attacks detection in the proposed e-Government framework will be developed. This will be achieved by transforming the concept of 'normal and abnormal contexts' to support multiple classes or 'contexts', which in the original DCA were designed to support binary classification by simply following the biological Danger model. Thus, softmax regression [69] will be employed in the context analysis phase to generate a probability distribution of each output context in the DCA for class label assignment. The performance of the multi-class DCA in the proposed framework will be evaluated on different benchmark multi-class classification IDS datasets from UCI machine learning repository [40] and then apply it to cybersecurity datasets to validate its applicability in e-Government systems.

3.5 Discussion

This section provides a theoretical qualitative analysis on security and privacy performance of the proposed e-Government framework. Every e-Government system must guarantee the confidentiality, integrity and availability. Confidentiality is achieved when information is not disclosed to unauthorised users; integrity is achieved by protecting information from any sort of modification, while availability means information is available when needed and is free from DoS or DDoS. For computational efficiency, user's devices will run lightweight client to store the transactions rather than the complete copy of the blockchain which is expensive in terms of storage. E-Government devices are expected to be computationally powerful with enough storage capacity to store and process user's records efficiently.

The records stored in the proposed system are secured through the public key cryptography that protects against adversarial attempts to alteration, whilst network users are assigned with private keys for validating and signing transactions. Encryption and digital signature used in the network ensure security, privacy and access control to the stored records. Moreover, most of the blockchain consensus algorithms including the PoS require an attacker to control at least 51% of the network peers in order to alter every instance of a record [146], which is generally impossible to achieve. More precisely, to change any block in the blockchain, an attacker is required to modify every copy of that block in the network and then convince all nodes that the new block is the valid one. Also, to increase the privacy of the data stored in the proposed e-Government P2P network, all user's blocks are hashed and an unreadable hashes of the transactions are stored in the blockchain.

The proposed P2P e-Government system within the framework is decentralised where user's data are stored in different nodes which guarantees the availability of the system by avoiding any single point of failure. Using consensus mechanism such as PoS, it is difficult for any adversary to launch DDoS or DoS attacks against the system as registration is required for a node to start sharing information with the rest of the network peers. Any transactions received from the network node are validated by a selected peer making it difficult for malicious nodes to initiate malicious connections. Additionally, the DCA-based IDSs can identify and mitigate unwanted traffics in e-Government system before gaining access to the stored information. The security services and corresponding common measures provided by the proposed framework is summarised in Table 3.1, which ensure adequate privacy and security of the transactions.

Table 3.1 Security services and common measures

Security service	**Countermeasure (s)**
Authentication	Blockchain address and digital signature.
Access Control	IDS, Digital Signature and encryption.
Confidentiality	Encryption.
Integrity	Encryption and digital signature.
Non-Repudiation	Encryption and digital signature.
Availability	Distributed/decentralised and IDS.
Trust	Decentralised, encryption and digital signature.

Apart from decentralisation, security and privacy-preserving, the potential design goals of the proposed e-Government framework are summarised in Table 3.2.

Table 3.2 The design goals of the proposed e-Government Framework

Design Goal	**Justification**
Reduced human error	User devices and identities are authenticated in advance before gaining access to the blockchain network.
Greater scalability	System can easily scale out as it allows new devices and users to be added to the network automatically following the consensus mechanism.
Better intrusion detection	DCA-based IDSs can effectively detect insider and external threats in the proposed e-Government framework to avoid any invalid operations.
Increased public trust	Individuals have direct control of their information and all network participants are authenticated.
Improved reliability	Data is stored in multiple place. Consensus protocol ensure that data can only be altered when all participants agree.
Increased resiliency	There is no single point of failure and hence resilience to malware, DoS and DDoS attacks.
Improved auditability	It is easy to trace back the history of all transactions since they remain unchanged in the network.
Greater verifiability	All new transactions are validated by all participating peers in the network before being added to the blockchain.
Information ownership	Individuals are responsible to authorise who will access their information.
Improved accessibility	Information is stored at multiple locations which enhances easy and speed access.
Increased data quality	All transactions and records stored in the system are validated in advance making the stored information authentic and with the required quality.
Greater transparency	All nodes in the network share the same copy of the blockchain and through consensus mechanism they agree on the new transactions to be added.
Low operational costs	There is no third party organisation needed to process transactions.
Improved efficiency	Anyone in the network has access to all records and new records are distributed to all participating nodes.

3.6 Summary

This chapter reported a decentralised e-Government framework with privacy preservation, and insider and external threat detection functionality, using blockchain technology and the DCA algorithm. The proposed e-Government framework is comprised of three main modules. Firstly, a decentralised e-Government module is comprised of a P2P network with each node representing a public department based on the blockchain technology. Secondly, an external attack detection module based on the DCA detects unexpected traffics coming from the Internet to the e-Government system for further investigation by the network administrator. Thirdly, an insider threat detection module based on the DCA identifies internal anomalies from legitimated accounts of the e-Government system for further investigation. The theoretical and qualitative analysis on security and privacy of the proposed framework shows that, encryption, immutability and the decentralised management and control offered by the blockchain technology can provide the required security and privacy in e-Government systems. Insider and external threats associated with the blockchain transactions from users are detected and reported by the DCA-based IDS to avoid any invalid operations to the blockchain database. Thus, it can be applied in Government organisations to implement a decentralised and secure e-Government systems to overcome design challenges such as interoperability, integration and complexity. Additionally, this framework has the potential to increase citizens' trust in the public sectors. As this chapter is limited in framework and theoretical discussion level, the practical evaluation of this framework is further conducted in Chapter 4 to 6 of this thesis.

Chapter 4

Consortium Blockchain for E-Government Decentralisation and Privacy-Preservation

As discussed in the previous chapters, the existing e-Government systems are centralised and thus subject to single point of failure. Blockchain technology provides the decentralised environment for secure transaction processing in trustless systems. It is designed as an immutable and distributed database for protecting privacy and security of the shared transactions among its trustless participants. Blockchain technology can be public (permissionless), private (permissioned), hybrid or consortium (semi-public and semi-private). One type of blockchain technology that is designed to meet the need of the enterprise is the consortium blockchain. The consortium blockchain technology is a semi-public and decentralised blockchain system which consists of a group of pre-selected entities or organisations responsible for consensus and decisions making for the benefit of the whole network of peers [52, 109, 37].

Therefore, this chapter proposes a decentralised and privacy-preserving e-Government system based on the consortium blockchain technology. The consortium blockchain is particularly chosen in this project because it has moderate computational cost which is crucial for e-Government systems. The consortium blockchain allows decentralised and flexible information access control, where the accessibility of the information stored in the consortium network can be limited to validators (e-Government departments), authorised users (registered citizens and shareholders), or not limited at all (public information). Also, unlike public blockchain where consensus process and transaction audit are carried out by all nodes with high computational cost, consortium blockchain performs the consensus process using pre-selected trusted nodes with moderate cost.

The proposed decentralised system was simulated and evaluated by using Ethereum Visualisations of Interactive, Blockchain, Extended Simulations (eVIBES simulator) [36]. eVIBES simulator was selected because it is open source and supports off-chain (sideDB) data storage such as images, PDFs, DOCs, contracts, and etc; and these are essential part of e-Government systems. The performance evaluation based on the number of transactions processed per second and on the time for processing a single transaction by varying the number of nodes (validators) in the consortium blockchain network have proved that, the proposed system is suitable for security and privacy assurance in e-Government systems. Consortium blockchain technology provides the decentralised environment and control required in the proposed e-Government system.

The rest of this chapter is structured as follows. Chapter 4.1 briefly presents the background about the consortium blockchain technology. Chapter 4.2 details the proposed consortium blockchain-based e-Government system. Chapter 4.3 presents the simulation of the proposed e-Government system on eVIBES simulator. Chapter 4.4 evaluates the performance of the proposed system, and finally, Chapter 4.5 briefly summarises the chapter.

4.1 Background of Consortium Blockchain

The consortium blockchain is defined as semi-public and semi-private blockchain since it offers the benefits of both private and public blockchains [109]. Its notable difference from either public or private is seen at the consensus level. Instead of an open network where any node can validate transactions like in public or a centralised network where only a single organisation validate transactions like in private, a consortium blockchain uses the pre-selected nodes (i.e.; validators) to validate the transaction and users. The visibility of the information store in the consortium network is limited to validators, authorised users, or by all (publicly). More precisely, the pre-selected nodes (organisations) dictate who can see and write to the consortium blockchain ledger. The pre-selected organisation of the consortium network are distributed and each one maintains a copy of the blockchain on their machines.

Generally, a fewer number of node on the consortium network allows it to perform much more effectively compared to public blockchains while simultaneously removing a single entity control compared to private blockchain. Consortium blockchain is recommendable and the best choice when a system requires a collaborative environment to be formed among the participating organisations like in e-Government systems [37, 109, 52]. It is highly secure because only authorised users can get access into the information stored in the consortium network. In fact, the consortium blockchain technology has been successfully exploited for decentralised controls and privacy-preservation in supply chain [126], healthcare system [177], IoT [109], energy trading [64], land registry [149], smart homes [109], and so on.

4.2 Consortium Blockchain for E-Government

The proposed consortium blockchain-based decentralised e-Government system is illustrated in Figure 4.1. It consists of e-Government users and a P2P consortium blockchain network of e-Government devices (pre-selected validators) responsible for validating transactions and authenticating e-Government users joining the network. The P2P consortium network is also responsible for provision of various services to users including consensus, P2P communication, and users' identity management. In order to add a new data to the blockchain, the devices that jointly make up the e-Government blockchain system vote for a validator who will validate the transactions and seal the block to the database. This approach is appropriate for security requirement, since random nodes cannot join the network, generate new tokens and set up a network node unless the rest of e-Government nodes approve it. The consortium blockchain system make sure that the stored records are trustworthy, auditable and transparent. Note that, any new e-Government devices or user's devices joining the consortium blockchain network, will be reviewed by peers of the network and assigned with a blockchain wallet and identity for setting up its network node. The assigned blockchain wallet is for user to collect and access his/her data. In this system, e-Government users can use various devices and technologies such as laptop, tablet, smartphones and etc, to access e-Government services online as well as to store their credentials locally.

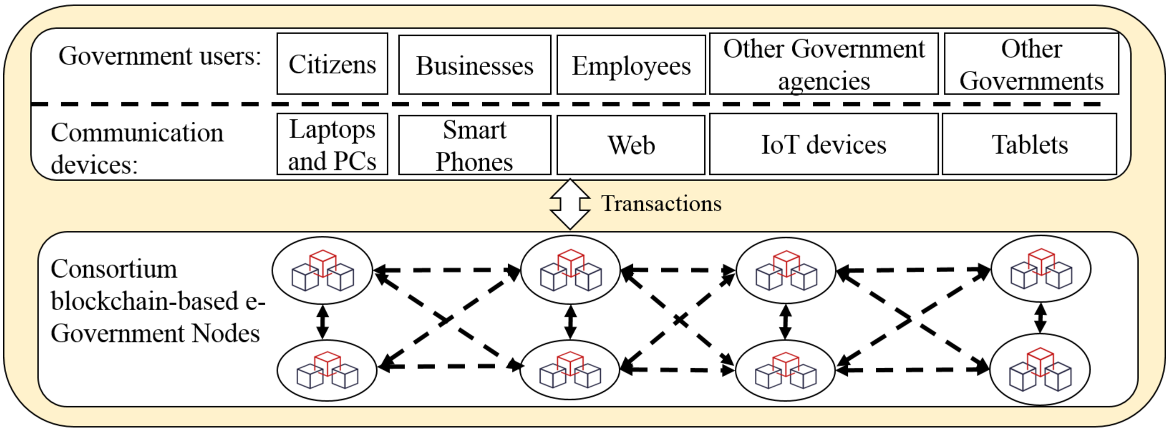


Figure 4.1 The proposed consortium blockchain e-Government system

Blockchain Structure

Two types of blockchain nodes are used in the proposed e-Government framework, including full nodes and light nodes. Full nodes store a copy of the entire blockchain; so, all e-Government devices from different departments are configured as full nodes and form the foundation of the blockchain network. The general users of the e-Government are registered and configured as light nodes. Light nodes do not store a copy of the entire blockchain. Instead, they use their account and wallet to connect to a full node for a copy of the blockchain upon demand. Therefore, e-Government users are required to register with one of the full nodes for authorisation and information access. Any new transaction for the user is relayed to one of the full nodes which is then propagated to other nodes in the network. Hence, full nodes are responsible for synchronising its local blockchain copy with the rest of the P2P network. Note that, a blockchain wallet is a digital storage which lets individuals to manage and store their account credentials including ID, passwords, private and public keys, and other information associated with their accounts. The wallet has a unique ID, similar to a bank account number, such that it can allow users and organisations to safely and securely transfer and exchange information between themselves. The wallet is stored in mobile and web applications, and accessible by using mobile phones and computers.

A number of validators among the e-Government consortium devices are required to validate any transaction before being added to the blockchain ledger. For instance, the consensus could be reached if a transaction is verified by twenty (20) departments in a consortium blockchain with fifty (50) departments. Therefore, a number of devices from selected e-Government departments are assigned as validators to authenticate transactions submitted to the network. The pre-selection is implemented by authorised entity from an e-Government agency, such as the Governance board, via the use of an approved application programming interface (API) based on a set of pre-defined operating rules. Other non-pre-selected e-Government devices are allowed to create, review and submit new transactions to the blockchain, but not contributing to the consensus and validation processes. A Government official can access the information to identify a particular user or agency from the blockchain. Note that, the proposed IDS discussed in the next Chapters 5 and 6 is deployed on all full nodes to detect intrusive activities in the consortium network.

A Government node (device) from a selected department that meets the requirement to be used as a full node is added to the consortium blockchain network using Algorithm 5. Once this new node joins the network, its blockchain wallet and address containing public and private keys are generated as indicated by lines 2 and 3 in Algorithm 5. The generated keys and wallet are for each node to sign and validate transactions therefore they must be stored safely using a function $storeKeysandWallet(K_{pub}, K_{pr}, Wal)$ in line 5. Thereafter, the registration information of the new node is broadcasted to the network peers by the selected node (line 6 to 8), allowing other network peers to receive its wallet information for sending transactions in the next cycle. Here, N in line 6 represents the total number of current registered devices in the network, and n is an individual device that is receiving the broadcasted information about the new node. Thus, the process of adding a new node is completed when a full network node is successfully set up and broadcasted to the network. After registration, the added node will receive a blockchain address, a blockchain wallet, and the private and public key pair for signing up and verifying transactions.

Algorithm 5 New device full node registration

```
input:  A new validating device $m$,
        A set of $N$ devices in the current consortium network
output: Registered device $m$
 1: generate public and private keys through $(K_{pub}, K_{pr}) \leftarrow keysGeneration(m)$;
 2: generate peer ID for $m$ through $pID \leftarrow peerIDGeneration(m)$;
 3: create a blockchain address for $m$ through $Baddr \leftarrow generateAddress(m, K_{pub}, K_{pr})$;
 4: create blockchain wallet for $m$ through $wallet \leftarrow generateWallet(m, K_{pub}, K_{pr})$;
 5: store the keys and wallet for $m$ through $storeKeysandWallet(K_{pub}, K_{pr}, Wal)$;
 6: for each $n \in \{N - m\}$ do
 7:     distribute $m$'s registration details to all nodes through $broadcastRegistration(n, m)$;
 8: end for
 9: $m \leftarrow verifiedValidator()$;
10: return $m$.
```

The efficiency of the proposed e-Government system is guaranteed by fewer full nodes thanks to the use of a consortium blockchain rather than a private blockchain, which ensures timely process of transactions. Additionally, e-Government devices in the blockchain network are designed to be able to store off-chain (sideDB) data such as images, PDF, DOC, text documents, contracts, and other non-transactional data files that are too large to be stored in the blockchain or subject to change or deletion in the future [52]. These files are encrypted from the user side and stored off-chain in the proposed decentralised consortium blockchain-based system. That is, the raw data of each file is hashed by using SHA256 algorithm to generate a hash value that is stored in the blockchain with a link to the original file stored in the off-chain database [118] (meaning that the information is not publicly accessible).

Off-chain transactions are processed faster than on-chain ones because they do not require the verification from all full nodes. Note that, when a transaction is performed on off-chain stored documents, the role of e-Government department is to verify the execution of the transaction and confirm that the blockchain protocols have been observed in the proposed system. For instance, when a user agrees the terms and conditions to share their off-chain documents with an external organisation, an e-Government blockchain full node will execute the transaction via the exchange of private keys. Personal data, such as identity number, social insurance number, names, nationality, are required when signing up a new user. In the proposed framework, personal data are hashed and cryptographically linked with the ledger, and thus the blockchain ledger only stores the hash of the personal data.

User Registration

Citizens, businesses and other users, can register to e-Government with the process summarised in Algorithm 6. An ID is firstly created for the user as described in line 2, and a new blockchain address is then generated for the user containing the public and private keys as shown in line 3, allowing the identification of the user. The generated private key and user ID are stored safely by the selected node (line 5) before his/her details being broadcasted to other e-Government network nodes. From this, a blockchain wallet for this new user is created and broadcasted with other details to all nodes in the blockchain network, as depicted from lines 6 to 8. The created blockchain wallet will be used to send and receive relevant transactions to this account. User IDs and private keys will be stored safely in the wallet file or the database of the user's device (line 5). The User can conveniently view their records and the new transactions available in their blockchain addresses through the wallet interface.

Algorithm 6 User Registration

```
input: New user request
output: Registered user $u$
 1: generate keys for $u$ through $(K_{pub}, K_{pr}) \leftarrow generateKeys(u)$;
 2: creating user ID for $u$ through $uID \leftarrow createUserID(u)$;
 3: create a blockchain address for $u$ through
    $Addr \leftarrow createBlockchainAddress(u, K_{pub}, K_{pr})$;
 4: create blockchain wallet for $u$ through $Walt \leftarrow createBlockchainWallet(u, K_{pub}, K_{pr})$;
 5: store the ID and private key for $u$ through $(uID, K_{pr}) \leftarrow storeIDandKeys(uID, K_{pr})$;
 6: for each $n \in N$ do
 7:     distribute $u$'s registration details through $distributeWallet(n, Walt)$;
 8: end for
 9: $u \leftarrow verifiedNewUser()$;
10: return $u$.
```

When a user submits a record to the e-Government network, the transaction is authenticated and initialised. From this, the block is updated to a new version which is broadcasted across the network for validation and then transferred to the user's blockchain address stored in the e-Government consortium network. The transferred record in the user's blockchain address includes the following data content: (1) the ID of the user, (2) the record value or the transaction, such as property registration, and (3) the record identification, such as tax registration number. Each data instance in a blockchain represents an asset.

When a third party organisation (e.g. business) requests to access a user's information for any official issues, the user needs to provide their blockchain address for verification. The organisation can then use the blockchain web API to access the blockchain data stored in user's address. All e-Government users are required to backup their private keys and keep them safe. If any user lost their private key, the user will be required to create a new blockchain address and request one of the e-Government department node to transfer the user's information/records from the old blockchain address to the newly created blockchain address.

The identity of a registered user through a registered device will be validated and authenticated when the user wants to access the blockchain network. Note that human error remains a great cause of cyber security breach in public and private organisations [59]; this access validation and authentication helps to significantly reduce human error which has always been considered as a cause of failure and a weak link to access information stored in information systems [110]. As a result, government information will flow securely and seemingly to the right individuals at the right time through a user's device, regardless of the user's location as long as the Internet is available.

Data Storage

E-Government devices in the blockchain network are designed to be able to store off-chain (sideDB) data such as images, PDFs, DOCs, contracts, and other files that are too large to be stored in the blockchain or that are required to be deleted or changed in the future. These types of files make it necessary to identify and mitigate intrusive activities originating from insiders before being added. Thus, User's data is stored and replicated in e-Government database servers, which are supported by the consortium network devices. Storage and replication of this data is so critical in case data reuse is required as the data that are store in the blockchain network are immutable and difficult to be reused. In blockchain technology, the off-chain storage of data is required for documentation and verification [24]. A hash value for the off-chain document will be produced and stored in the consortium blockchain ledger while the real document is stored in the ledger storage layer. Additionally, the data that are processed by the participants of the consortium blockchain will be replicated and stored in this layer in order to guarantee more storage spaces for the blockchain network devices.

Devices Networking

Connectivity is required between users, e-Government consortium devices and storage servers, especially in real world scenarios. There is a wide choice of technologies that can be used on this layer such as WiFi, Ethernet LAN or cellular network. The participants of the consortium blockchain can use these devices to communicate transactions to/from the users and store new blocks in the ledger storage servers. Wireless devices can benefit from wireless communication technologies while interacting with the e-Government consortium blockchain network; thanks to the fact that many Governments around the globe are trying to incorporate wireless broadband networks across different cities using Wi-Fi technology [165].

4.3 Simulation of E-Government System

It is not practical to implement the proposed e-Government design in this research due to the significant hardware requirement, this study therefore has simulated the system using eVIBES simulator [36]. Briefly, the eVIBES is a configurable and open source system for simulating large-scale Ethereum networks so as to observe the empirical behaviours and dynamic properties of P2P nodes [36]. The eVIBES was adopted due to its less deployment efforts and low running cost compared to a real deployment of Ethereum which is a complex and difficult task for this research [36]. The eVIBES is a broadcast-based, an event-driven, scalable, message-oriented and concurrent blockchain simulator. Its scalability enables the network to accommodate a large number of nodes, without degrading the simulation speed or effectiveness. Note that, in eVIBES, nodes operate the same way as in actual Ethereum applications [23, 36].

The eVIBES allows users to configure the blockchain parameters for a specified blockchain network and study the behaviour of blockchain systems. The configurable simulation parameters include: the minimum and maximum number of P2P nodes, the number of transactions, the rate of transaction generation, the smart contract mode to allow smart contracts to be uploaded for execution during simulation, and customised initialisation of the genesis block for a new simulation [36]. Briefly, a smart contract is a piece of computer code which contains a set of rules under which the participating parties agree upon and run on the top of

Ethereum blockchain. The outputs of the eVIBES simulator include the overall execution time, the total number of transactions processed, the number of transactions per second (throughput), the block propagation delay and a log of all transactions. All simulation input and output metrics are observed on the web-based interface. Tendermint and Ethermint Ethereum protocols that are used to develop decentralised app [58, 57] were adopted to implement the eVIBES simulator.

Tendermint is a protocol used by the Ethereum blockchain platform to handle networking and consensus between nodes in the blockchain network. It consists of a blockchain consensus engine and an application interface as its main technical components for developers. Tendermint uses an application interface known as Application Blockchain Interface (ABCI) to allow transactions to be processed in any programming language [58, 37]. With tendermint, developers need to focus only on the application layer of the blockchain application since it handles the consensus and the networking for the application. Tendermint uses the PoS consensus algorithm in its core engine for handling the consensus and networking between nodes [58]. Additionally, tendermint enables the blockchain developers to define how many pre-selected validators are required in the system. Note that, a blockchain application implemented via the tendermint protocol can provide high performance since tendermint engine has a block creation time of just 1 sec compared to that of bitcoin of 10 minutes [58].

In order to run the blockchain application on the web, Ethereum provides Ethermint protocol which is fully compatible with web interfaces [57, 37]. Ethermint implement web compatible API layer thus allows the blockchain application developed by using the smart contract to load via the URL and run on the web [57]. The execution of the smart contracts on Ethereum platform normally happens in the Ethereum Virtual Machine. Thus, Ethermint enables blockchain developers to deploy their smart contracts on the web and facilitate communication between Ethereum nodes and users on the network. Therefore, Ethermint plays a crucial role in smart contracts compatibility on the web.

Note that, in order to simulate a consortium blockchain as used in the proposed e-Government system, the default eVIBES (i.e. private blockchain) is reconfigured. The sidechains (i.e. off-chain database) provided by the eVIBES facilitate this reconfiguration simulation, and the same implementation of database and servers as reported in [36] is adopted in this work; and the smart contracts facilitate the exchange of users' information, documents, property, etc., between P2P nodes without the involvement of a trusted third party. Each node in the proposed e-Government system represents a peer in the simulated consortium network which continuously exchanges messages with other peers.

The architecture of the configured eVIBES simulator for the proposed e-Government system is illustrated in Figure 4.2. The main component of the simulator is the orchestrator, which manages the entire simulated network by sending configuration parameter settings and messages to all nodes in the system. It also monitors and records the system states, and communicates with the browser for system configuration and simulation results visualisation. In other words, the orchestrator regulates the entire simulated e-Government system including the generated blockchain. The HTTP module enables the interaction between the browser and the orchestrator in an event driven manner.

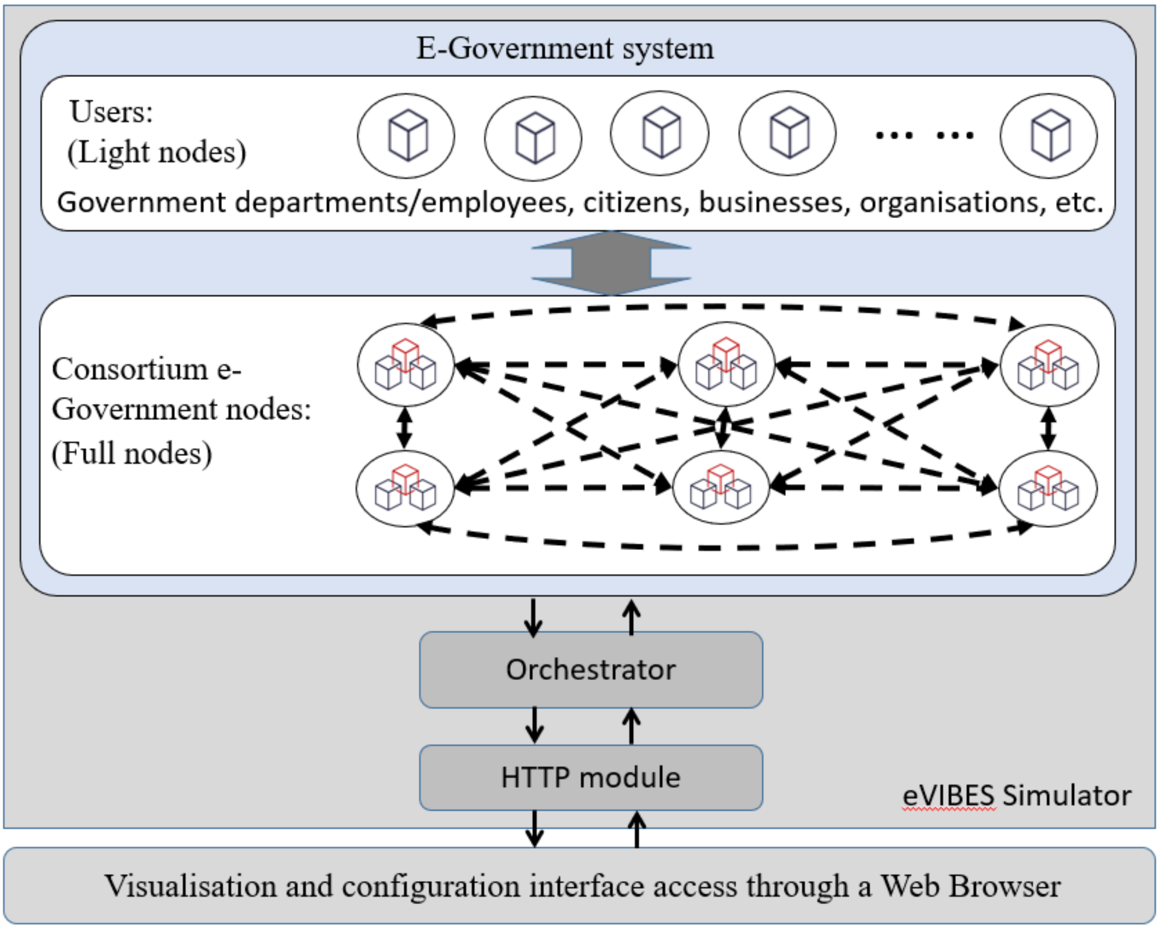

E-Government system
Users:
(Light nodes)
Government departments/employees, citizens, businesses, organisations, etc.
Consortium e-Government nodes:
(Full nodes)
Orchestrator
HTTP module
eVIBES Simulator
Visualisation and configuration interface access through a Web Browser

Figure 4.2 The simulation of the proposed e-Government design

4.4 System Evaluation

This section details the the performance evaluation and working of the proposed consortium blockchain-based e-Govrnment system through the eVIBES simulator. The experiment was performed by using an HP workstation with Intel$^{®}$ Xeon^{TM} E5-16030 v4 CPU @3.70 GHz and 32GB RAM. The main purpose of experiments was to evaluate the performance of the proposed system based on the number of transactions processed per second, block creation and propagation time, and on the time for processing a single transaction by varying the number of nodes in the consortium blockchain network to prove that the proposed system is scalable and suitable for security and privacy assurance in e-Government systems.

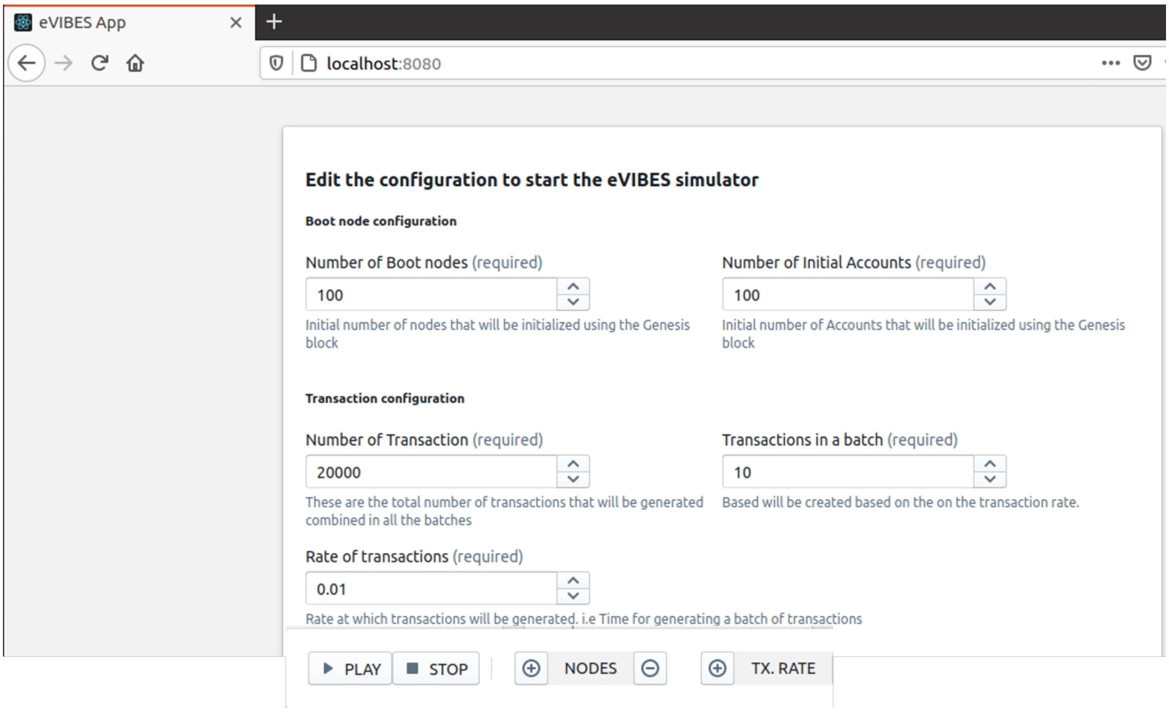
4.4.1 Experiment Design

This experiment is implemented through the eVIBES simulator, running in VMware with 8GB RAM and Ubuntu 20.04 Long Term Support operation system. The parameters of the e-Government blockchain network were configured as presented in Table 4.1. The maximum number of nodes that need to boot with the genesis block during initialisation in the network was set to 100, with flexibility to be decreased or increased during the simulation. Each initialised node is assigned with an initial account, including node ID, blockchain wallet, private and public keys for signing and validating transactions. The number of transactions that an individual node can process in a batch and add into the blockchain was set to 10, and the rate of transaction in the network was set to 0.01 sec for fast and efficient block creation and propagation in the network [36, 23, 24]. The initial number of user accounts for generating transactions in the blockchain network was set to 100 users. The total number of transactions that the blockchain network can store from all nodes was set to 20000 for this simulation. Similarly, the initial number of peers per node was set to 5 to allow peer-to-peer communication between themselves during transaction execution in the blockchain network. Sample user interface configuration for the simulation is illustrated by Figure 4.3.

Transactions are executed in form of smart contracts in the eVIBES simulator, and the smart contracts mode is set to be independent execution. Independent execution mechanism of smart contracts enables each node to maintain its own state of database while executing the assigned transactions before sharing and adding them to the blockchain [36]. Before the simulation starts, the default genesis block is overridden with the custom Ethereum genesis block values to maintain consistency and to serve as the starting point for its ledger in the consortium blockchain [36, 23]. All nodes periodically send their transactions to the consortium blockchain storage every 10 seconds, in order to capture the throughput, the rate at which transactions are validated, the block propagation time, and the processing time for each transaction.

Table 4.1 eVIBES parameters of the e-Government blockchain network simulation

Parameter	Setting
Number of boot nodes	100
Number of transactions in the network	20000
Rate (in sec) of transaction generation	0.01
Initial number of peers per node	5
Transactions in a batch	10
Initial number of user accounts	100
Smart contracts mode	Independent execution
Genesis block	Override



Edit the configuration to start the eVIBES simulator

Boot node configuration

Number of Boot nodes (required)
100
Initial number of nodes that will be initialized using the Genesis block

Number of Initial Accounts (required)
100
Initial number of Accounts that will be initialized using the Genesis block

Transaction configuration

Number of Transaction (required)
20000
These are the total number of transactions that will be generated combined in all the batches

Transactions in a batch (required)
10
Based will be created based on the on the transaction rate.

Rate of transactions (required)
0.01
Rate at which transactions will be generated. i.e Time for generating a batch of transactions

Figure 4.3 eVIBES user interface configuration for simulation

4.4.2 Experimental Results

This section presents the performance results based on block propagation time in the proposed consortium blockchain-based e-Government system, throughput in the system and a single transaction validation time.

Blocks Propagation Time

The purposes of this experiment is to demonstrates the scalability and robustness of the proposed system for adoption in e-Government systems particularly when there is a spike in the number of blocks/transactions in the network. In this experiment, the number of Ethereum nodes (validators) was increased linearly while simultaneously measuring the average propagation time of blocks in seconds with the result shown in Figure 4.4. Here, the number of Ethereum nodes represent the number of dedicated e-Government full nodes. As it can be noticed in Figure 4.4, the block propagation time increase linearly as the number of Ethereum nodes increase in the network, which demonstrates the scalability of the proposed e-Government system. Also, since as the number of nodes increase in the network so as the block propagation time, this further proves that, blocks generated by peers in the network are being successfully propagated to all nodes.

Transactions Throughput

The purposes of this experiment is to provide a detailed technical analysis about transaction throughput of the consortium blockchain network under different number of nodes for adoption in e-Government systems. The performance of the consortium blockchain network based on the number of transactions processed per second (throughput) with increasing number of Ethereum nodes (validators) in the network, is illustrated in Figure 4.5. Note that, the ideal case happens when all transactions from the users are validated in one second or less. It is clear from the figure that, the number of average transactions validated per second decreases as the number of Ethereum nodes increases in the consortium blockchain network. This is because the communication overhead required to pre-vote for a node will process the transactions, create a block, and append it to the ledger. Thus, if a given consortium blockchain system requires a large number of Ethereum nodes, the transaction processing speed will decrease. In e-Government systems, this may not be the case since all participating department save the same goal of proving public services, fewer validators will be enough and trusted to process and validate transactions while other Ethereum nodes communicating to users.

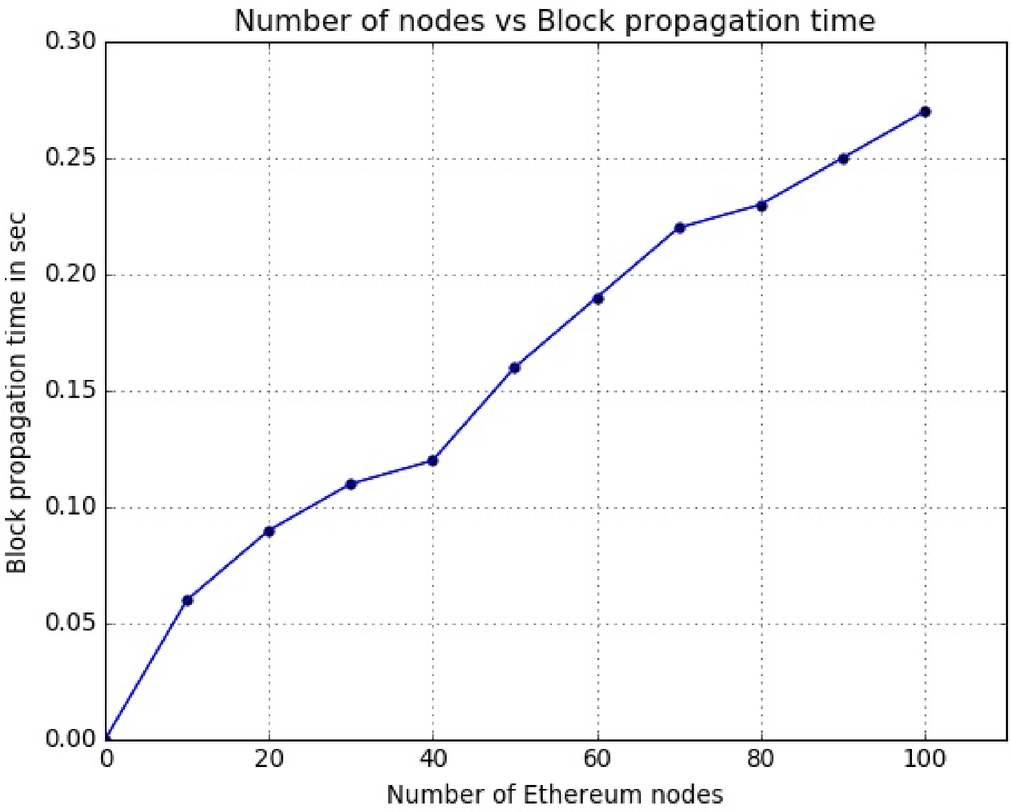


Figure 4.4 Block propagation time against the number of nodes

From Figure 4.5, a consortium network with 10, 20, 30 or 40 validators can validate 100 transactions in a few seconds, which is close to the ideal case. However, for a consortium network comprised of 50, 60, 70 or 80 validators, the network requires several seconds to validate 100 transactions, which deviates somehow far from the ideal situation. So, as the number of nodes increase in the network (e.g. up to 200 nodes), the network requires more time to validate transactions. Apparently, a compromise needs to be made between the desired network performance, the number of transactions sent per second, and the number of nodes in the consortium blockchain network, which should be fully considered during the design stage. More specifically, an extra care must be taken to select the appropriate number of validating Ethereum nodes in the consortium network for a balance between security and transaction throughput.

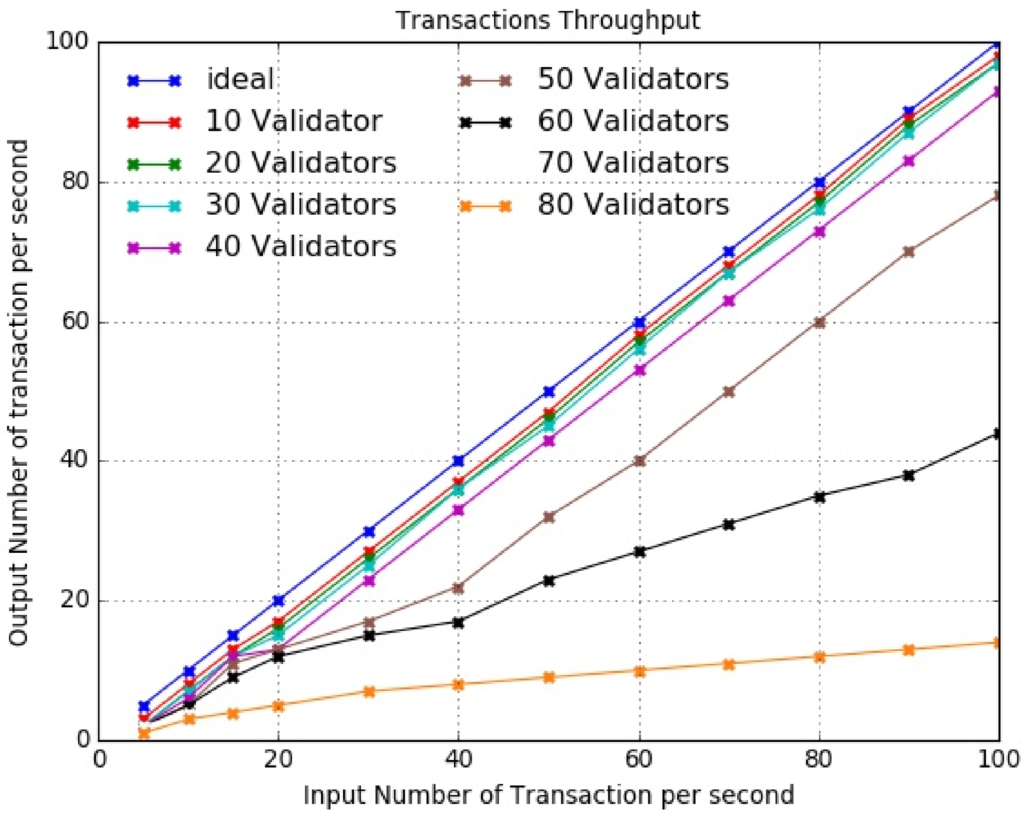


Figure 4.5 The transactions per second against Ethereum nodes

Single Transaction Validation Time

The purposes of this experiment is to provide a detailed technical analysis on a single transaction validation time (in seconds) in the consortium blockchain for adoption in e-Government systems. The standard average validation time per transaction in a consortium blockchain network comprised of 100 validators is usually between 0.1 sec to 1 sec [109, 38, 37]. It has been observed in this experiment that more time was required to validate a transaction as the number of Ethereum nodes was increased in the blockchain network, as shown in Figure 4.6. If a blockchain network contains less than 90 validators, the average validation time was less than 0.1 sec; however, the validation time is increased up to 0.12 sec when a network is comprised of more than 90 Ethereum nodes in the simulated environment. Therefore, during the design stage of a consortium blockchain, a compromise needs to be made between the desired network performance, the number of transactions sent per second, and the number of Ethereum nodes in the network. Thus, the proposed system can be adjusted depending on the requirements of a country where it will be applied.

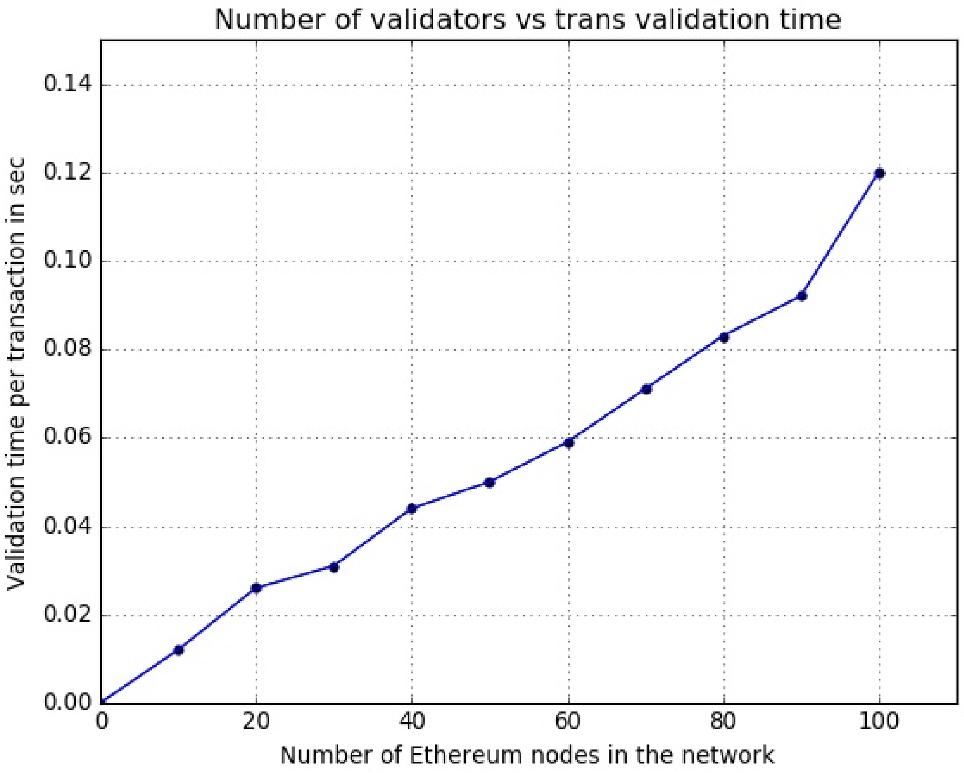


Figure 4.6 Validation time against the number of Ethereum nodes

4.4.3 Discussion

This section discusses how the proposed consortium blockchain-based system can prevent common threats to security and privacy in e-Government systems.

- **Prevention against DDoS attacks** – Usually occurs when attackers flood online services with massive fake traffic in order to render the service unavailable. Generally, DDoS attacks tend to consume a significant amount of bandwidth and resources until the service is down. In the proposed consortium blockchain-based system, no centralised server that can be a direct target for DDoS attacks. The decentralised nature of the proposed consortium blockchain will allocate data and bandwidth to the less overloaded peers in the network to absorb DDoS attacks when it happens.

- **Authentication and Authorisation attacks** – This may occur when the adversarial users try to take control of the consortium blockchain network so that the can authorise themselves or introduce fake nodes for authorising users while taking control of the network. This is impossible in the proposed system since all peers of the consortium network are pre-selected by an authorised entities from e-Government agency of a particular government beforehand.

- **Threats to Anonymity** – Anonymity in blockchain-based applications may open the door for criminals to carry out unauthorised activities [7]. In the proposed system, the blockchain information is only available to the consortium peers and authorised users. Therefore, any adversary trying to set up an anonymous connection will be detected instantly since the validators verify any user trying to access the information stored in the consortium network.

Note that, more time is required to validate transaction as the number of validators increases in the consortium blockchain network. Therefore, a compromise needs to be made between the desired network performance, the number of transactions sent per second, and the number of validating peers in a consortium blockchain network, which should be fully considered during the design stage. More specifically, an extra care must be taken to select the appropriate number of validators in the consortium network for a balance between security and transaction throughput.

The Advantages of the proposed e-Government consortium blockchain are summarised in Table 4.2.

Table 4.2 Advantages of the consortium blockchain-based e-Government network

Advantage	**Justification(s)**
Fast transaction speed	Only a selected group of participants process transactions.
High scalability	New participants can be added to a controlled number of consortium nodes when needed.
Low transaction costs	The transactions validators process transaction without incentives like in public blockchain.
Low energy consumption	A voting mechanism is used to pre-select the validating nodes, no computation is needed.
Low risk of cybersecurity attack	Random participants are not allowed to join the consortium unlike in public blockchain.
High transparency	Participants of the network know their peers within the consortium (more of enterprise).
High data integrity	Information is maintained in a consistent manner within the consortium network.
High collaboration	Different department and agencies will be able to share information on demand.

4.5 Summary

This chapter proposed a decentralised e-Government system based on the consortium blockchain technology. The consortium blockchain allows decentralised and flexible information access control, where the accessibility of the information stored in the consortium network can be limited to validators (e-Government departments), authorised users (registered citizens and shareholders), or not limited at all (public information). Also, unlike public blockchain where consensus process and transaction audit are carried out by all nodes with high computational cost, consortium blockchain performs the consensus process using pre-selected trusted nodes with moderate cost. The proposed decentralised system was simulated and evaluated by using VIBES simulator since it is open source and supports off-chain (sideDB) data storage such as images, PDFs, DOCs, contracts, and etc. The performance evaluation based on the number of transactions processed per second and on the time for processing a single transaction by varying the number of nodes (validators) in the consortium blockchain network have proved that, the proposed system is suitable for security and privacy assurance in e-Government systems. Consortium blockchain technology provides the decentralised environment and control required in the proposed e-Government system.

Chapter 5

E-Government Security and Privacy-Preserving Using Enhanced Dendritic Cell Algorithms

The most daunting and challenging task in the proposed e-Government system presented in this project as described in Chapter 3 and 4, is to distinguish between normal and malicious traffics effectively before granting them access to the blockchain ledger. Note that, blockchain database is append-only and immutable; so, once the information is added cannot be deleted or changed in the future. Adding unwanted transactions to the blockchain network can be deadly due to its immutability nature. Unwanted traffics such as spyware, worms, ransomware, spam and etc, must be detected and stopped from getting into the e-Government blockchain network. Hence, there is a need of developing an cybersecurity attack detection system that can detect insider and external threats in the proposed e-Government system.

Therefore, this chapter develops three different cybersecurity attacks detection systems based on enhanced DCA algorithm for identifying and mitigating unwanted traffics in e-Government systems. DCA is employed in this project due to its intrusion detection capability in computer networks and beneficial properties such as self-organisation, scalability, decentralised control, and adaptability [26, 75]. In order to enhance the attacks detection performance of the original DCA, this chapter has significantly revised and enhance it in three ways. Firstly, a new parameters optimisation approach for the DCA was implemented by using GA; since the original DCA uses manual method to pre-defined the weights for its objective function. Secondly, fuzzy inference systems approach was used to developed an approach which can solve nonlinear relationship that may exist between input features and the resultant three DCA's signals during its pre-processing stage. Thirdly, a new signal categorisation method for the DCA was proposed based on Partial Shuffle Mutation of GA to automatically categorise the input features into the three DCA's signal categories; given that the original DCA uses manual categorisation technique based on domain or expert knowledge of the domain. The enhanced DCA approaches will be used to function as an entrance to the decentralised e-Government system, to address any security issues due to malicious traffics in an e-Government system.

The performance evaluation of the proposed DCA approaches in e-Government systems were conducted by using three cybersecurity datasets namely the CERT (for insider threat study) [67], KDD99 (for external threat study) [92] and UNSW_NB15 (for external threat study) [117]. The experimental results show that the proposed DCA approaches are capable of detecting both external and insider threats with better performances compared to the original versions of the DCA and some state-of-the-art classifiers such as Decision Trees (DT), Naive Bayes (NB) Random Forests (RF), Artificial Neural Network (ANN) and Support Vector Machines (SVM).

The remainder of this chapter is structured as follows: Chapter 5.1 presents the proposed weights ptimised DCA. Chapter 5.2 presents the proposed fuzzy inference enhanced DCA. Chapter 5.3 presents the proposed GA-based signal categorisation for DCA. Chapter 5.4 demonstrates the experimental evaluation process of the proposed approaches as well as the discussion of the results, and finally Chapter 5.5 summarises this chapter.

5.1 Weights Optimised Dendritic Cell Algorithm

This section develops a DCA-based IDS by optimising its learning parameters for detecting malicious traffics in the proposed e-Government system. Conventionally, in its life cycle, the DCA goes through a number of phases including features categorisation into three artificial signals (i.e.; SS, DS and PAMP), context detection of data items, context assignment and finally labeling of data into either normal or anomaly class. During the context detection phase, the DCA requires user to manually pre-define the parameters used by its weighted function of Equation 5.1 to process the signals and data items. Manual derivation of the parameters of the DCA has been criticised due to the fact that it cannot guarantees that the optimal set of weights is obtained [53, 26, 29]. Therefore, in order to improve the detection performance of the DCA for e-Government systems, a new weight optimisation approach implemented by using the GA is developed.

$$Context[csm, smDC, mDC] = \sum_{d=1}^{m} \frac{\sum_{i,j=1,1}^{3}(c_j * w_i^j)}{\sum_{i,j=1,1}^{3} w_i^j}, \tag{5.1}$$

where $c_j(j = 1,2,3)$, represent the PAMP, DS and SS signal values respectively; and $w_i^j(i, j = 1,2,3)$ represent the weights of *csm*, *mDC* and *smDC* context, regarding PAMP, DS and SS, respectively.

Compared to other evolutionary search techniques, the GA was chosen in this PhD project due to the following reasons [176, 78, 68]

- Its ability to easily achieve the proper balance between exploitation and exploration of search space simply by setting well its parameters (i.e.; mutation, crossover, population size, fitness function, number of iterations and elitism rate).

- GA can easily escape from being trapped in local optimal solutions compared other evolutionary methods through crossover and mutation that help it to search for the optimal solution from a population of points while other optimisation methods normally search from a single point (can easily get trapped in local maxima).

- Compared to other evolutionary search methods, GA simply uses objective function without any need of its derivative or auxiliary information to generate the fitness score for its solutions which makes it simple and fast. More precisely, GA doesn't require any prior information about the structure of the objective function to be optimised.

- GA uses probabilistic selection process rather than deterministic ones which are exploited by other evolutionary optimisation algorithms which makes it more suitable for the search problems.

- GA parameters can be easily modified and adaptable to fit different computational problems compared to other evolutionary algorithms.

- GA is suitable for multi-objective optimisation problems compared to other methods since it has the ability to search through large and wide solution space. Therefore, in the future GA could be used to optimised the DCA algorithm by treating it as a multi-objective problem with weights to be optimised and signals to be categorised, concurrently.

5.1.1 The proposed Dendritic Cell Algorithm

The proposed optimised DCA-based IDS is illustrated in Figure 5.1. Feature selection step is firstly employed to select the most informative features from the input dataset. The selected features are then categorised into three groups of either *PAMP*, *SS* or *DS*, representing the three input signals. The optimal set of parameters associated with the three signals are effectively searched by employing the GA during context detection phase as detailed in Section 5.1.2. The last two phases of the optimised GA-DCA are exactly the same as the original DCA versions, and thus the rest of this section focuses only on the optimisation of the parameters using GA during the context detection phase.

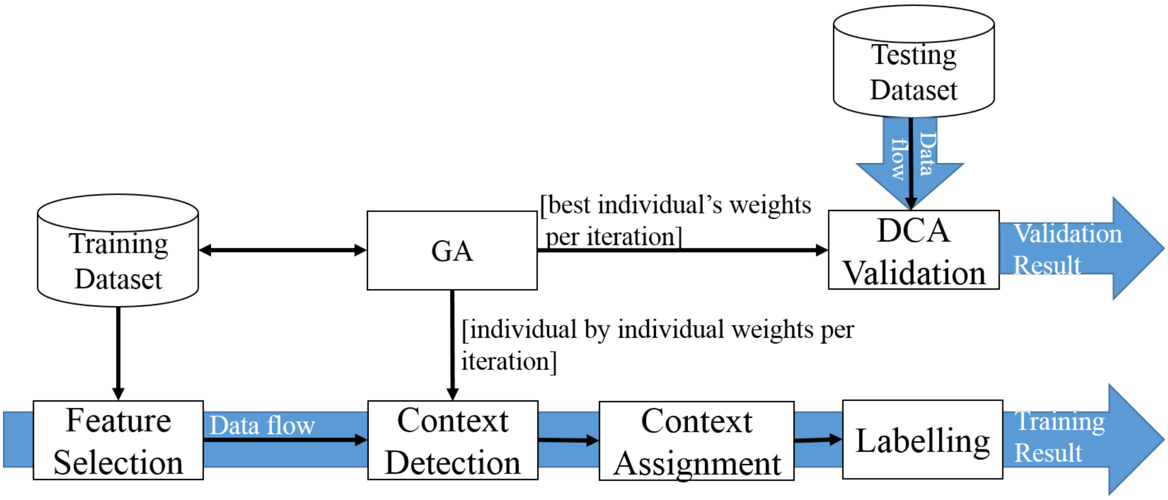


Figure 5.1 Intrusion detection in e-Government systems by using DCA

5.1.2 Parameters Optimisation Using Genetic Algorithm

GA is a family of computational models inspired by natural evolution which heuristically search for optimal or near-optimal solutions to optimisation problems [78]. GA has been employed for parameters optimisation in machine learning models, such as neural networks [68], rule base optimisation in fuzzy inference systems [107], and etc, with significant performances. GA starts by randomly initialising a population of individuals (solutions) in a pool for a given problem. Then, it uses the techniques inspired by evolutionary biology such as selection, mutation and crossover to evaluate each individual in every iteration. Gradually, more effective individuals are evolved over a number of iterations until a specified level of performance or maximum number of iterations is reached.

The weights required by the DCA's Equation 5.1 can be summarised as a matrix:

$$W = \begin{bmatrix} w_{smDC,1} & w_{smDC,2} & w_{smDC,3} \\ w_{mDC,1} & w_{mDC,2} & w_{mDC,3} \\ w_{csm,1} & w_{csm,2} & w_{csm,3}. \end{bmatrix}$$

In this section, the main steps followed by GA to optimise the DCA's parameters for the DCA's Equation 5.1 are summarised below.

1) Individual representation: In this chapter, an individual (I) within a population is designated as a possible solution that contains all the weights involved in Equations 5.1. Therefore, the individual is represented as $I = \{w^1_{smDC}, w^2_{smDC}, w^3_{smDC}, w^1_{mDC}, w^2_{mDC}, w^3_{mDC}, w^1_{csm}, w^2_{csm}, w^3_{csm}\}$, where 1,2,3 represent the three signals categories extracted from the selected features during the pre-processing step.

2) Initialisation of individual's parameters: The population $\mathbb{P} = \{I_1, I_2, ..., I_N\}$ is randomly initialised from a Gaussian distribution with a mean of 0 and a standard deviation of 5. Here, N is the size of population which is a problem-specific modifiable parameter, with 10-50 being widely used [107, 120].

3) Objective function: GA requires an objective function to determine the quality of individual's solutions within a population. In this chapter, the objective function is taken as the classification accuracy of the DCA, which is determined by:

$$Accuracy = \frac{\text{Total number of correctly classified data instances}}{\text{Total number of data instances}} \tag{5.2}$$

4) Selection: In this work, the fitness proportionate selection method is adopted for selecting individuals who reproduce. Generally, in fitness proportionate method, the probability of an individual to become a parent is proportional to its fitness. Thus, the fittest individuals have the higher chances of being selected for reproduction for the next iteration. Assuming that f_i is the fitness of an individual I_i in the current population $\mathbb{P}$, its probability of being selected to generate the next generation is:

$$p(I_i) = \frac{f_i}{\sum_{j=1}^{N} f_j}, \tag{5.3}$$

The fitness f_i of an individual I_i in the work is determined as follows [12]:

$$f_i = r_i^{\frac{1}{4}} \cdot 20 \tag{5.4}$$

where r_i is the ranking position of individual I_i in a ordered population $\mathbb{P}$.

5) Reproduction: Usually, GA explores the problem space through crossover and mutation. In this work, the single point crossover and mutation operations are applied with probability of α and μ respectively. Crossover and mutation help GA to reduce the likelihood of being trapped in local maxima and increase the likelihood of obtaining the global optimal solution. Crossover operation is applied to increase the exploitation of search space whereas mutation explores the search space so as to increase diversity in population. After these operations, the newly bred individuals and some of the best individual in the current generation $\mathbb{P}$ jointly form the next generation of the population $\mathbb{P}^{'}$.

6) Iteration and termination: In this chapter, GA converges when the classification performance of the DCA exceeds the pre-specified threshold of the optimum accuracy or the pre-defined maximum number of iterations is reached. Subsequently, when GA terminates, the optimal weights are taken from the fittest individual in the current population. From this, the DCA can perform validation using this set of weights and get applied in the proposed e-Government framework.

Note that, the GA is used during the training stage to generate and optimise the weights and therefore no large computational effort is imposed when applied to online in-time performance in e-Government system. Additionally, GA-DCA IDS does not add any dependency on the training dataset after getting the best weights, so the quality of the original DCA is not compromised.

5.2 Fuzzy Inference Enhanced Dendritic Cell Algorithm

Conventionally, the DCA takes three signals as inputs, including SS, DS and PAMP, which are generated in its pre-processing and initialisation phase. In particular, after a feature selection process for a given training data set, each selected attribute is assigned to one of the three input signals. Then, these input signals are calculated as the aggregation of their associated features, usually implemented by a simple average function followed by a normalisation process. If a nonlinear relationship exists between a signal and its corresponding selected attributes, the resulting signal using the average function may negatively affect the classification results of the DCA.

Therefore, this section proposes an approach named TSK-DCA to address such limitation by aggregating the assigned features of a signal linearly or non-linearly depending on their inherit relationship using the TSK+ fuzzy inference systems [107]. In order to implement the proposed TSK-DCA, a data-driven rule base generation method is firstly employed to generate three sub-TSK fuzzy rule bases, corresponding to the three input signals. Then, the TSK+ fuzzy inference approach is applied to compute the value of each input signal from the assigned features for each data instance, before the application of the DCA.

5.2.1 Fuzzy Inference Systems

Fuzzy inference systems are built upon fuzzy logic to map from the input space to the output space. They have been widely applied in solving either linear or non-linear problems of arbitrary complexity, such as [106, 172]. The two most widely used fuzzy inference systems are the Mamdani fuzzy model and TSK fuzzy model. Compared with the Mamdani fuzzy model, which is more intuitive and commonly utilised to deal with human natural language, the TSK fuzzy model is more convenient to be employed when crisp output values are required. Both of these conventional fuzzy inference systems are only workable with a dense rule base by which the entire input domain is fully covered. Fuzzy interpolation enhances the power of the conventional fuzzy inference systems by relaxing the requirement of dense rule bases [97, 96]. In other words, the conventional fuzzy inference systems fail to generate a conclusion when a given observation does not overlap with any rule antecedents in the rule base, but fuzzy interpolation can still approximate the conclusion. Various fuzzy interpolation methods have been developed in the literature, such as [82, 173, 175, 174, 119, 169, 108].

The original TSK inference system generates a crisp inference result as the weighted average of the sub-consequences with the firing strength of the fired rules as weights [147]. Obviously, no rule will be fired if a given input does not overlap with any rule antecedent. As a consequence, the TSK inference cannot be performed. TSK+ was proposed to address such issue which generates a consequence by considering all the rules in the rule base [105]. Suppose that a sparse TSK rule base is comprised of n rules:

$$
\begin{aligned}
R_1 : & \textbf{ IF } x_1 \text{ is } A_1^1 \text{ and } \cdots x_j \text{ is } A_j^1 \cdots \text{ and } x_m \text{ is } A_m^1 \\
& \textbf{ THEN } z = f_1(x_1, \cdots, x_m), \\
& \ldots \ldots \\
R_n : & \textbf{ IF } x_1 \text{ is } A_1^n \text{ and } \cdots x_j \text{ is } A_j^n \cdots \text{ and } x_m \text{ is } A_m^n \\
& \textbf{ THEN } z = f_n(x_1, \cdots, x_m),
\end{aligned}
\tag{5.5}
$$

where $A^i_j, (i \in \{1, 2, \cdots, n\}$ and $j \in \{1, 2, \cdots, m\})$ represents a normal and convex polygonal fuzzy set that can be denoted as $(a^i_{j1}, a^i_{j2}, \cdots, a^i_{jv})$, v is the number of odd points of the fuzzy set. Given an input $I = (A_1^*, A_2^*, \cdots, A_m^*)$ in the input domain, a crisp inference result can be generated by the following steps:

Step 1: Identify the matching degrees between the given input $(A_1^*, A_2^*, \cdots, A_m^*)$ and rule antecedents $(A_1^i,\ A_2^i,\ A_3^i, \cdots, A_m^i)$ for each rule R_i by:

$$S(A_j^i, A_j^*) = \left(1 - \frac{\sum\limits_{q=1}^{v} |a_{jq}^i - a_{jq}^*|}{v}\right) \cdot (DF) \ , \tag{5.6}$$

where DF is a $distance\ factor$, which is a function of the distance between the two concerned fuzzy sets:

$$DF = 1 - \frac{1}{1 + e^{(-cd+5)}} \ , \tag{5.7}$$

where c is a sensitivity factor, and d represents the Euclidean distance between the two fuzzy sets for a given defuzzification approach. in particular, c is a positive real number. Smaller value of c leads to a similarity degree which is more sensitive to the distance of two fuzzy sets, and vice versa.

Step 2: Determine the firing degree of each rule by aggregating the matching degrees between the given input and its antecedent terms by:

$$\alpha_i = S(A_1^*, A_1^i) \wedge S(A_2^*, A_2^i) \wedge \cdots \wedge S(A_m^*, A_m^i) \ , \tag{5.8}$$

where $\wedge$ is a t-norm operator usually implemented as a minimum operator.

Step 3: Generate the final output by integrating the sub-consequences from all rules by:

$$z = \sum_{i=1}^{n} \alpha_i \cdot f_n(x_1, \cdots, x_m) \ / \ \sum_{i=1}^{n} \alpha_i \ . \tag{5.9}$$

5.2.2 TSK+ Enhanced Fuzzy Inference System for DCA

The proposed TSK-DCA system is depicted in Figure 5.2. In particular, given a training dataset, a feature selection process is first performed to select the most significant features. The selected features are then categorised into three groups, representing the three input signals. From this, three TSK+ fuzzy models can be generated using the given training data set for the aggregation of the three input values. Given an input, the TSK+ inference systems take place to aggregate the given inputs to the three DCA input signals. Then, the output of the TSK-DCA classifier is generated by the DCA model. Each of these key components of the proposed system is detailed in the following subsections.

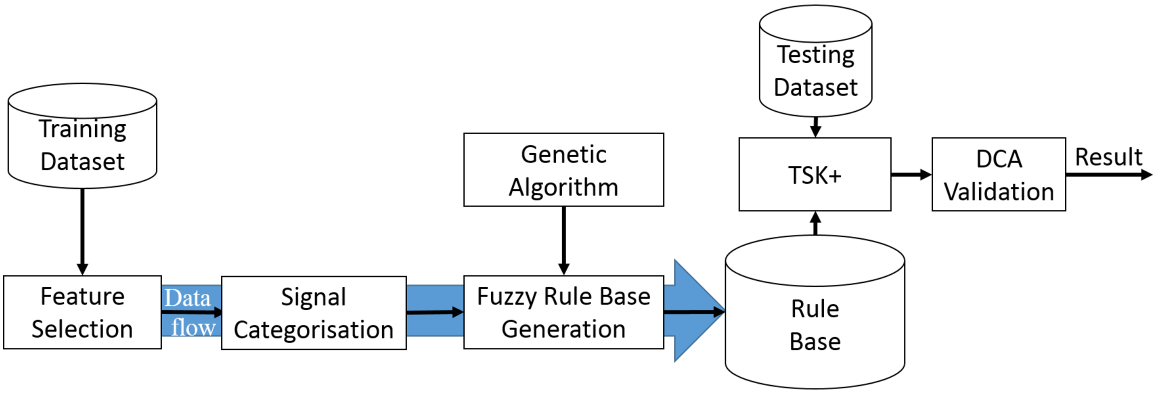


Figure 5.2 The overall enhanced TSK-DCA system

Signal Aggregation Using TSK+ Fuzzy Inference System

Feature selection and signal categorisation method was first applied to the dataset to select the most informative features and categorise them into the three signals of either SS, DS or PAMP. Then, the TSK+ approach was applied to aggregate the input signals of the DCA. In order to apply the TSK+ approach as introduced in Section 5.2.1, a rule base needs to be generated first, which is outlined in Figure 5.3 in two key steps as detailed below.

Clustering: The K-Means clustering algorithm is employed to each sub-dataset to obtain the rule clusters based on the Euclidean distance. Assume that k_s clusters are needed for sub-data set $T_s, (s \in \{1,2,3\})$. The algorithm starts with the initialisation of k_s cluster centroids, which may be generated randomly or based on some strategies. Then each data object is allocated to the cluster whose centroid is the closest from the data object. After that, the algorithm updates the cluster centroids and reassign each data object to its closest cluster again. This process is reiterated until the centroids stabilised or the sum of squared error (SSE) is minimised. In particular, SSE is defined as follows:

$$SSE = \sum_{j=1}^{n_i^s} \sum_{i=1}^{k^s} (\| x_{ij}^s - v_i^s \|)^2 \,, \tag{5.10}$$

where x_{ij}^s is the j^{th} data point in i^{th} cluster in the sub data set T_s; v_i^s is the centre of the i^{th} cluster in the sub dataset T_s; n_i^s is the number of data points in i^{th} cluster of the subset T_s; and $\| x_{ij}^s - v_i^s \|$ is the Euclidean distance between x_{ij}^s and v_i^s.

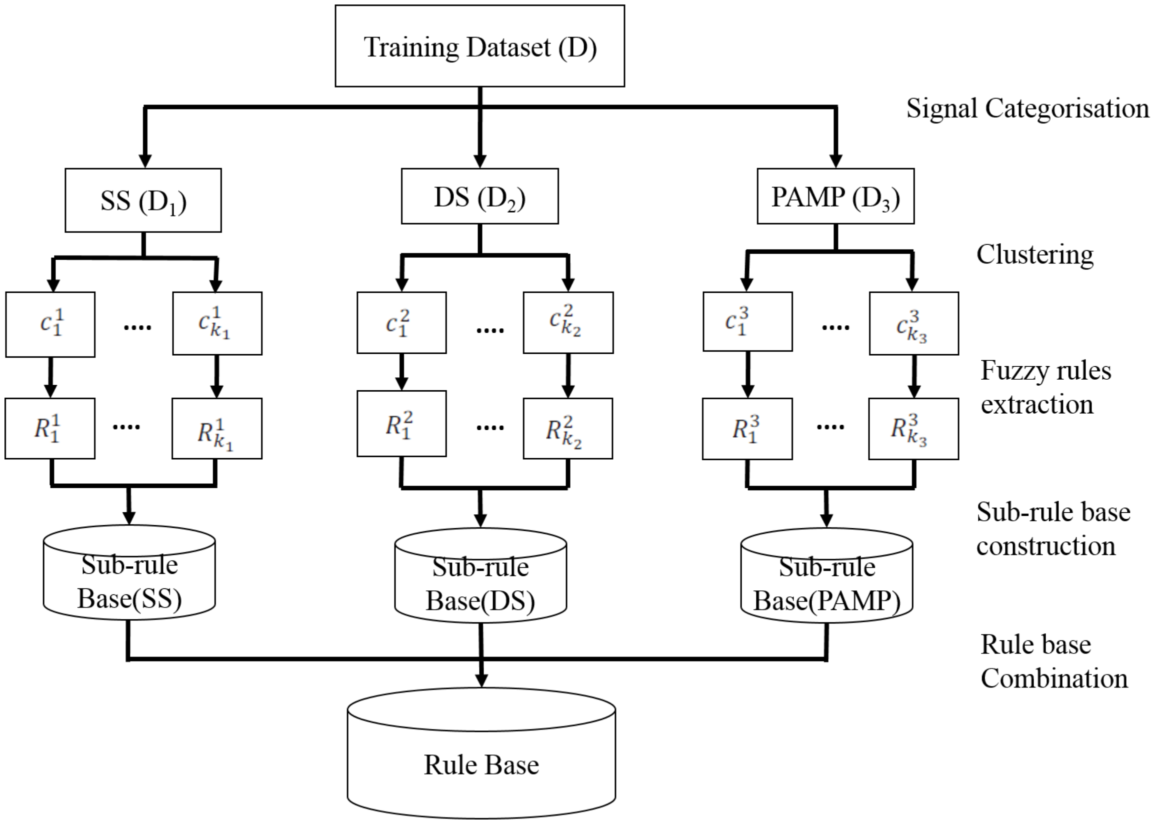


Figure 5.3 The TSK+ fuzzy rule base generation

Note that the value of k^s has to be pre-defined to enable the application of the K-Means algorithm. The Elbow method, which has been used in [107, 172], is employed in this work to determine the number of clusters.

Fuzzy Rule Extraction: Each obtained cluster is expressed as one TSK fuzzy rule. Assume that a determined cluster for a signal is associated with d features, then a TSK fuzzy rule R_i can be formed as:

$$\begin{aligned} R_i : \textbf{IF } x_1 \text{ is } A_1^i \text{ and ... and } x_d \text{ is } A_d^i \\ \textbf{THEN} y = f_i(x_1, ..., x_d) \ , \end{aligned} \tag{5.11}$$

where A_r^i $(r = \{1,...,d\})$ is a fuzzy set as a rule antecedent. For simplicity, triangular membership functions are utilised in this work, that is $A_r^i = (a_{r1}^i, a_{r2}^i, a_{r3}^i)$. Without loss of generality, take a rule cluster c_k as an example, which contains p_k elements, such as $c_k = \{x_k^1, x_k^2, ..., x_k^{p_k}\}$. The core of the fuzzy set is set as the cluster centre which is $a_{r2}^i = \sum_{q=1}^{p_k} x_k^{qr}/p_k$; and the support the fuzzy set is expressed as the span of the cluster, i.e. $(a_{r1}^i, a_{r3}^i) = (\min\{x_k^{1r}, x_k^{2r}, ..., x_k^{p_k r}\}, \max\{x_k^{1r}, x_k^{2r}, ..., x_k^{p_k r}\})$. The consequent of a TSK fuzzy rule is the DCA input signal values. In particular, the consequent is expressed as a first-order polynomial in this work, which can be represented as $y = f_i(x_1, ...x_d) = \beta_0^i + \beta_1^i x_1 + \beta_2^i x_2 + ... + \beta_d^i x_d$, where β_i^d is a constant parameter of the linear functions.

The rule base is optimised by employing the genetic algorithm (GA). The algorithm firstly initialises the population with random individuals. It then selects a number of individuals for reproduction by applying the genetic operators, that is mutation and crossover. The offspring and some of the selected existing individuals jointly form the next generation. The algorithm repeats this process until a satisfactory solution is generated or a maximum number of generations has been reached.

In this project, an individual (I) in a population ($\mathbb{P}$) is used to represent a potential solution that contains all the parameters of the polynomial functions in the TSK rule consequent, represented as $I = \{\beta_0^1, ..., \beta_d^1, ..., \beta_0^i, ..., \beta_d^i, \beta_0^n, ..., \beta_d^n\}$, where n denotes the total number of rules in the current rule base. Given a population, represented as $\mathbb{P} = \{I_1, ..., I_{|\mathbb{P}|}\}$, where $|\mathbb{P}|$ is the numbers of individuals, the next generation of a population is produced by applying a single point crossover and a mutation, on selected individuals. The DCA classification accuracy is used to evaluate the quality of individuals in the new generation of population. After the algorithm is terminated, the fittest individual in the current population is the optimal solution. From this, all the extracted rules are grouped together to form the final rule base.

Once the rule bases are generated for all three input signals, the TSK+ inference approach as introduced in Section 5.2.1 is applied with the training dataset, which generates the signal inputs for the DCA as illustrated in Figure 5.4.

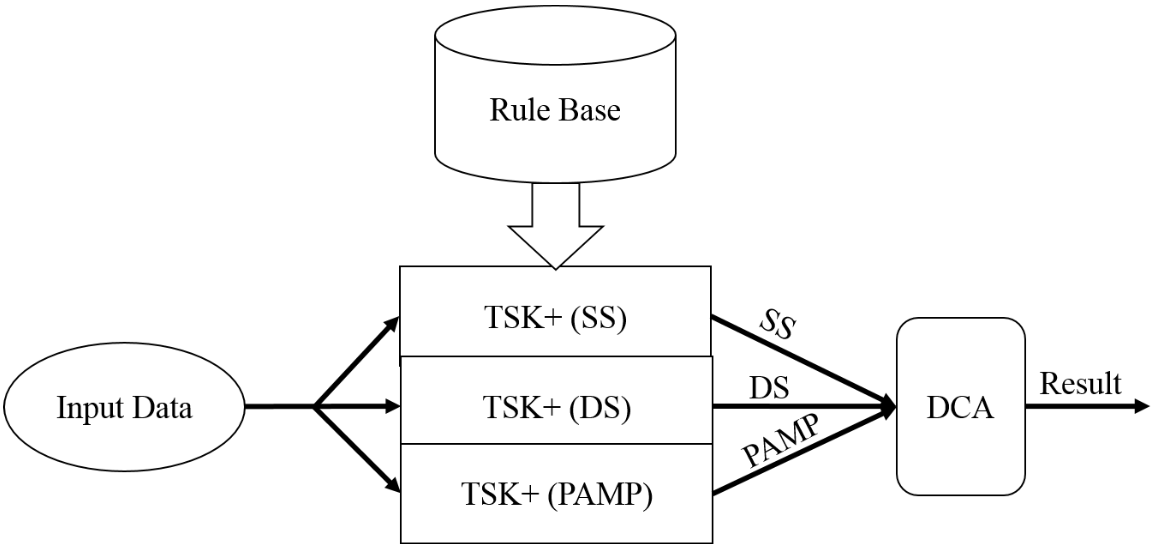


Figure 5.4 Inputs generation for DCA

Note that, the cluster numbers for each sub-dataset identified are listed in Table 5.1. Based on the results of the Elbow method, there are 21 TSK fuzzy rules supposed to be generated to contribute the final rule base. For instance, rule antecedents of one fuzzy rule in DS sub-rule base can be express by Equation 5.12:

Table 5.1 The number of clusters for each sub-dataset

	DS	**SS**	**PAMP**
Number of clusters	7	7	7

$$x_1 = (0, 0.39, 3.52) \text{ and } x_2 = (0, 0.39, 31.51) \tag{5.12}$$

Taking Equation 5.12 as an example, the optimised fuzzy rule is shown by Equation 5.13

$$\begin{aligned} R_1 : \textbf{IF } x_1 = (0, 0.39, 3.52) \text{ and } x_2 = (0, 0.39, 31.51) \\ \textbf{THEN } f_1(x_1, x_2) = 18.2x_1 - 4.05x_2 - 5.5 \ . \end{aligned} \tag{5.13}$$

5.3 GA-based Signal Categorisation for Dendritic Cell Algorithm

This section utilises GA based on partial shuffle mutation (PSM) [1] to automatically map the input features into three signals of the DCA. The initial studies on the DCA used expert knowledge of the problem domain to manually pre-determine the mapping between selected features and the three signal categories of either SS, PAMP or DS [73, 29]. Manual pre-processing phase has been criticised as it may over-fits the data to the algorithm; it is thus application dependent and requires a deep understanding of the problem domain [75]. The principal component analysis (PCA) was also applied to DCA for automatic feature categorisation [75] to overcome the limitations of the manual method. However, the PCA does not produce satisfactory classification performance and destroys the underlying meaning behind the initial features presented in the input dataset by generating a new set of features via dimensionality reduction. Another automatic signal categorisation approach is based on fuzzy-rough set theory (FRST), termed as FRST–DCA, which was proposed in [28] to overcome the shortcomings of the PCA approach. However, the FRST–DCA is an expensive solution to signal categorisation task due to information loss during the discretisation process; also, it is only practically applicable to simple datasets thanks to its computational complexity [29].

5.3.1 The Proposed Signal Categorisation Approach

The proposed PSM-DCA based system is illustrated in Figure 5.5. Firstly, given a dataset, feature selection process is applied to select the most significant features. Then, the PSM-DCA takes place to categorise the selected features into SS, PAMP or DS, each with s, p and m being the maximum number of features respectively. From here, the three signals are fed into the DCA by going through the signal processing phases to classify the data instances presented in the dataset. Note that the accuracy of the classification result was used during the training process as the fitness function of the GA, as demonstrated in Figure 5.5. The key component of the proposed system, that is signal categorisation using GAPSM, is detailed below.

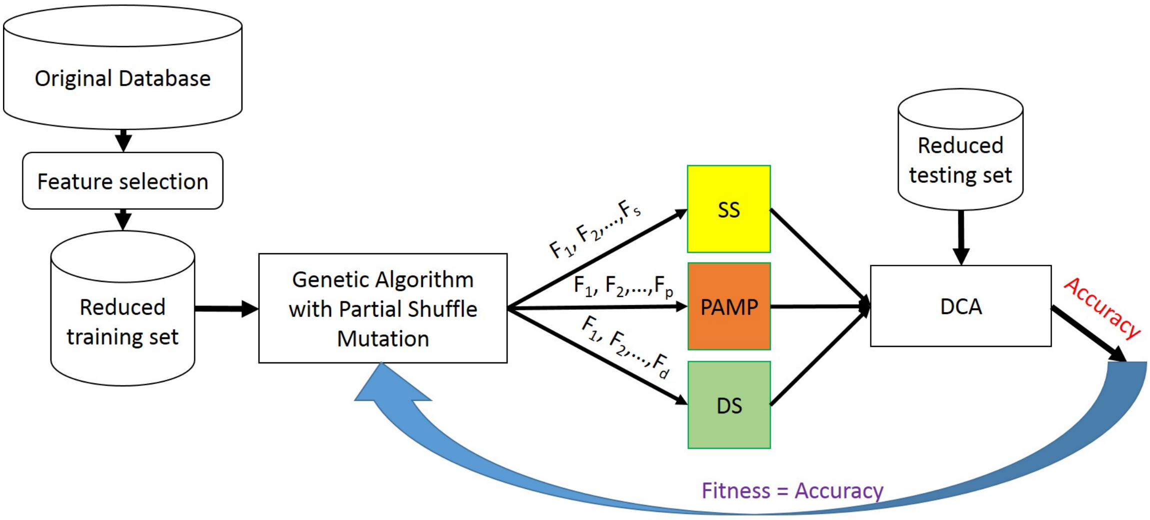


Figure 5.5 The proposed PSM-DCA based system

5.3.2 Signal Categorisation Using Partial Shuffle Mutation

Note that, the general requirements for the signal categorisation process are based on the definition of each signal type as follows [73]:

- *SS* is a confidence indicator of a normal condition associated with a particular feature.
- *PAMP* is a confidence indicator of an anomalous situation associated with a particular feature.

- *DS* is associated with an anomalous situation when its concentration is higher, however, when its concentration is lower under normal circumstances it is associated with the normal condition (i.e.; may or may not indicate abnormality).

According to these definitions, both *SS* and *PAMP* are positive indicators of normal and an anomalous situation respectively, whereas the *DS* indicates the situations where the risk of anomalousness may be high or the risk of normality may be present with a feature. Therefore, it can be seen that both *SS* and *PAMP* are more informative than *DS* meaning that both of them are indispensable in the order of importance; reflecting the first and the second-ranking in terms of signal priorities.

In the proposed method, given m number of selected features, the GA initialises a pre-defined number of individuals in a population each with m random values ranging from 1 to m. Note that, each value within an individual represent an index of a feature in the dataset, so, every value is unique. In order to be able to map each feature to its appropriate signal category, the following formulas were used to define the maximum number of features in each signal category based on different experiments of feature permutations (i.e.; empirical study).

s = 80% of percentage of normal samples within the dataset times the selected features (m) = 0.8 * percentage of normal samples * m.

p = 80% of percentage of anomaly samples within the dataset times the selected features (m) = 0.8 * percentage of anomaly samples * m.

d = the remaining set of features.

Where s, p and d represent the total number of feature in *SS*, *PAMP* and *DS* category respectively. Therefore, based on this empirical study, the proposed method assigns many features to *SS* and *PAMP* than to *DS* since both *SS* and *PAMP* are considered the most informative as it can be reflected from their definitions.

Accordingly, the first s indexes of features within an individual in a GA population belong to SS, the next p features belong to PAMP and the rest belong to DS. In order to perform GA-PSM operation on an individual after crossover, the GA select a number of values (indexes) with a probability of β and permute them which generate new arrangement of features within an individual. As a result, the new arrangement generates a new features-to-signal mapping and different classification performances. The same process repeats over a number of iterations until the GA converges. The following steps summarises these mechanisms followed by the GA:

1) Individual representation: An individual (I) within a population ($\mathbb{P}$) is designated as a possible solution that comprises of all m indexes of the selected features, where the first s

features belong to SS, followed by p features for PAMP and finally the last d features for DS; which is given by
$I = \{F_1^1, F_2^2, .., F_s^e, ..., F_1^f, F_2^g, .., F_p^i, ..., F_1^j, F_2^k, .., F_d^m\}$.

2) Population initialisation: In this study, the initial population $\mathbb{P} = \{I_1, I_2, ..., I_N\}$ is formed by initialising N individuals each containing m unique random numbers ranging from 1 to m, where each number represents an index of a feature within the database.

3) Objective function: The objective function is defined as the classification accuracy of the DCA.

4) Selection: The fitness proportionate selection method is employed for selecting a number of individuals who reproduce during crossover and partial shuffle mutation so as to evolve better individuals for the next iterations. In this technique, the probability of an individual to become a parent is proportional to its fitness.

5) Crossover: A single point crossover genetic operation is applied with a probability of α to increase the exploitation of search space.

6) Partial shuffle mutation (PSM): The PSM was proposed by [1] for fast solving the Travelling Salesman Problem. As its name suggests, in this study, the PSM shuffles the index of features within an individual in the population based on randomly selected indexes of the features with a probability p_m as detailed in Algorithm 7. This process is equivalent to the trying of a different combination or rearrangement of index of features. Suppose that the dataset is of 10 features, 55% of anomaly data samples and 45% of normal samples; then, s = 0.8*0.45*10=3, p = 0.8*0.55*10=4 and d = 3); the PSM process for this example is illustrated in Figure 5.6. When the PSM-DCA is applied the order of indexes of features within an individual is changed, so as the order and indexes in each signal category as shown in Figure 5.6. The following equations shows the percentage distribution of feature by the PSM.

$$s = 0.8 * percentage\_of\_normal\_samples * m \tag{5.14}$$

$$p = 0.8 * percentage\_of\_anomaly\_samples * m \tag{5.15}$$

$$d = the\_rest\_of\_the\_features. \tag{5.16}$$

After applying the PSM genetic operator to the individuals, the newly generated individuals and some of the best individual in the current iteration $\mathbb{P}$ jointly form the next generation of the population $\mathbb{P}'$. Note that, the performance of each individual is evaluated by the classification accuracy of the DCA.

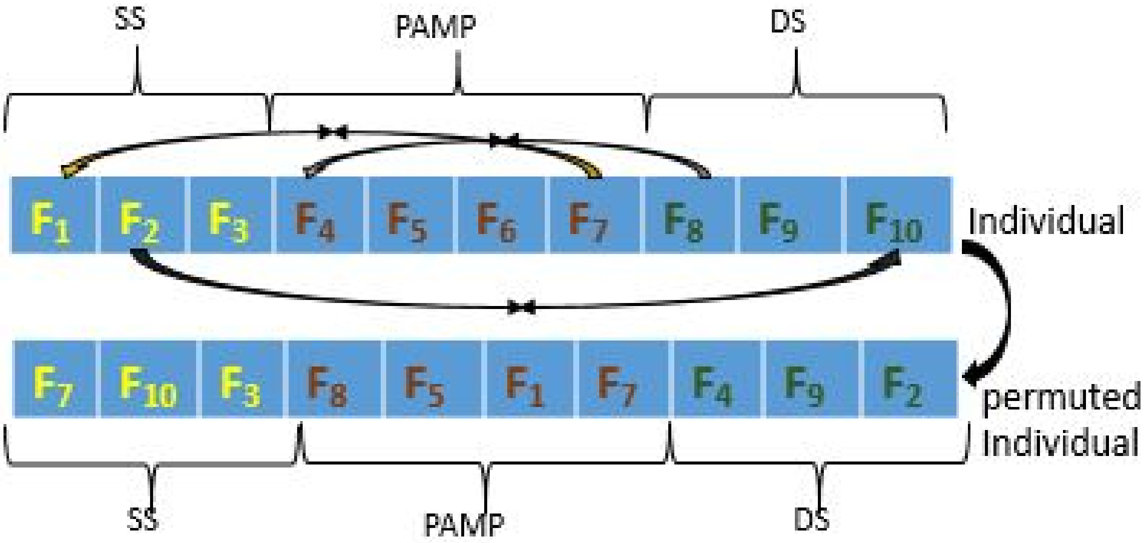


Figure 5.6 The partial shuffle mutation example

Algorithm 7 Signal Categorisation using Partial Shuffle Mutation

```
input: an individual $I$ and mutation probability $\beta$
output: a permuted individual $I'$
$i = 1$;
while $i \leq m$ do
    choose $p_m$ a random number between 0 and 1;
    if $p_m < \beta$ then
        choose $j$ a random number between 1 and $m$;
        permute $(F_i, F_j)$;
    end if
    $i = i + 1$;
end while
return a permuted individual $I'$.
```

7) Iteration and termination: The GA terminates when the pre-defined maximum number of iterations is reached or the classification accuracy of the DCA exceeds the pre-defined threshold of the optimum accuracy. When the GA terminates, the optimal solution of the categorised features is taken from the fittest individual in the current population. From this, the optimal features-to-signals mapping is applied to the DCA for classification tasks.

For instance, after applying the GA-PSM to the IDS datasets (presented in Sec 5.4.1) used in this work, the features were categorised as follows.

For the KDD99 dataset, *DS*={count and srv_count}, *SS*={logged_in, srv_different_host_rate and dst_host_count}, *PAMP*={serror_rate, srv_serror_rate, same_srv_rate, dst_host_serror and dst_host_rerror_rate}.

For the UNSW_NB15 dataset, *DS*={sbytes, dbytes, dload and dmean}, *SS*={dpkts, sttl, smean, ct_state_ttl, ct_dst_sport_ltm and ct_srv_dst }, *PAMP*={dur, rate, dttl, sload, ct_srv_src, ct_src_dport_ltm and ct_dst_src_ltm}.

For the CERT insider dataset, *DS*={number HTTP site visited and number of sensitive files accessed}, *SS*={number of log-on/log-off and number of emails sent/received}, *PAMP*={number of connecting and disconnecting external devices}.

5.4 Experimental Evaluation

This section describes the setup of experiments, validation and discussion about the results. Firstly, the performance of the three enhanced DCA approaches for cybersecurity attack detection in e-Government systems were evaluated by using publicly available cybersecurity datasets. Secondly, the performance of the proposed approaches were compared with some state-of-the-art classification methods. The comparative analysis was performed in order to investigate and validate the effectiveness of the proposed approaches in comparison to other classification methods. All experiments were implemented by using JAVA in NetBeans IDE 8.2. Then, the performance evaluation were performed by using an HP workstation with Intel® Xeon™ E5-16030 v4 CPU @3.70 GHz and 32GB RAM.

5.4.1 Benchmark Datasets

The proposed enhanced DCA approaches were validated by using one insider dataset namely the CERT [67] and two external intrusion detection datasets namely KDD99 [92] and UNSW_NB15 [117]. The description of each dataset is given in the next three subsections.

CERT Insider Threat Datasets

The CERT insider threat dataset V4.2 was used to validate the performance of the enhanced DCA approaches on insider threat detection [67]. The CERT dataset is a synthetic dataset that describes insiders' computers daily activities over a period of 17 month. The data were collected based on 1000 user accounts, of which 70 were performing malicious activities in the organisation. There are five different activities that insiders performed during this time period, including log-on/log-off to the computers, sending and receiving emails, connecting and disconnecting external devices to the computers, type of file accessed, and HTTP site visited. This dataset is pre-processed; in this work, the CERT dataset was split into two sets for training (80%) and testing (20%).

KDD99 External Dataset

KDD99 is an intrusion detection dataset which was generated for the third International Knowledge Discovery and Data Mining Tools Competition [92]. The goal was to build a network intrusion detector as a predictive model with an ability of distinguishing between bad (intrusion or attack) and good (normal) connections. The dataset is divided into two sets of training and testing, each set having 41 features. The training dataset contains 494,021 records (97,278 normal and 396,743 anomalous) while the testing dataset comprises of 311,029 records (60,593 normal and 250,436 anomalous).

There are four attack types included in both training and testing set of the KDD99 dataset which affect large number of networked systems like e-Government globally daily. These attack types are grouped as follows:

1. **DOS**: Denial of service attacks which attempt to shut down the system to make it inaccessible to its intended users– e.g. syn flooding, teardrop and smurf
2. **Probes**: An attempt of gaining access to a computer and its files by exploiting the weak points available through surveillance and other probing techniques, e.g. port scanning.
3. **U2R**: Unauthorised attempt to gain super user privileges by exploiting vulnerabilities that allow normal user to gain a root privileges, e.g. buffer overflow and rootkit attacks.
4. **R2L**: Unauthorised access of a computer resources from a remote machine, e.g. password guessing and ftp_write attacks.

UNSW_NB15 External Dataset

In 2015, UNSW_NB15 dataset was publicly published to support the evaluation of IDS [117]. It contains nine (9) new different moderns attack types which are not present in the KDD99 dataset. The new attack types include Reconnaissance, Shellcode, Exploit, Fuzzers, Worm, DoS, Backdoor, Analysis and Generic; and they are deadliest and lethal to organisations like e-Government systems so they need to be detected in advance. This dataset is further divided into training and testing sets, each set having a total of 49 features including the class label. The training dataset contains 175,341 records (56,000 normal and 119,341 anomalous) while the testing dataset comprises of 82,332 records (37,000 normal and 45,332 anomalous).

Furthermore, features in UNSW_NB dataset are categorised into six groups namely the flow features, basic features, content features, time features, additional generated features and labeled features. Flow features comprise of traffic flow captured between a client and server and the packet header which covers the behaviour of the network traffic. Basic features comprise of the attributes that characterise protocols connections. Content features contains the attributes of TCP/IP and http services. Time features comprise of the timing attributes such as arrival round trip time of TCP protocol and time between packets. The additional generated features are divided into two subgroups called general purpose features (features numbering from 36-40) and connection features (numbering from 41-47) [117].

5.4.2 Dataset Pre-processing

All three enhanced DCA approaches proposed in this chapter require feature selection process to be applied to the datasets to select the most informative features. Therefore, the Information Gain (IG) [166] method was exploited for feature selection. Additionally, signal categorisation process is require by the optimised DCA-GA and TSK-DCA approaches, so, feature-to-signal categorisation was performed by maximising the feature-to-class mutual information [166]. If an attribute has a higher mutual information with the normal class (maximising) and significant lower mutual information with the anomalous class (minimising), it is categorised as SS. If an attribute has higher mutual information with the anomalous class (maximising) but significant lower mutual information with the normal class (minimising), it is categorised as PAMP. Otherwise, the feature is categorised as DS.

5.4.3 Parameter Settings

As commonly used in the previous DCA studies [73, 29], a population of 100 DCs was initialised in the sampling pool and the size of mature pool was set to 10 DCs. The migration thresholds of DCs were assigned by a Gaussian distribution with a mean of 5.0 and standard deviation of 1. The anomaly threshold for both datasets was computed as the percentage of anomalies in the datasets. The parameter values used for the GA for validation are listed in Table 5.2 as they are being widely used [120, 107].

Table 5.2 The employed GA parameters

Number of Individuals	50
Number of Iterations	250
Crossover Rate	0.95
Mutation Rate	0.1

5.4.4 Measurement Metrics

The performances of proposed enhanced DCA approaches for e-Government system was evaluated via accuracy and sensitivity or true positive rate (TPR). The higher the sensitivity, the higher the true positives generated by the proposed enhanced DCA approach and the better it can detect cybersecurity attacks in the proposed e-Government system.

Accuracy and Sensitivity are defined as follows:

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\ Sensitivity(TPR) &= \frac{TP}{TP+FN}, \end{aligned} \tag{5.17}$$

where TP, FP, TN, and FN refer respectively to true positive, false positive, true negative and false negative, respectively.

Additionally, the performances of the three approaches were further validated by using precision and F-score to investigate how best they can perform when attacks in e-Government system contains an uneven traffics distribution between normal and malicious samples. Note that, high accuracy indicates that the enhanced DCA is doing better only when the dataset has equal or near equal samples between classes. F-measure is more effective than accuracy when a dataset is imbalanced.

Precision and F-score are computed as follows:

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} \\ Recall(TPR) &= \frac{TP}{TP+FN} \\ F-score &= \frac{2*Recall*Precision}{Recall+Precision}. \end{aligned} \tag{5.18}$$

5.4.5 Results and Discussion

This section presents the experimental results and discussion on the performances of the three enhanced DCA approaches.

Results on Sensitivity and Accuracy

The evaluation results and discussion on the sensitivity and accuracy for the three enhanced DCA approaches in comparison to the original DCA are presented in this subsection. Table 5.3 presents the sensitivity and accuracy performances of the three approaches from the three used datasets. The best performing approach's result for each dataset is marked in bold.

From Table 5.3, it can be noticed that, in comparison to the original DCA, the three proposed approaches for enhancing DCA have produced higher sensitivity and accuracy on all three datasets which make them more suitable and effective for detecting cybersecurity attacks in the proposed e-Government system and thus enforce data privacy and integrity. In fact, the three enhanced DCA approaches are more convenient in handling the imprecision and complexity that may be caused by manual derivations. Note that, the optimised GA-DCA has outperformed the other two approaches (i.e.; TSK-DCA and PSM-DCA) mainly because optimised GA-DCA search for optimal set of weights to enhance DCA performance, while, the TSK-DCA and PSM-DCA require pre-determined weights in order to enhance DCA so as to reduce computational complexity. More precisely, the optimised GA-DCA has performed better on all three datasets compared to the other approaches both on sensitivity and accuracy. The effective performance of the GA-DCA prove that it is more suitable for detecting cybersecurity attacks during blockchain transactions in the proposed e-Government system. So, once the enhanced DCA IDS is integrated in the proposed framework in this project, it will make it more robust to mitigate and eliminate the possible future cybersecurity risks and threats targeting e-Government systems.

The GA-DCA (best approach) fine-tuning process of the testing accuracies for the three datasets over 250 iterations is captured in Figure 5.7. The testing time required by the optimised approach in each single iteration is the same as that of the original DCA, the difference lies on the training time where the GA-DCA approach requires more time. The fine-tuning processes for the TSK-DCA and PSM-DCA follow similar characteristics but with different results that they produce.

Results on Precision and F-score

Cybersecurity datasets are often imbalanced due to the fact that attackers often try to inject few attacks to the network to avoid being detected. Therefore, F-score is the best measure for imbalanced datasets than accuracy. Table 5.4 presents the results on the precision and F-score. It can be seen that, the three proposed approaches in this project have produced better results on F-score, indicating that, if attackers inject few attacks in traffics going

Table 5.3 Results on Sensitivity and Accuracy

Dataset	Sensitivity (%)				Accuracy (%)			
	GA-DCA	TSK-DCA	PSM-DCA	orig. DCA	GA-DCA	TSK-DCA	PSM-DCA	orig. DCA
KDD99	**99.98**	99.64	98.25	89.88	**95.73**	94.05	93.44	83.67
UNSW_NB15	**97.48**	95.12	96.51	90.14	**92.59**	91.87	91.06	78.01
CERT	**97.65**	96.82	96.82	93.30	**96.52**	94.22	94.22	92.01

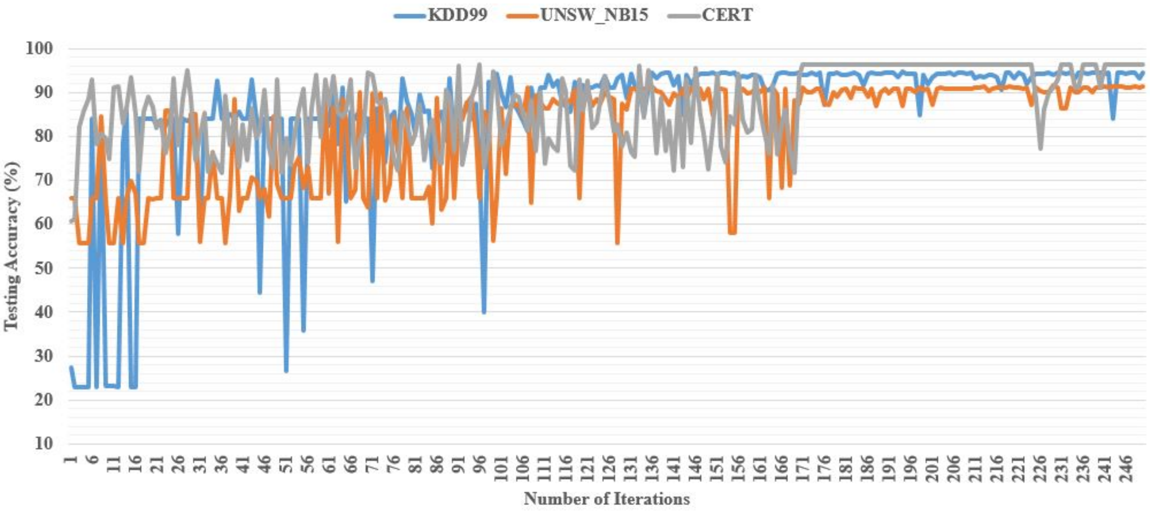


Figure 5.7 Optimised GA-DCA fine-tuning the testing accuracies

to e-Government network, they will be detected effectively. Again, since the optimised GA-DCA has performed much better than the other approaches including the original DCA version, it is more suitable for detecting anomalous traffics during blockchain transactions in the proposed e-Government system. Additionally, all three enhanced DCA approaches have outperformed the original DCA, which implies that the original DCA needed improvement to be much more effective in detecting attacks in networked system including e-Government systems.

Comparison between GA-DCA and other published studies based on different sampling strategies

Note that, in practice, the value of the migration threshold should be a compromise between DCs' sampling time and classification accuracy. If the migration threshold value is too low, the DC will migrate too quickly without sampling enough data items and features [71]. If the value of the migration threshold is set too high, the DC will migrate too slowly and misclassify the sampled data items. Therefore, the results of classification are not strictly comparable if different sampling strategies are used. Compared to other studies attempted to use DCA with different sampling strategies on the KDD99 dataset, the result produced by the proposed approach is better. For instance, the work of [135] named **Prob-DCA** uses GA to optimise the probabilities of SS and DS (i.e., p_safe and p_danger) with randomly assigned migration thresholds to DCs recorded a classification accuracy of 91.23% compared to the proposed GA-DCA which has produced a classification accuracy of **95.73%**. Also, the work of [28] used Fuzzy Rough Set Theory based DCA (FRST-DCA) for automatic feature

Table 5.4 Results on Precision and F-score

Dataset	Precision (%)				F-score (%)			
	GA-DCA	TSK-DCA	PSM-DCA	orig. DCA	GA-DCA	TSK-DCA	PSM-DCA	orig. DCA
KDD99	**88.96**	86.68	88.54	81.93	**94.15**	92.71	93.14	85.72
UNSW_NB15	89.19	**90.92**	87.31	85.11	**93.19**	92.97	91.68	87.55
CERT	92.93	89.66	**94.33**	74.35	**95.21**	93.10	95.56	82.31

selection and signal categorisation based on QuickReduct algorithm [27] with randomly assigned migration thresholds to DCs and produced a classification accuracy performance of 92.56% while the work of [75] used principal component analysis (PCA) to automatically categorise the features into signals with randomly generated migration thresholds for DCs produced a classification accuracy of 87.67%. Another work by [76] used information gain method (IG-DCA) for automatic feature selection and signal categorisation with randomly assigned migration thresholds to DCs and recorded a classification accuracy of 86.88%. From these results, it is clearly that the proposed approach of GA-DCA has performed better (**95.73%**) than other published works in the literature based on random DCs' sampling strategies. Table 5.5 summarises the comparison between the GA-DCA and other published works detailed here based on KDD99 dataset which was commonly used in all of them.

Table 5.5 Comparison between GA-DCA and other published studies on sampling

Approach	**Prob-DCA [135]**	**FRST [28]**	**PCA [75]**	**IG [76]**	**Proposed**
Accuracy (%)	91.23	92.56	87.67	86.88	**95.73**

Comparison with Other State-of-the-art Classifiers

Furthermore, since the optimised GA-DCA approached has performed better than the other two on all three datasets, its testing results were compared with five state-of-the-art classifiers namely DT, NB, SVM, RF and ANN. The experiments for DT, NB, SVM, RF and ANN classifiers were conducted by using Weka software [77] with the parameter values of the algorithms set to the default.

The comparison was performed in terms of the overall testing accuracies and the results are captured in Figure 5.8. It can be noticed that, in comparison to DT, NB, SVM, RF and ANN, the GA-DCA is capable of producing comparable overall classification result which make it suitable and effective for detecting and mitigating attacks targeting the proposed e-Government system.

Application of the enhanced DCA to other binary classification Datasets

The best approach (i.e.; optimised GA-DCA) proposed in this chapter was further applied to other binary classification datasets in order to investigate and validated its performance apart from the cybersecurity problems. The GA-DCA was applied to ten benchmark binary classification datasets from the UCI machine learning repository [40]. The properties of these datasets are provided in Table 5.6.

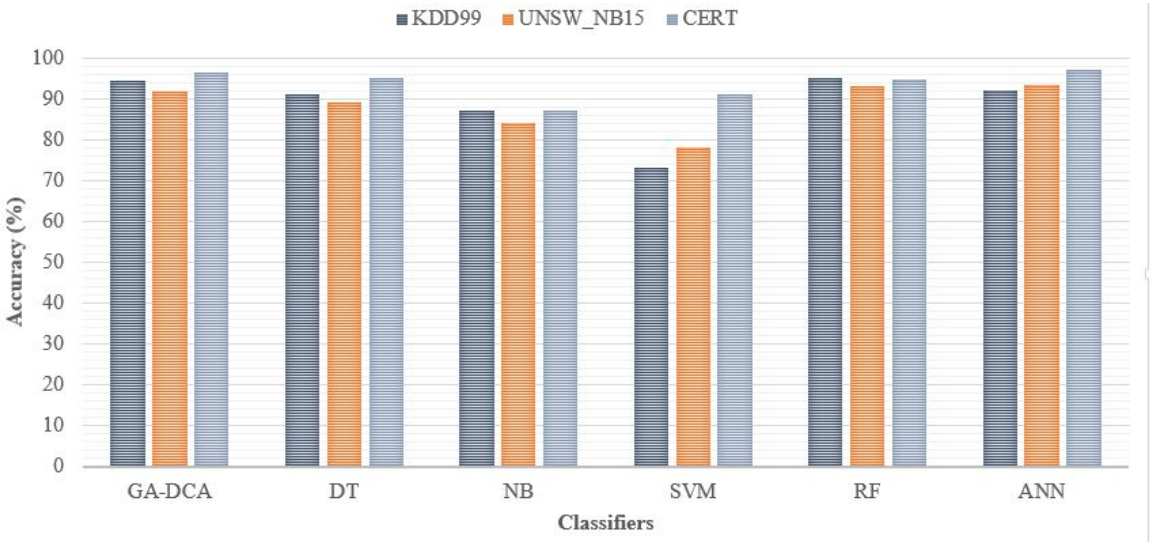


Figure 5.8 Comparison with other classifiers in terms of testing accuracy

Table 5.6 Benchmark datasets

Dataset	#Samples	#Features
Mammographic Mass (MM)	961	6
Pima Indians Diabetes (PID)	768	8
Blood Transfusion Service Center (BTSC)	748	5
Wisconsin Breast Cancer (WBC)	699	9
Ionosphere (IONO)	351	34
Liver Disorders (LD)	345	7
Haberman's Survival (HS)	306	4
Statlog (Heart) (STAT)	270	13
Sonor (SN)	208	61
Spambase (SB)	4601	58

Table 5.7 presents the comparison of the testing results on sensitivity, specificity and accuracy for the GA-DCA approach and the original DCA. The best performing result among the two approaches for each dataset are marked in bold.

From Table 5.7, the sensitivity, specificity and accuracy results indicate that, GA-DCA overall performs better than the original DCA. The GA-DCA generated better testing accuracies on all datasets except for MM dataset where it was slightly outperformed by the original DCA. Also, the sensitivity and specificity results of GA-DCA for most datasets outperformed

Table 5.7 Comparison of the GA-DCA with the original DCA

Dataset	Sensitivity (%)		Specificity (%)		Accuracy (%)	
	GA-DCA	orig. DCA	GA-DCA	orig. DCA	GA-DCA	orig. DCA
MM	**98.22**	97.10	95.73	**97.03**	96.75	**97.06**
PID	**96.29**	93.05	**97.36**	94.73	**96.73**	93.75
BTSC	**99.20**	98.47	**98.65**	97.87	**98.73**	97.93
WBC	**96.96**	96.53	**99.09**	97.68	**98.93**	97.64
IONO	96.18	**97.04**	**98.67**	97.33	**98.86**	97.15
LD	**97.55**	89.50	**99.37**	90.34	**98.55**	89.85
HS	84.38	**88.88**	**97.77**	94.22	**95.10**	92.81
STAT	**90.00**	82.50	**91.33**	84.55	**90.74**	83.75
SN	**99.77**	98.40	**99.89**	98.49	**9.84**	98.45
SB	**98.03**	94.12	**97.71**	91.93	**97.83**	94.23

the original DCA except on IONO, HS, MM and SN datasets. Therefore, these results further demonstrate that, GA-DCA can perform better in other binary classification problem.

The precision, recall and F-score for the proposed GA-DCA on all datasets are summarised in Figure 5.9. The results for the majority of datasets are effective, which further indicates that, optimised GA-DCA is applicable to imbalanced binary datasets with effective performances.

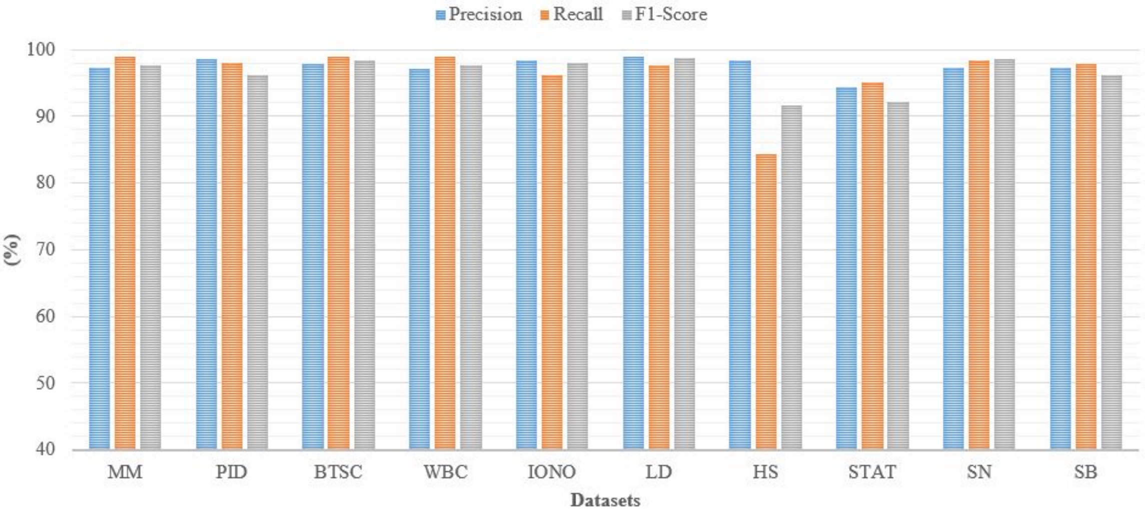


Figure 5.9 Precision, Recall and F-score for optimised GA-DCA

5.5 Summary

This chapter proposed three different cybersecurity attacks detection systems based on enhanced DCA for identifying and mitigating unwanted traffics in e-Government system. Firstly, a new parameters optimisation approach for the DCA was implemented by using GA; since the original DCA uses manual method to pre-defined the weights for its objective function. Secondly, fuzzy inference systems approach was used to developed an approach which can solve nonlinear relationship that may exist between input features and the resultant three DCA's signals during its pre-processing stage. Thirdly, a new signal categorisation method for the DCA was proposed based on partial shuffle mutation of GA to automatically categorise the input features into the three DCA's signal categories; given that the original DCA uses manual categorisation technique based on domain or expert knowledge of the domain. The experimental results show that the enhanced DCA approaches are capable of detecting cybersecurity attacks in e-Government system with better performances while simultaneously ensuring privacy to blockchain transactions. In particular, the optimised GA-DCA as produced much better performance and hence can be adopted in the proposed e-Government system to address any network traffic and transaction based malicious attacks.

Chapter 6

E-Government Multi-Attack Detection using Multi-Class DCA

Despite the excellent real-time performances produced by the enhanced DCA approaches developed in chapter 5, they are only applicable to binary classification problems. Note that, many practical classification problems including computer network data are associated with the classification of information which contain multiple classes. For instance, computer network data usually contains multiple attacks such as DDoS, DoS, malware, worm, SQL injection, information theft, and etc, which often attack the network concurrently. Although, all these attacks are categorised in one group as anomalies to make it easier for binary classifiers, it is of paramount importance for an IDS to detect and alert the system about each attack independently. This may help to identify an attack type(s) which is more common to a particular networks. Thus, using an IDS that can detect multiple attacks in e-Government systems will be beneficial for making informed decisions on preventive and mitigation measures regarding certain common occurring cybersecurity attacks.

Therefore, this chapter proposes an innovative generalisation of the original DCA to facilitate multi-class classification (McDCA) for multiple attacks detection in e-Government systems. This is achieved by transforming the concept of 'normal and abnormal contexts' to support multiple classes or 'contexts', which in the original version were designed to support binary classification by simply following the biological danger model. GA is used to determine the weighted sum functions bestowed upon each DC cell. Also, softmax regression [166] is employed in the context analysis phase to generating a probability distribution of each output context in the DC for class label assignment, given that the softmax regression is a generalisation of the logistic regression for the case when the input dataset has more than two classes [166]. Of course, the class with the highest probability is assigned to the DC as its context. From this, the same with the original DCA, a voting function is applied to each data instance for class assignment based on the contexts of its hosting DCs.

This chapter has also further transformed the proposed McDCA by stacking the DCs procedures in a multi-layer structure, termed as multi-layer McDCA, edging the algorithm closer to operating within the deep learning paradigm, in an effort to further enhance multi-attack detection performance in e-Government system and thus enable the exploration of the full potential of proposed McDCA. Despite some similarity may present between multi-layer McDCA and neural network, the fabric of any DCA implementation clearly provides a contrast with the neural network approaches, which clearly distinguish the two family of approaches. The multi-layer McDCA uses data transfer between agents in the algorithm, a novel innovation in DCA research, forming distinct pre- and post- processing layers; this facilitates the McDCA to be fully developed as a deep learning approach, which remains an important piece of active future work.

In order to validate and evaluate the proposed McDCA, firstly, ten widely used benchmark multiclass datasets from UCI machine learning repository [40] were used. Secondly, two cybersecurity datasets with multiple attacks namely KDD99 [92] and UNSW_NB15 [117] were adopted to evaluate its multiple attacks detection in e-Government system. The results support that McDCA can be used to detect multiple attacks targeting e-Government system. Additionally, the proposed McDCA was compared with some of existing multiclass classification techniques. The comparative results show that McDCA can perform competitively in reference to other existing state-of-the-art multiclass classification methods. The competitiveness of the proposed McDCA is demonstrated both in terms of prediction accuracy and computational efficiency. The contribution of this chapter is four-folds: 1) proposing the novel McDCA, 2) proposing the extended multi-layer McDCA, 3) validating and proposed approaches using popular benchmark datasets and 4) applying cybersecurity datasets to the proposed approach to investigate its multiple attacks detection performance in e-Government system.

The remainder of this chapter is structured as follows: Chapter 6.1 presents the proposed Multi-attack detection DCA. Chapter 6.2 demonstrates the experimentation process followed to validate and evaluate the proposed approach. Chapter 6.3 gives the discussion of the results, and finally Chapter 6.4 summarises the chapter.

6.1 Multi-Class DCA

The proposed McDCA is a general supervised multi-class classifier, which is able to perform classification tasks based on a given dataset or live data streams. As commonly used in conventional machine learning approaches, pre-processing measures, such as feature selection, are applied first to select the most significant features in discriminating the presented classes. Note that the traditional binary DCA requires signal categorisation, which groups the features into three categories to simulate the three types of signals emitted by an antigen, but the proposed McDCA herein takes all the selected features as system inputs by following the seminal work as reported in [44]. The consequence of this is that the signal of a class may be determined by a number of features or a feature may determine multiple classes, but the impact of this is compensated by the proposed weighting summation mechanism during the context detection stage as detailed in Section 6.1.2. The weights in this mechanism are determined using the GA algorithm in this chapter, which effectively optimises the aggregation mechanism based on the intrinsic feature of the training dataset.

6.1.1 System Overview

The proposed McDCA is summarised in Algorithm 8, as illustrated in Figure 6.1. Without losing generality, given a dataset with n data items, u selected features and k classes, the ith data instance $d_i, 1 \leq i \leq n$ can be represented as a pair $(\vec{x_i}, y_i)$, where $\vec{x_i}$ is a normalised feature vector $\vec{x_i} = \{x_1, x_2, \cdots, x_u\}$, and y_i $(y_i \in \{1, 2, \cdots, k\})$ is the class label. The McDCA first initialises a population of immature artificial DCs (i.e., iDCs) in a sampling pool. Then, the initialised iDCs sample data items and develop to mature DCs, and then the McDCA performs classification by going through three phases, including context detection, context assignment, and classification. During the context detection phase, the context of each mature DC regarding each class is determined using a generalised weighted summation function, as detailed in Subsection 6.1.2. In its context assignment phase, the detected context values for all possible classes are normalised using the softmax regression function; and then the most likely class of each mature DC is determined using the argmax function, as discussion in Subsection 6.1.3. This is followed by the final voting stage for data item classification, which is exactly the same with the original DCA in the Analysis and Classification phase. The two main phases of context detection and classification are detailed in the following subsections.

Algorithm 8 McDCA

```
input: the data stream $D$, DCs pool size $z$, mature pool size $q$, migration threshold $th$
output: predicted class
/** DCs initialisation**/
initialise sampling pool with $z$ iDCs;
/** Context Detection**/
while data instance $d$ in $D$ do
    select $q$ iDCs from the sampling pool;
    for 1 to $q$ do
        copy and store data item $d$ in iDC$_q$;
        get features $\vec{x_i} = \{x_1, x_2, \cdots, x_u\}$;
        compute $c_{csm}$;
        compute the cumulative value of $c_{csm}$;
        if cumulative value of $c_{csm} > th$ then
            move iDC to the mature pool;
            compute the concentration for each class $c_i$ $(1 \leq i \leq k)$;
            compute the cumulative value of each class $c_i$ from its concentration;
        end if
    end for
    /*Context Assessment */
    for each DC in the mature pool do
        calculate $c_i$ using Equation 6.2;
        apply Eqs. 6.4 and 6.5;
        attach the class index with the highest context value to all the data items in the DC;
        flush DC with $th$ assigned;
    end for
end while
/* McDCA classification*/
for each processed data item $d$ above do
    count the number of attachments representing each class $c_i$;
    label $d$ by class with the highest number attachments;
end for
return predicted class.
```

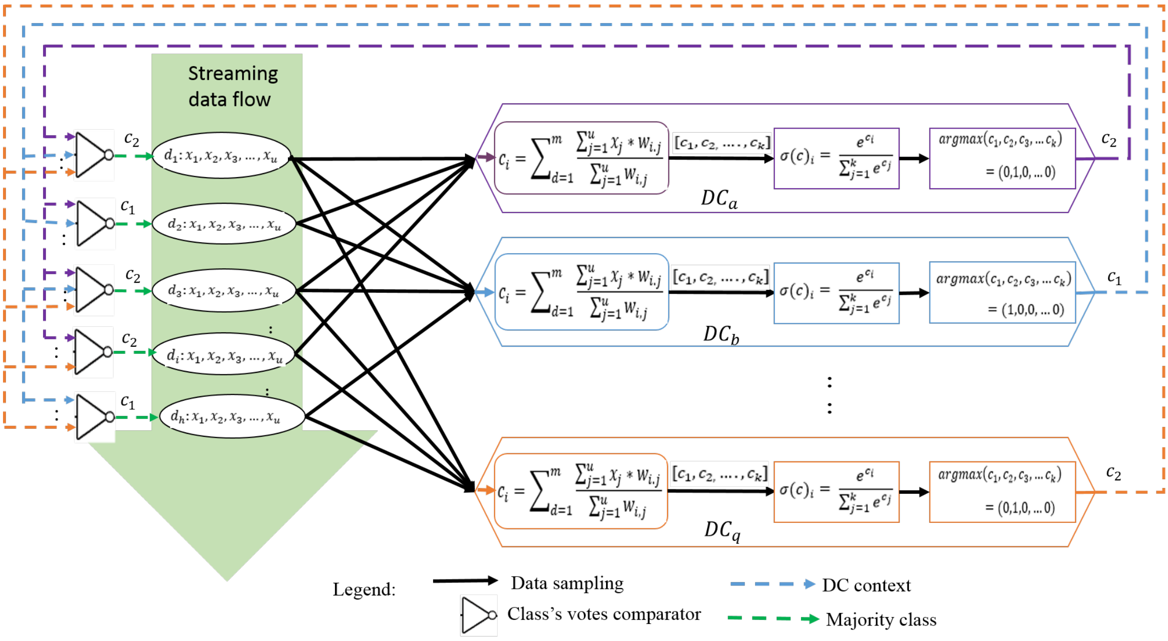

Streaming data flow
c_2
c_1
$d_1: x_1, x_2, x_3, \dots, x_u$
$d_2: x_1, x_2, x_3, \dots, x_u$
$d_3: x_1, x_2, x_3, \dots, x_u$
$d_i: x_1, x_2, x_3, \dots, x_u$
$d_h: x_1, x_2, x_3, \dots, x_u$
$c_i = \sum_{d=1}^{m} \frac{\sum_{j=1}^{u} x_j * W_{i,j}}{\sum_{j=1}^{u} W_{i,j}}$
$[c_1, c_2, \dots, c_k]$
$\sigma(c)_i = \frac{e^{c_i}}{\sum_{j=1}^{k} e^{c_j}}$
$argmax(c_1, c_2, c_3, \dots c_k)$
= (0,1,0, ... 0)
= (1,0,0, ... 0)
DC_a
DC_b
DC_q
Legend:
Data sampling
Class's votes comparator
DC context
Majority class

Figure 6.1 The proposed McDCA system

6.1.2 Parameters Optimisation

The proposed McDCA works for both static dataset and dynamic data stream. For static datasets, the size of the iDC sampling pool is pre-defined, whilst for streaming data, the sampling pool can be very big, and a pre-defined number of iDCs are randomly selected for processing in a batch. The size of the sampling pool in this work is determined empirically, but more investigation is required in the future. Suppose the size of the sampling pool is z; that is the sampling pool is initialised with z empty iDCs. For a given static dataset, all the iDCs sample data items randomly, and they move to the mature pool once they reach certain criterion based on the *csm* threshold th. Each iDC is assigned with a different migration threshold regarding the cumulative value of *csm*, and *csm* is a simulation of the concentration of biological co-stimulatory molecule [73]. This mechanism effectively limits the amount of time spent on data sampling. In this project, the values of th for iDCs are initialised in a Gaussian distribution determined from the characteristic behaviour of the dataset and the amount of data items that the iDCs can sample.

Context Detection

Suppose that m data instances have been sampled by the iDC overtime; its cumulative value of *csm*, denoted as c_{csm}, is determined using a generalised weighted function as expressed below:

$$c_{csm} = \sum_{d=1}^{m} \frac{\sum_{j=1}^{u}(x_j * w_{csm,j})}{\sum_{j=1}^{u} w_{csm,j}}, \tag{6.1}$$

where x_j is the normalised value of attribute j, $w_{csm,j}$ is the weight of attribute j regarding the csm. The weights $w_{csm,j}$ is determined during the training process using any general search algorithm and GA is used this work as discussed in Section 6.1.2. The value of m varies for different iDCs depending on the assigned th. Note that Equation 6.1 is a generalisation of the binary DCA equation (function).

As soon as the c_{csm} value of the iDC exceeds its assigned threshold th, the iDC has developed to a mature DC which ceases sampling and moves to the mature DC pool, and they are ready for contexts calculation. Once the mature DC pool reaches a certain size, the cumulative context value of each iDC regarding each class is calculated. Given a mature DC, its cumulative context value regarding class i, denoted as c_i, can be calculated using a generalised weighted function:

$$c_i = \sum_{d=1}^{m} \frac{\sum_{j=1}^{u}(x_j * w_{i,j})}{\sum_{j=1,1}^{u} w_{i,j}}. \tag{6.2}$$

Although this equation shares similar form with Equation 6.1, it takes very different set of weights. All the weights are generated using a GA in this chapter. The weights required for the calculation can be summarised as a matrix:

$$W = \begin{bmatrix} w_{1,1} & w_{1,2} & ..w_{1,j} & ..w_{1,u} \\ w_{2,1} & w_{2,2} & ..w_{2,j} & ..w_{2,u} \\ \vdots & & & \\ w_{k,1} & w_{k,2} & ..w_{k,j} & ..w_{k,u} \\ w_{csm,1} & w_{csm,2} & ..w_{csm,j} & ..w_{csm,u} \end{bmatrix}, \tag{6.3}$$

where u indicates the number of features of the dataset, and k represents the number of classes. Apparently, Equation 6.2 is also a generalisation of binary DCA function.

Weights Optimisation by GA

The optimal set of weights for Eqs. 6.1 and 6.2 can be generated using any general optimisation algorithm based on a training dataset; and the GA is utilised here in this seminal work as it has been successfully utilised for weights optimisation for binary DCA [53]. The GA algorithm starts by initialising a population of random individuals, with each representing a possible solution. Then, the population evolves through a number of operations, such as elitism, mutation and crossover; and more effective individuals are evolved over a number of iterations until a specified level of performance or maximum number of iterations is reached.

In this work, an individual (I) is a vector comprising of all the weights as listed in Equation 6.3. The size of the population is a problem-specific adjustable parameter, typically in a range from tens to thousands, with 20-50 being widely used [120]. The objective function is used to guide the elitism process and also determine the termination of the algorithm, which is simply defined as the McDCA classification accuracy in this work. The roulette wheel selection method is implemented for selecting a number of individuals who reproduce during crossover and mutation so as to evolve better individuals for the next iterations. When the GA terminates either by reaching the maximal number of iterations or pr-defined accuracy requirement, the optimal solution is taken from the fittest individual in the current population as the weights for Equation 6.3.

6.1.3 Classification

A softmax regression function is employed in this phase to assess the context of each matured DC using the cumulative values regarding all possible classes, that is which class the DC most likely belongs to. In logistic regression, the outcome of class label prediction is always binary or dichotomous (i.e., true or false) [166]. The softmax regression is the generalisation of logistic regression for multiclass classification under the assumption that the classes are mutually exclusive. It is widely used in neural network to convert the outputs of input processing function into the probability distribution of input classes present in the dataset so as to simplify the multiclass classification tasks [166]. The softmax regression does not process input features from the dataset rather it aids linear or nonlinear machine learning functions to perform prediction on multiple classes. In other words, the softmax regression function takes a k-dimensional vector of arbitrary real values and generates an output of another k-dimensional vector with real values in the range [0, 1] that add up to 1.

Note that, during this phase, if the data items in the data source are finished and the DCs haven't attained the migration thresholds, they are forced to migrate from the sampling pool, then queried for their context values before being flushed. Denote the vector of k cumulative context values, calculated using Equation 6.2, as $\vec{c}$ and $\vec{c} = \{c_1, c_2, \cdots, c_k\}$. Take vector $\vec{c}$ as the input, the softmax function calculates the probability distribution of the cumulative contexts as expressed below:

$$\sigma(\vec{c})_i = \frac{e^{c_i}}{\sum_{j=1}^{k} e^{c_j}}. \tag{6.4}$$

This softmax function generates a k-dimensional vector of probabilities associated with the likelihood of the context of the DC. Then, the context of the DC is calculated using the argmax function:

$$context = \arg \max_{i \in \{1,\dots,k\}} \sigma(\vec{c})_i. \tag{6.5}$$

Based on this equation, the class with the highest probability will be assigned to the DC as its context. Accordingly, all the sampled data instances are also assigned with the same class.

Once the context of each data item in each mature DC is determined, the proposed McDCA uses exactly the same approach as the conventional DCA uses, that is the majority votes, to determine the lable of each data item. Briefly, the class that has the highest votes by the DCs that have sampled the data item will be labelled as the class for the data item.

Runtime Complexity Analysis

The runtime complexity of the McDCA system is performed in this sub-section based on Algorithm 8. The McDCA algorithm is presented by combining the procedural operations with the while loop, for loop or if statements. Generally, runtime analysis is performed by calculating the number of steps or standard primitive operations executed by the algorithm [31]. Let n be the number of data instances within the dataset, u be the number of features selected, z be the size of the DC population and q the size of mature pool. Assume that, the GA population size is p and r is the number of iterations it has gone through.

Let $T_1(n)$, $T_2(n)$, $T_3(n)$, $T_4(n)$ denote the runtime complexity of of the four phases of McDCA, thus overall runtime complexity of the McDCA can be evaluated as $T(n)$= $T_1(n)$ + $T_2(n)$ + $T_3(n)$ + $T_4(n)$.

i Initialisation phase

The runtime of the initialization phase is independent of the number of instances n, it is executed only once and determined by the population size z. So, the runtime at this phase is calculated as

$T_1(n) = \mathcal{O}(z)$

ii Context detection phase

The runtime of the context detection phase depends on the number of data instances n, the number of features u, the DC population size z, the number of times a data items is sampled by multiples DCs q. Note that, the training and utilising runtime differ in this phase. The McDCA training requires the GA for optimisation whilst testing does not. For training, the GA population size is p, the number of iterations is r and the crossover and mutations rates which are done once per iteration, and thus: $T_2(n) = \mathcal{O}(n) + \mathcal{O}(qun) + \mathcal{O}(przn)$. This can be simplified as $T_2(n) = \mathcal{O}(nzpr)$. Without the GA, the runtime of the detection phase for testing becomes $T_2'(n) = \mathcal{O}(nz)$.

iii Context assessment phase
The runtime of the context assessment phase depends on the size of DC population z, the number of data instances n, and the number of times a data item is processed by DCs q. The training runtime is dependent on the number of iterations r and population size p of the GA. The runtime for the training phase thus can be calculated as: $T_3(n) = \mathcal{O}(pr(zun))$. Assuming u is very small in comparison to n then, $T_3(n) = \mathcal{O}(przn)$. The testing runtime of the assessment phase is given as: $T_3'(n) = \mathcal{O}(zn)$.

iv Classification phase
The runtime of the classification phase depends on data size n and the DCs population size z. Similarly, the training runtime complexity depends on the GA operations. In this phase, the data items are analysed n times to determine the number of presentations by multiple DCs in the context detection and assessment phases: $T_4(n) = \mathcal{O}((n)przn)$ $= \mathcal{O}(przn^2) = \mathcal{O}(prn^2)$. The testing runtime of the classification phase is done once without GA and is given as: $T_4'(n) = \mathcal{O}((n)nz) = \mathcal{O}(n^2)$.

From this, the total training runtime complexity becomes: $T(n) = T_1(n) + T_2(n) + T_3(n)$ $+ T_4(n) = \mathcal{O}(z) + \mathcal{O}(nzpr) + \mathcal{O}(prn^2) + \mathcal{O}(przn) = \mathcal{O}(prn^2))$. The total testing runtime complexity becomes: $(T'(n)) = T_1(n) + T_2'(n) + T_3'(n) + T_4'(n) = \mathcal{O}(z) + \mathcal{O}(nz) + \mathcal{O}(n^2) + \mathcal{O}(nz) = \mathcal{O}(n^2)$. Therefore, the worst case training and testing runtime complexities of the McDCA are quadratic given by $\mathcal{O}(prn^2))$ and $\mathcal{O}(n^2)$ respectively with n the dataset size, p the GA population size and r the number of iteration performed by the GA.

6.1.4 Multi-layer DCA

This section extends the proposed McDCA by using a multi-layer structure, to investigate the full potential of the proposed McDCA. The configuration and structure of the first layer of multi-layer McDCA is similar to that of McDCA while the succeeding layer(s) use outputs from the preceding layer of DCs to generate their context values as depicted in Figure 6.2. The number of DCs within different layers may vary depending on the specific application. The index number of each data instance is used to trace it while being processed within the system. The final prediction results are generated from the last layer (context assignment and labeling) as briefed in the following subsections.

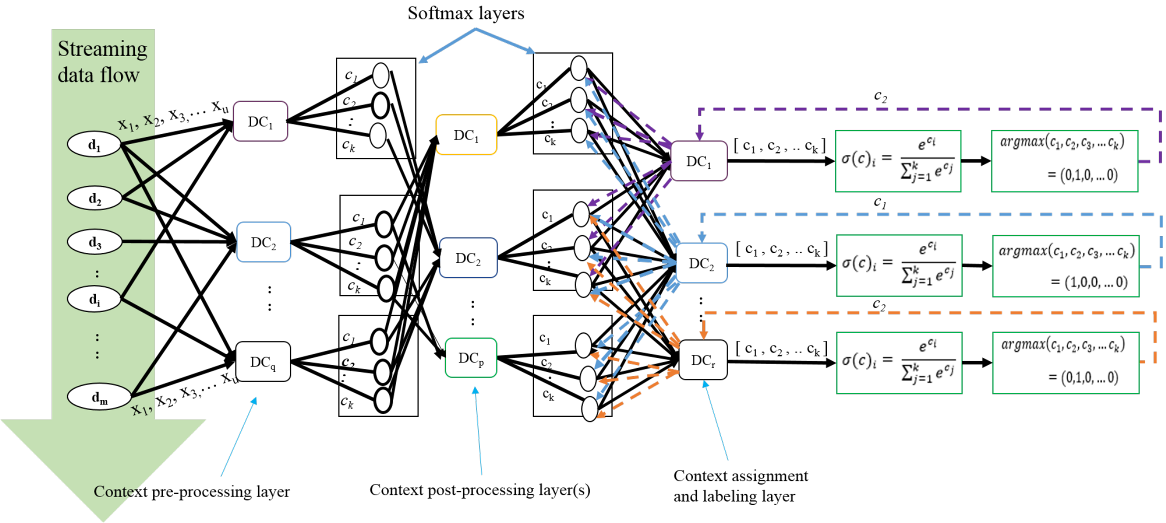

Softmax layers
Streaming data flow
$x_1, x_2, x_3, \dots x_u$
d_1
d_2
d_3
d_i
d_m
DC_1
DC_2
DC_q
c_1
c_2
c_k
DC_1
DC_2
DC_p
DC_1
DC_2
DC_r
$[c_1, c_2, .. c_k]$
$\sigma(c)_i = \frac{e^{c_i}}{\sum_{j=1}^{k} e^{c_j}}$
$argmax(c_1, c_2, c_3, \dots c_k) = (0,1,0, \dots 0)$
$argmax(c_1, c_2, c_3, \dots c_k) = (1,0,0, \dots 0)$
Context pre-processing layer
Context post-processing layer(s)
Context assignment and labeling layer

Figure 6.2 The proposed multi-layer McDCA system

Context Pre-processing Layer

In this layer, each DC uses Equation 6.1 to compute the cumulative value, c_{csm}. Once the cumulative c_{csm} value of a DC reaches the migration threshold (th), it stops data sampling and uses the weighted function of Equation 6.2 to process the input features for different data items it has sampled and generate k output cumulative context values. Then, it uses softmax function given by Equation 6.4 to normalise each of the k cumulative context values in a range of 0 to 1. Ultimately, DC forwards the normalised context values to the post-processing layer of DCs. Note that, migration thresholds were applied to DCs in the pre-processing layer only to ensure synchronisation between the post-processing layer(s) and context assignment and labeling layer. Once the DCs in the pre-processing layer stop sampling the data items, the DCs in the post-processing layers process the data and perform prediction before the McDCA start its next cycles of picking DCs for context pre-processing. The weights associated with features were generated and optimised using GA, the same as that described in Section 6.1.2.

Context Post-processing Layer(s)

For the sake of synchronisation, a fix number of DCs without any assigned migration thresholds are used in this layer(s). As soon as DCs in this layer receive input contexts from the preceding layer, the following equation with the support weights generated by GA are used to compute the cumulative context values corresponding to the k classes:

$$c_i = \frac{\sum_{DC=1}^{v}(y_{i,DC} * w_{i,DC})}{\sum_{DC=1}^{v} w_{i,DC}}, \tag{6.6}$$

where, $y_{i,DC}$ is the cumulative context value of a DC corresponding to class i and $w_{i,DC}$ is the weight associated with the input cumulative context $y_{i,DC}$ of a DC.

Then, DCs use softmax function to normalise the generated cumulative context values and produce outputs that are forwarded to the succeeding layer:

$$y_i = \frac{e^{c_i}}{\sum_{j=1}^{k} e^{c_j}}, \tag{6.7}$$

where, c_i is the cumulative context value corresponding to class i and y_i is the normalised cumulative context value.

Context Assignment and Labeling Layer

This layer is responsible for assigning the final context and class label to each data item by counting the number of context attachments made by the hosting DCs regarding each class. The same as that in the post-processing layer, a fixed number of DCs are directly employed without assigning them migration thresholds. DCs in this layer take the outputs from the post-processing layer as their inputs and uses Equation 6.6 to generate the k cumulative values. Then, they use the softmax function, i.e., Equation 6.7 and argmax function, i.e., Equation 6.5 as described in Section 6.1.3 to assign the context and label to data item from the final cumulative context values.

6.2 Experimental Evaluation

Firstly, the proposed McDCA was applied to 10 multiclass benchmark datasets, that are commonly employed in the literature, to validate and evaluate its performance in comparison to that of the existing multiclass classification approaches. Secondly, the McDCA was applied to two cybersecurity datasets to investigate its multiple attacks detection performance in e-Government system. All experiments were performed using an HP workstation with Intel$^{®}$ Xeon^{TM} E5-16030 v4 CPU @3.70 GHz and 32GB RAM, with code implementations done by using Java in NetBeans IDE 8.2.

6.2.1 Benchmark Datasets

The benchmark multiclass datasets were taken from the UCI machine learning repository [40] and two computer anomaly detection datasets namely KDD99 and UNSW_NB15 were taken from [92] and [117] respectively. Note that, the attack types present in the KDD99 and UNSW_NB15 datasets were used to represent class labels. The selected datasets covers a wide range of sizes from small, to very large as listed in Table 6.1:

- The datasets with small size and low dimensions are Iris, Wine, Seeds, Glass, Vowels, Segment and Vehicle.
- The datasets with relatively large sizes and low dimensions are PedDigits and SATimage.
- The datasets with large size and high dimensions are Covertype, KDD99 and UNSW_NB15.

If training and test datasets are not provided separately, ten-fold cross-validation was used; with the testing performance result for each dataset measured as the mean accuracy of ten-folds; otherwise, the training sub datasets were used for training and the test datasets were employed for testing, which follows the general practice as reported in the referenced papers for consistent comparisons.

Table 6.1 Multiclass benchmark datasets

Dataset	Training	Testing	Classes	Features
Iris	150	-	3	4
Wine	178	-	3	13
Seeds	210	-	3	7
CoverType	464,810	116,202	7	54
Glass	214	-	7	10
PenDigits	7,494	3,498	10	16
Vowels	528	-	11	10
Segment	210	2,100	7	19
Vehicle	846	-	4	18
SATimage	4,435	2,000	6	36

Description of Cyberscurity Datasets

The two cybersecurity datasets, KDD99 and UNSW_NB15 are briefed below:

KDD99 Dataset contains 494,021 training records (97,278 normal and 396,743 anomalous) and 311,029 testing records (60,593 normal and 250,436 anomalous); each set with 41 attributes [92]. Also, it is comprised of four (4) different attack types and the normal class. The four attacks present in the KDD99 dataset include **DoS**, **Probe**, **user to root (U2R)**, and **remote to local (R2L)**. Therefore, this makes up a multiclass dataset with 5 classes (4 attacks and 1 normal class)

UNSW_NB15 Dataset contains 175,341 training records (56,000 normal and 119,341 anomalous) and 82,332 testing records (37,000 normal and 45,332 anomalous); each set with 49 attributes [117]. It further comprises nine (9) different moderns attack types and the normal class. The 9 attack types include the **Reconnaissance**, **Shellcode**, **Exploit**, **Fuzzers**, **Worm**, **DoS**, **Backdoor**, **Analysis** and **Generic**. Therefore, this makes up a multiclass dataset with 10 classes (9 attacks and 1 normal class).

6.2.2 Experimental Setup

This section describes the experimental design and setup followed to validate and evaluate the performance of the proposed McDCA in e-Government system.

Data Pre-processing

Information Gain (IG) [166] was employed to perform feature selection due to its simplicity and efficiency. Then, the selected features were normalised using the min-max normalisation.

DCs Initialisation and Sampling

For all experiments, the size of the mature pool was set to 10, which were determined based on some empirical study. Based on empirical study, the activation thresholds (ths) were generated from a Gaussian distribution with mean of 7.5 and standard deviation of 2.5 to ensure that the McDCA generates best classification performances from the DCs in each cycle. The identification number of each data item within the dataset was used to track it while being processed by multiple DCs.

Weight parameters optimisation using GA

The GA was employed to generate the optimal weights for the weighted function. The parameters for the GA are listed in Table 6.2 below.

Table 6.2 The employed GA parameters

Number of Individuals	50
Number of Iterations	250
Crossover Rate	0.95
Mutation Rate	0.1

6.2.3 Results

This section presents the evaluation of the experimental results.

Analysis of the Migration Thresholds

In practice, the value of the migration th should be a compromise between iDCs' sampling time and classification accuracy. If the value of th is too low, an iDC will migrate too quickly without sampling enough data items [70]. If the value of th is set too high, an iDC will migrate too slowly and misclassify the sampled data items.

In order to determine the best performing th values for each dataset, seven experiments were performed using seven Gaussian distributions of th. Table 6.3 presents the testing results obtained for seven Gaussian distributions of th starting from mean of 2.5 and standard deviation of 1 to mean of 17.5 and standard deviation of 1.0 with the best performing th for each dataset marked in bold. The result can further be visualised in Figure 6.3.

The testing classification performances for majority of datasets were stable when the th values were initialised in a Gaussian distribution with mean of 2.0 to 10 and standard deviation of 1.0. Therefore, for the rest of experiments in this chapter, the values of th were initialised in a Gaussian distribution with mean of 7.5 and standard deviation of 2.5 to ensure that the values of th vary between 5.0 and 10.0 and iDCs can persist over multiple iterations. The values of th in this distribution were found to allow iDCs to sample enough data instances without ceasing sampling too fast or too slower. It is clearly that when the values of migration thresholds for iDCs are set to more than 10, DCs makes more classification error which in turn generate poor accuracy. Additionally, from this result it can be concluded that, the appropriate selection of the th's values is critical to the classification performance of the McDCA.

Table 6.3 Analysis of the migration threshold and test accuracy

Dataset	Different Gaussian distributions of the th						
	mean=2.5	mean=5.0	mean=7.5	mean= 10.0	mean=12.5	mean=15.0	mean=17.5
	stdev=1.0	stdev=1.0	stdev=1.0	stdev=1.0	stdev=1.0	stdev=1.0	stdev=1.0
Iris	**96.67**	**96.67**	**96.67**	**96.67**	90.33	91.55	89.34
Wine	97.14	**99.44**	**99.44**	**99.44**	98.89	93.33	90.12
Seeds	**99.05**	98.58	**99.05**	95.22	95.22	91.67	89.44
CoverType	**77.71**	**77.71**	74.37	73.43	73.43	62.70	60.12
Glass	**86.45**	**86.45**	**86.45**	**86.45**	78.63	71.23	68.44
PenDigits	78.20	**77.60**	**77.60**	**77.60**	71.10	72.77	69.89
Vowels	70.20	**69.40**	**69.40**	67.29	63.55	58.19	59.71
Segment	96.53	**96.62**	**96.62**	94.91	92.68	88.46	87.62
Vehicle	**88.89**	**88.89**	**88.89**	86.90	82.56	81.46	75.62
SATimage	89.36	**91.70**	**91.70**	86.36	82.56	82.56	82.56

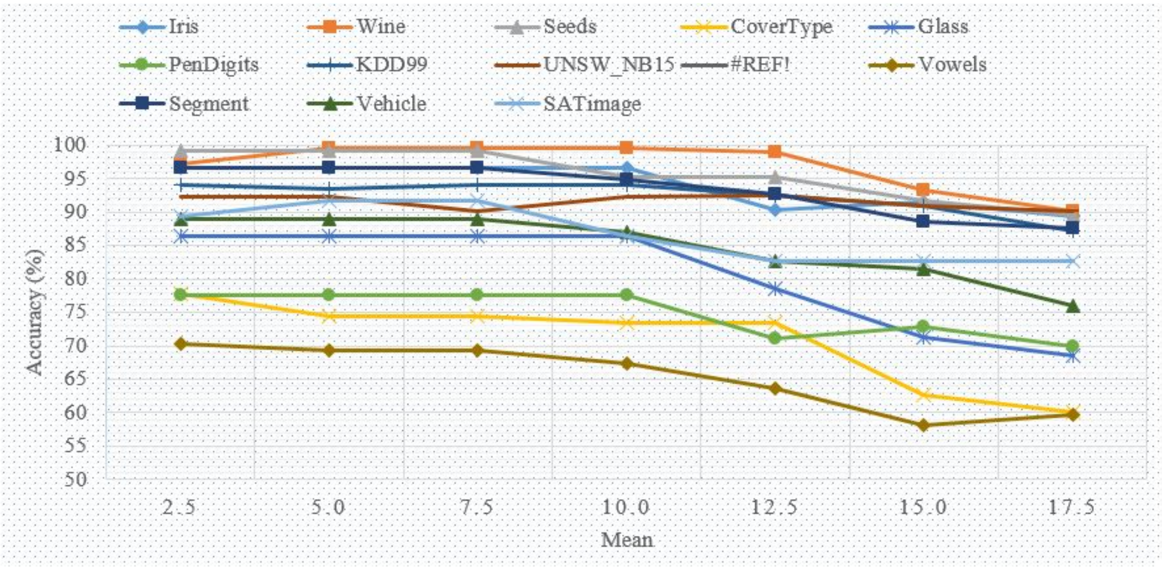


Figure 6.3 The range of migration thresholds VS test accuracy

The validation process of the testing accuracies for four datasets over 250 iterations of the GA is captured in Figure 6.4. The Iris, KDD99, Glass and Wine datasets were taken as the samples to show that 250 iterations are enough to optimise the weights of the McDCA. The rest of the datasets follow similar trends. It can be noticed that, from iteration number 150 to 250 the optimised weights produce almost no variation on the testing accuracies which indicates that new individuals weights getting generated by the GA cannot introduce any further improvements to the testing accuracies.

Comparison with Other Multiclass Approaches from the literatures

Based on the ten benchmark multiclass datasets, this section compares the performance of McDCA and other multiclass techniques which were recently proposed by different scholars in the literatures including Least Square and Proximal SVM based on Extreme Learning Machines (SVM-ELM) [81], Multiclass Classifier Based on Immune System Principles (AISLFS) [43], Directed Acyclic Graph SVM (DAGSVM) based on one-against-one [79], Sparse Extreme Learning Machines (SELM) [11], Vector-Valued Regularized Kernel Function Approximation (VVRKFA) [65], Multiclass Relevance Vector Machines (mRVM) [132], Sparse Bayesian Extreme Learning Machine (SBELM)[111] and Discriminative Clustering via Extreme Learning Machine (DC-ELM) [80].

The comparison results are presented in Table 6.4 where the performance results of McDCA are shown in the first column and the rest of the columns present the best accuracies

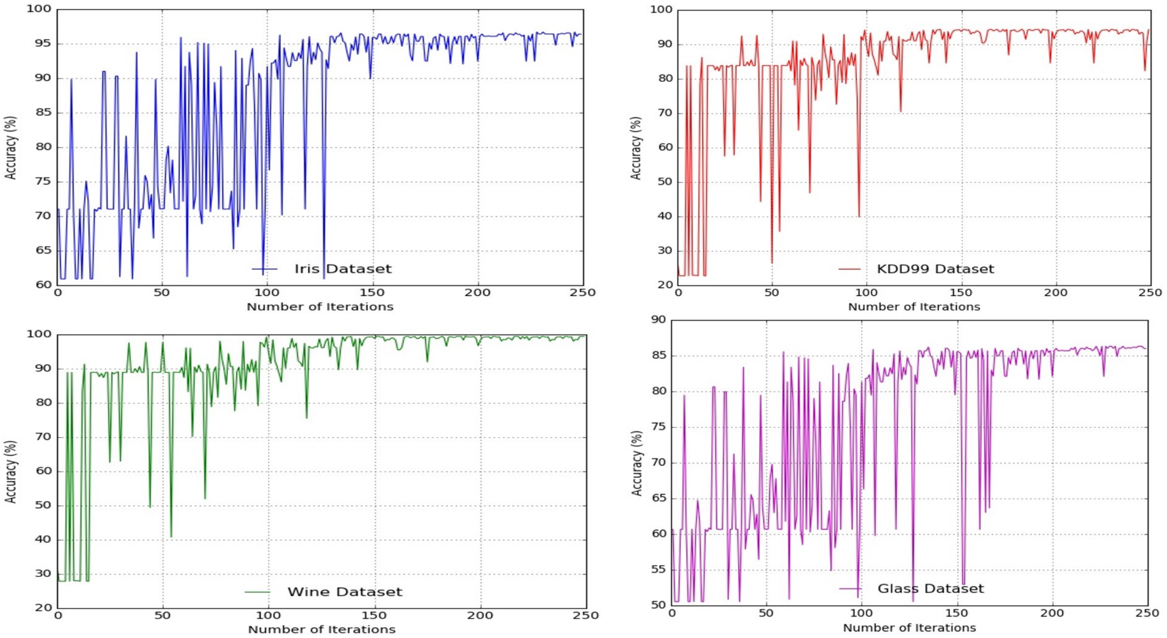


Figure 6.4 Testing accuracies validation over 250 iterations

taken from the studies of the compared approaches. The best performing accuracy among these approaches and McDCA for each dataset is marked in bold. Some datasets used in this work were not used by authors of the compared approaches, therefore, the values are left blank in Table 6.4.

From Table 6.4, the results indicate that, the classification accuracies of the McDCA compare well with other other multiclass approaches in the literature, which confirms that the McDCA is applicable to multiclass problems with competitive performances. It can be noticed that, the McDCA outperforms the rest of other approaches on the Wine and Breast cancer datasets. Except on the Vowels and Glass datasets where there is a significant differences between the best classifier's accuracy and the accuracy from the McDCA, for the rest of the datasets, there is not significant difference between the best classifier's accuracy and the McDCA. This proves that, the GA is able to well optimise the required weights with good generalisation ability between the training and testing datasets.

Comparison with State-of-the-art machine learning algorithms

Similarly, based on the ten benchmark multiclass datasets, this section compares the McDCA with the well-known state-of-the-art algorithms in machine learning which are the Support Vector Machine based on Sequential Minimal Optimisation algorithm (SVM-SMO), Artificial Neural Network (ANN), Decision Tree (DT) and Naive Bayes (NB). The parameters of SVM,

Table 6.4 Comparison with other multi-class classification approaches in the literatures

Dataset	Testing Accuracy (%)								
	McDCA	ELM[81]	LFS[43]	DAG[79]	SELM[11]	KFA[65]	mRVM[132]	SELM[111]	DLM[80]
Iris	96.67	96.28	95.71	97.33	97.33	97.48	93.87	**98.00**	89.59
Wine	**99.44**	98.48	97.76	**99.44**	98.89	99.38	96.24	99.41	96.98
Seeds	**99.05**	-	-	-	-	-	-	-	-
CoverType	**74.37**	-	-	-	-	-	-	-	-
Glass	86.45	68.41	75.42	73.83	-	70.67	67.49	**95.26**	51.05
PenDigits	**77.60**	-	-	-	-	-	-	-	-
Vowels	69.40	58.66	-	99.05	63.55	**99.19**	-	95.35	-
Segment	96.62	96.53	-	**97.68**	96.10	97.22	-	97.40	76.66
Vehicle	**88.89**	84.37	-	87.47	-	86.16	76.30	85.17	43.41
SATimage	91.70	92.35	-	92.35	**96.88**	91.60	-	90.00	75.17

ANN, DT and NB were set to the most suitable values for these classifiers using the Weka software [77]. The classification results are compared in Table 6.5 with the best performing accuracy for each test dataset marked in bold.

Table 6.5 Comparison with the state-of-the-art classifiers

Dataset	Testing Accuracy (%)				
	McDCA	SVM-SMO	DT	ANN	NB
Iris	96.67	96.00	96.00	**97.33**	96.00
Wine	**99.44**	98.10	93.82	97.19	96.63
Seeds	**99.05**	93.81	91.91	95.24	91.42
CoverType	74.37	69.58	**92.33**	71.53	63.05
Glass	**86.45**	57.48	65.89	67.29	49.53
PenDigits	77.60	**94.94**	92.05	92.30	82.13
Vowels	69.40	63.25	78.79	**83.14**	65.90
Segment	**96.62**	86.95	91.00	91.67	80.0
Vehicle	**88.89**	74.46	72.57	82.62	44.80
SATimage	**91.70**	86.56	86.07	88.77	79.50

The results indicate that, for the most datasets, the performances generated by McDCA are notably better than those from the state-of-the-art classifiers. However, the McDCA performs a bit worse than the rest of classifiers on Covertype, Pendigits and vowels datasets while slightly outperformed by ANN on the Iris dataset. These encouraging results further proves that, the McDCA is able to produce competitive results when applied to small and large size multiclass datasets.

Runtime comparison

As it was shown in Section 6.1.3 of this work, the McDCA has a quadratic worst case runtime complexity of $\mathcal{O}(n^2)$ and $\mathcal{O}(RPn^2)$ for testing and training respectively, which compares well with other multiclass approaches such as SVM, DT, ANN, NB, and AISLFS [43]. The training runtime complexities of SVM, DT, ANN, NB, and AISLFS are $\mathcal{O}(n^2 f)$ [21], $\mathcal{O}(nf \log n)$ [42], $\mathcal{O}(n^3)$ [99], $\mathcal{O}(nf)$ [99], and $\mathcal{O}(n^4)$ [43] respectively, with n the number of input data items and f is the dimensionality (features).

Also, in Figure 6.5, the testing time of McDCA (in seconds) for seven datasets were compared with three multiclass approaches from Table 6.4 including SVM-ELM [81], SELM

[11] and VVRKFA [65], since the testing time were pointed out in these studies. The seven datasets include Iris, Wine, Glass, Vowels, Segment, Vehicle, and SATimage. It can be observed that, for all datasets, McDCA takes comparable testing time to that of other multiclass approaches. Also, unlike SVM-ELM and SELM approaches, the McDCA is observed to use lower testing time on relative large dataset such as SATimage which compares well with VVRKFA approach. This is due to the fact that the McDCA requires relatively smaller time for testing after optimisation.

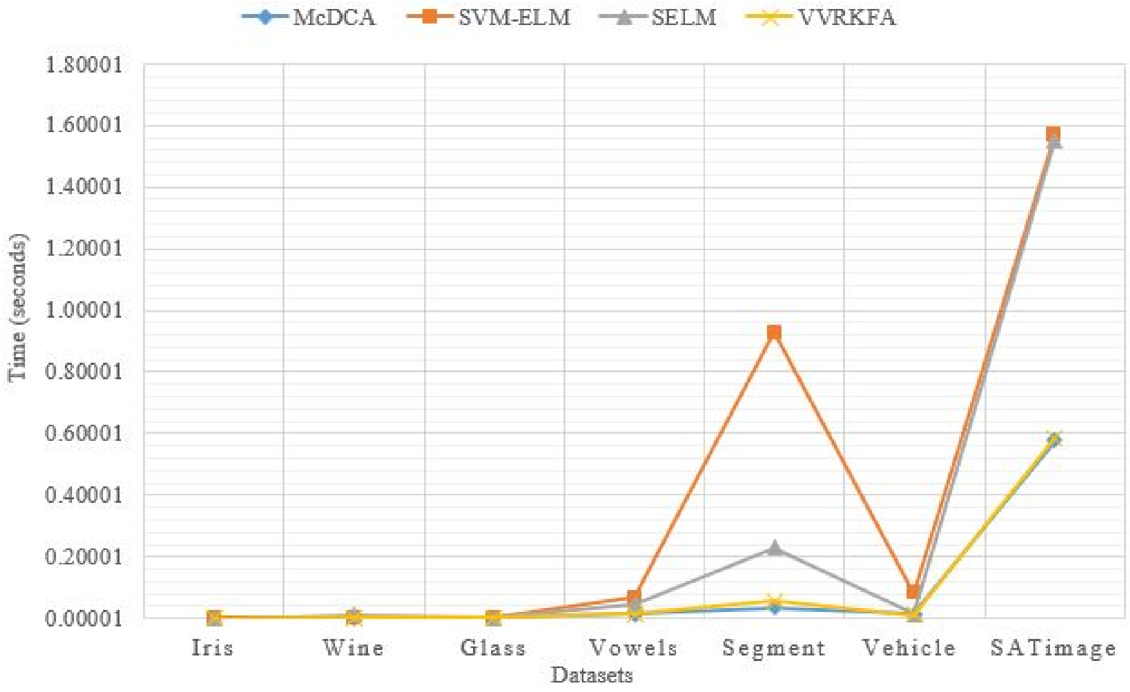
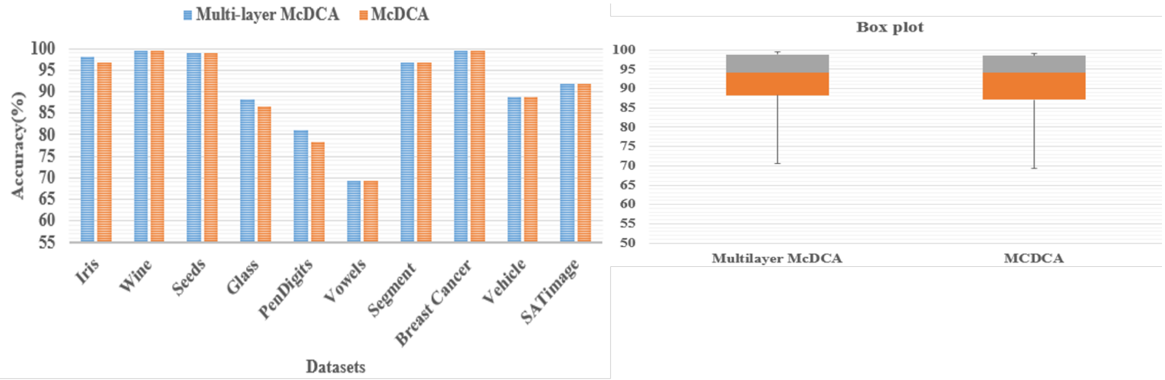


Figure 6.5 Running time (seconds) comparison with other approaches

Result on Multi-layer McDCA

The general experimental arrangement of multi-layer McDCA followed in this section is as it is presented in Figure 6.2. For simplicity, three (3) layers of DCs were employed in all experiments; with ten (10) DCs in each layer. Other parameters of the McDCA were kept unchanged. Nine (9) datasets from Table 6.1 were used to validate the performances of multi-layer McDCA while simultaneously comparing the results with that of the McDCA. The testing results obtained are presented in Figure 6.6 (a) with box plot analysis of the testing result between multilayer and single layer McDCA on Figure 6.6 (b).



(a)

(b)

Figure 6.6 Testing performance comparison for multi-layer McDCA

From Figure 6.6 (a), two observations can be drawn. Firstly, it can be noticed that, multi-layer McDCA performs considerably better on four datasets namely Iris, Glass, Pendigits and Breast cancer compared to the single layer McDCA. Secondly, the performances of multi-layer McDCA is similar to that of the single layer McDCA for the rest of datasets; which requires more investigation in the future studies of McDCA. Accordingly, it can be concluded that, multi-layer McDCA is more effective in learning the optimal weights for some datasets which significantly improves its classification performance compared to the single layer McDCA. Box plot analysis on Figure 6.6 (b) indicates that, multilayer McDCA is considerably better than the McDCA on all datasets used. In terms of running time, multilayer McDCA requires more training and testing time than the single layer McDCA to process data. However, availability of high performance and super computers can lessen processing timing tremendously nowadays.

6.2.4 E-Government Multi-Attack Detection

Evaluation of the proposed McDCA on multiples attacks detection in e-Government system is presented in this section. The performance were measured in terms of accuracy which is equal to the correctly classified data instances including the normal, and average detection rate which was computed from true positive rate of each attack type (excluding the normal). The datasets used for evaluation are KDD99 and UNSW_NB15. The results obtained are presented in Table 6.6.

From Table 6.6, both datasets have produced effective results on the classification accuracy (95.64% and 93.24%) and average attacks detection rate (93.05% and 90.82%). The McDCA was able to detect DoS/DDoS attacks from the KDD99 and UNSW_NB15 dataset

Table 6.6 E-Government multi-attacks detection using McDCA

Dataset	**Accuracy (%)**	**Avg. Detection Rate (%)**
KDD99	95.64	93.05
UNSW_NB15	93.24	90.82

at a detection rate of 98.62% and 97.70% respectively. DoS and DDoS are one of the most popular attack techniques that attackers can easily implement and use to disrupt and destroy networked systems such as e-Government. Hence, the effective classification performance of McDCA proves that, it can detect different kind of attacks targeting e-Government systems and thus ensuring privacy and integrity of data stored in the blockchain database. Therefore, in the proposed framework, McDCA can function as an entrance to the e-Government system in order to identify individual attack associated with user's transactions. Consequently, when user submit a transaction to the system, firstly the McDCA will inspect the traffic for any attack types before being accepted and recorded to the blockchain database.

6.3 Discussion

The experiments have proved that, McDCA is able to generate competitive classification performance with some of the existing multiclass classification techniques. The competitiveness of McDCA is demonstrated both in terms of prediction accuracy and computational efficiency. The better performances of the McDCA is attributed by the fact that it can generate the best set of optimal weights associate with features without categorising them into three signal as require by the original DCA. This help to facilitate its generalisation ability and thus being successfully applied to different multiclass classification problems. Moreover, the McDCA potentially eliminates the time, impression and complexity required to map the features to signal categories, which often leads to performance improvement.

Muti-attack detection rate results obtained from the two cybersecurity datasets have proved that, McDCA can detect multiple attacks in the proposed e-Government system. Note that, the core challenge when developing a multi-attack detection system is to be able to train the classifier well to differentiate between different kind of attacks and normal data present in the dataset. Since the McDCA has produced better detection performance, it can function as an entrance to the e-Government system in order to identify individual attack associated with user's transactions before being recorded to the blockchain database. So, once the McDCA IDS is integrated in the proposed framework presented in chapter 3, it will mitigates the current cybersecurity attacks and emerging security threats targeting e-Government systems; thus improves privacy, integrity, confidentiality and availability of the sensitive information processed, shared and stored, processed with public sector.

The runtime complexity analysis of the McDCA for training and testing were found to be $\mathcal{O}(RPn^2)$ and $\mathcal{O}(n^2)$ respectively. Generally, in many practical application, the time required for testing is of far much importance than the time required for training [79]. It is clear that, the McDCA requires more runtime to optimise the parameters of the weighted function using GA during the training process but after weights optimisation, the testing runtime required is reasonable since the McDCA goes through a single iteration to process and classify the testing data.

6.4 Summary

This chapter proposed a multi-attack detection system for e-Government system by transforming the computing metaphor of the human immune system, i.e., the binary classifier DCA, to support multi-class classification. The McDCA was implemented by generalising the natural behaviors of DCs to allow multiple situations to be considered rather than simply normal and anomaly. To further prove the potential of the proposed McDCA, a multilayer McDCA is also proposed and simply implemented by allowing the use of DCs in a layered structure; this ultimately opens the door for its further extension to be implemented as a deep learning approach. The experimental results based on the simple implementation of the proposed McDCA and multilayer McDCA demonstrated the working and efficacy of the system, with overall better performance than those from the commonly used and recently proposed conventional multi-class classifiers. The results obtained from two cybersecurity datasets prove that, McDCA is able to simultaneously detect multiples attacks targeting e-Government system. Hence, McDCA can function as an entrance to e-Government system in order to identify individual attack associated with user's transactions. So, when user submit a transaction to the e-Government system, McDCA will inspect the traffic for any attack before being accepted and recorded by e-Government devices and servers.

Chapter 7

Conclusion

This chapter concludes the thesis and points out the possible future developments. The summary of the thesis as detailed in the previous chapters is presented. After reviewing the privacy and security issues in the exisiting e-Government systems in the literature, this PhD work has developed a decentralised secure and privacy-preserving e-Government system by innovatively using blockchain technology. Blockchain technology enables the implementation of highly secure and privacy-preserving decentralised applications where information is not under the control of any centralised third party. The developed secure and decentralised e-Government system is based on the consortium type of blockchain technology, which is a semi-public and decentralised blockchain system consisting of a group of pre-selected entities or organisations in charge of consensus and decisions making for the benefit of the whole network of peers. Ethereum blockchain solution was used in this project to simulate and validate the proposed system since it is open source and supports off-chain data storage such as images, PDFs, DOCs, contracts, and other files that are too large to be stored in the blockchain or that are required to be deleted or changed in the future, which are essential part of e-Government systems.

This PhD work also has developed an IDS based on the enhanced DCA for detecting unwanted internal and external traffics to support the proposed blockchain-based e-Government system, because the blockchain database is append-only and immutable. The IDS effectively prevent unwanted transactions such as virus, malware or spyware from being added to the blockchain-based e-Government network. Three significant improvements have been implemented for DCA-based IDS. Firstly, a new parameters optimisation approach for the DCA is implemented by using the GA. Secondly, fuzzy inference systems approach is developed to solve nonlinear relationship that exist between features during the pre-processing stage of the DCA so as to further enhance its anomaly detection performance in e-Government systems.

In addition, a multiclass DCA capable of detection multiple attacks is developed in this project, given that the original DCA is a binary classifier and many practical classification problems including computer network intrusion detection datasets are often associated with multiple classes.

The effectiveness of the proposed approaches in enforcing security and privacy in e-Government systems were demonstrated through three real-world applications: privacy and integrity protection of information in e-Government systems, internal threats detection, and external threats detection. Privacy and integrity protection of information in the proposed e-Government systems is provided by using encryption and validation mechanism offered by the blockchain technology. Experiments demonstrated the performance of the proposed system, and thus its suitability in enhancing security and privacy of information in e-Government systems. The applicability and performance of the DCA-based IDS in e-Government systems were examined by using publicly accessible insider and external threat datasets with real-world attacks. The results show that, the proposed system can mitigate insider and external threats in e-Government systems whilst simultaneously preserving information security and privacy. The proposed system also could potentially increase the trust and accountability of public sectors due to the transparency and efficiency that are offered by the blockchain applications. Although promising, further research is needed to enhance the proposed e-Government system. Therefore, short-term and long-term developments are also presented in this chapter.

7.1 Summary of Thesis

E-Government employs ICTs to offer public services to citizens, employees and other shareholders. Due to its complex nature and sensitive information it stores, e-Government requires to be secured, privacy-preserved and decentralised. The existing e-Government systems such as websites and eIDs are faced with the potential privacy issues, security vulnerabilities and suffer from a single point of failure due to centralised databases and servers. Centralised management and validation system always presents a single point of failure and make the system a target to cyber attacks such as malware, ramsonware, DDoS, DoS, and etc. Recently, blockchain technology has appeared to be one of the core technologies for secure data sharing and storage over trustless and decentralised systems. Therefore, the main goals of this thesis was to develop a decentralised secure and privacy-preserving e-Government system by using the blockchain technology. Then, develop an IDS based on the DCA algorithm for detecting unwanted internal and external traffics to support the proposed blockchain-based e-Government system, because the blockchain database is append-only and immutable. The DCA-based IDS effectively prevent unwanted transactions such as virus, malware or spyware from being added to the blockchain-based e-Government network.

Before presenting the proposed e-Government system and IDS approaches, a detailed literature review of the e-Government systems, blockchain technology, intrusion detection systems and artificial immune systems has been presented in Chapter 2. In particular, privacy and security issues in e-Government systems are presented. Also, public-key cryptography, symmetric-key cryptography, digital signature and cryptographic hash function which are essential components for the implementation of the blockchain technology are reviewed. Additionally, types of blockchain technology and their advantages and limitations are provided. Finally, the background of the DCA algorithm which is a class of AIS used in this thesis to develop an IDS is presented.

Chapter 3 proposed a decentralised e-Government framework with privacy preservation, and insider and external threat detection functionality, using blockchain technology and the DCA algorithm. The proposed e-Government framework is comprised of three main modules. Firstly, a decentralised e-Government module is comprised of a P2P network with each node representing a public department based on the blockchain technology. Secondly, an external attack detection module based on the DCA detects unexpected traffics coming from the Internet to the e-Government system for further investigation by the network administrator. Thirdly, an insider threat detection module based on the DCA identifies internal anomalies from legitimated accounts of the e-Government system for further investigation. The theoretical and qualitative analysis on security and privacy of the proposed framework shows that, encryption, immutability and the decentralised management and control offered by the blockchain technology can provide the required security and privacy in e-Government systems. Insider and external threats associated with the blockchain transactions from users are detected and reported by the DCA-based IDS to avoid any invalid operations to the blockchain database. Thus, it can be applied in Government organisations to implement a decentralised and secure e-Government systems to overcome design challenges such as interoperability, integration and complexity. Additionally, this framework has the potential to increase citizens' trust in the public sectors.

Chapter 4 proposed a decentralised e-Government system based on consortium blockchain for effective and secure information sharing. The proposed system was simulated and evaluated by using the eVIBES simulator since it is open source and supports off-chain (sideDB) data storage such as images, PDFs, DOCs, contracts, and etc. The performance evaluation based on the number of transactions processed per second and on the time for processing a single transaction by varying the number of nodes (validators) in the consortium blockchain network have proved that, the proposed system is suitable for security and privacy assurance in e-Government systems. Additionally, the proposed system can offer advantages such as high scalability, high transaction speed, high data integrity, and high collaboration, low risk of cybersecurity attack, low energy consumption, low transaction cost and anonymity; while simultaneously ensuring the required level of trust in e-Government systems.

Note that, blockchain database is append-only and immutable; so, once the information is added cannot be deleted or changed in the future. Unwanted traffics such as spyware, worms, ransomware, and etc, should be prevented from being added to the proposed blockchain-based e-Government network. Therefore, Chapter 5 has proposed three different cybersecurity attacks detection systems based on enhanced DCA for identifying and mitigating unwanted traffics in e-Government systems. Firstly, a new parameters optimisation approach for the DCA was implemented by using GA; since the original DCA uses manual method to pre-defined the weights for its objective function. Secondly, fuzzy inference systems approach was used to developed an approach which can solve nonlinear relationship that may exist between input features and the resultant three DCA's signals during its pre-processing stage. Thirdly, a new signal categorisation method for the DCA was proposed based on Partial Shuffle Mutation of GA to automatically categorise the input features into the three DCA's signal categories; given that the original DCA uses manual categorisation technique based on domain or expert knowledge of the domain. The experimental results show that the enhanced DCA approaches are capable of detecting cybersecurity attacks in e-Government system with better performances while simultaneously ensuring privacy to blockchain transactions. In particular, the optimised GA-DCA as produced much better performance and hence can be adopted in the proposed e-Government system to address any network traffic and transaction based malicious attacks.

Chapter 6 proposed the multi-attack detection system for e-Government system by transforming the computing metaphor of the human immune system, i.e., the binary classifier DCA, to support multi-class classification. The McDCA was implemented by generalising the natural behaviors of DCs to allow multiple situations to be considered rather than simply normal and anomaly. To further prove the potential of the proposed McDCA, a multilayer McDCA is also proposed and simply implemented by allowing the use of DCs in a layered structure; this ultimately opens the door for its further extension to be implemented as a deep learning approach. The experimental results based on the simple implementation of the proposed McDCA and multilayer McDCA demonstrated the working and efficacy of the system, with overall better performance than those from the commonly used and recently proposed conventional multi-class classifiers. The results obtained from cybersecurity datasets of KDD99 and UNSW_NB15 prove that, McDCA is able to simultaneously detect multiples attacks targeting e-Government system. Hence, McDCA can function as an entrance to e-Government system in order to identify individual attack associated with user's transactions. So, when user submit a transaction to the e-Government system, McDCA will inspect the traffic for any attack before being accepted and recorded by e-Government devices and servers.

In summary,two main contributions have been achieved during the development of this PhD research work: 1) the development of a decentralised secure and privacy-preserving e-Government system which enforces information security and privacy in the public sectors, and, 2) the development of a DCA-based IDS for detecting and mitigating cyber attacks and anomalies targeting the proposed blockchain-based e-Government framework. Experimental evaluation and validation of both approaches have demonstrated the potential of the proposed system in mitigating threats in e-Government systems while simultaneously preserving privacy and security of information. Nonetheless,further research is also required to enhance the proposed approached for enahancing privacy and security in e-Government system.

7.2 Future Works

Although the results show that the proposed e-Government system could potentially eliminate privacy issues and security vulnerabilities facing the existing e-Government systems, much can be done to further improve the work presented in this thesis. The following subsections present the short-term and long-term developments that can be done to further improve the proposed e-Government system.

7.2.1 Short-term Developments

This subsection explores further developments that could be readily realised with the current PhD project on a more robust foundation in a short-term plan.

Blockchain's Block Creation Time

The performance evaluation of the proposed blockchain-based network based on the number of transaction processed per second, time for processing a single transaction and block creation time (sec) have proved that, the proposed system is suitable for security and privacy assurance in e-Government systems. It is desirable to investigate how machine learning and other artificial intelligence techniques can be used to increase the block creation time when there is a spike in the number of transactions in e-Government network so as to make the system more scalable and robust. This is because AI algorithms can be able to process incredibly large amounts of data and variables while making best decisions [166, 69].

Improve the Analysis and Assignment Phase of the Enhanced DCA

The context assignment phase of the DCA is performed by comparing the signal concentration values between *mDC* (abnormality) and *smDC* (normality) contexts. The most controversial question about the DCA is the existence of crisp separation between semi-mature and mature cumulative context values. The context regarding the collected data item will be hard to be separated if the difference between the two contexts is small, which negatively affects the classification accuracy. To address this challenge, fuzzy-rough set theory (FRST-DCA) context assignment approach was proposed in [28]. However, the FRST–DCA is an expensive solution due to information loss during the discretisation process; also, it is only practically applicable to simple datasets thanks to its computational complexity [29, 71]. Recently, K-Means clustering algorithm [54] was used to address this problem but nonetheless, it is computationally an expensive solution [54]. Therefore, it is desired to investigate how other methods such as Linear Discriminant Analysis (LDA) [166], PCA [166] and logistic regression [69] can solve this limitation for the DCA so that it can detect attacks in e-Government system with much higher accuracy.

DCA-based IDS Parameters Optimisation by Using Other Techniques

The proposed DCA-based IDS was optimised by using GA due to its ability to achieve the proper balance between exploitation and exploration of search space simply by setting well relatively fewer number of adjustable parameters. GA has proved to be effective for the DCA parameters optimisation even when it was compared with the particles swarm optimisation [53, 46]. However, the performance of GA could be further compared with gradient descent optimisation algorithms [166] and other meta-heuristic optimisation algorithms such as simulated annealing, artificial bee colony optimisation, ant colony optimisation, to fully investigate the potential of optimisation algorithms for the optimised DCA.

Multiclass DCA improvement

Given that the information aggregation in McDCA was simply implemented using a weighted summation, the application of other more complex aggregation approaches need to be investigated. Also, the parameters were determined using the popular GA, and other optimisation approaches may generate more optimal solutions. In addition, the DC pool size, the migration threshold were all fixed and empirically determined based on limited trials in this work; thus further investigation is required. What is more, study on multi-layer McDCA is in demand by examining the number of layers, the number of DCs, and alternatively information processing approaches in each DC; interestingly, the stack of many layers may turn the McDCA as a deep learning approach for more challenging problems. Last but not least, theoretical analysis from mathematical perspective will support the full comprehension of not only the proposed McDCA, but also its biological counterpart.

7.2.2 Longer-term Developments

This subsection proposes several possible future research works which could be achieved in long-term plan.

Blockchain-based Collaborative Intrusion Detection

A single IDS in a network such as e-Government system can have false positives (false attacks) or false negatives (missed attacks) [4, 170]. Therefore, it is very crucial to develop a Collaborative IDS (CIDS) that can aggregate data from multiple blockchain nodes or peers in order to make informed decisions about potential intrusions or anomalies in e-Government system. Note that, CIDS can be able to counter a variety of attacks, especially large-scale coordinated attacks [4, 114]. Thus, it would be ideal if the blockchain technology will be exploited to develop a CIDS for the proposed e-Government system. CIDS in the proposed system will replace the use of the DCA-based IDS which is computationally expensive due to the time required to train its parameters. Additionally, using the blockchain-based CIDS will be the best solution as it will ensure that the proposed system is only made up of blockchain technology components. Lastly, a carefully designed and configured CIDS can increase the detection accuracy compared to a single IDS, without a substantial degradation in performance [4].

Integrating the Proposed e-Government framework with Smart Cities

Smart Cities are now a reality. The United Nations Population Fund indicated that about 3.3 billion (54% of world population) lived in urban areas in 2014. This number will increase to about 5 billion (about 66%) by 2030 [157]. This dramatic growth in urban populations is generating a significant increase in the number of interconnected devices and subsequently e-Government is required to be at the centre of Smart Cities expansion. Smart Cities present new economic opportunities to governments but with technological developments there will be security and privacy threats. Smart cities will require large scale networks to accommodate these diverse devices. All cities are managed by government authorities and they will be responsible for delivering e-Government services with strict governance policies. Therefore, it would be ideal if the smart cities and IoTs are integrated in the proposed e-Government framework.

Efficient Blockchain Consensus Algorithm for e-Government system

To validate transactions, nodes in the blockchain network run a consensus algorithm such as PoW, PoS, DPoS, PoD, etc. This thesis has exploited DPoS since it saves energy that is required in PoW to solve mathematical puzzle. However, when there is a spike in number of transactions in the blockchain network, there is a trade-off between security and performance in designing DPoS mechanisms [23, 178]. It would be interesting to develop a more efficient and effective blockchain consensus algorithm for the proposed blockchain-based e-Government system. In particular, the Proof-Of-Authority (PoA) consensus method [139] could be exploited to implement a more effective consensus algorithm that can give a small and designated number of blockchain nodes the power to validate transactions or interactions with the network and to update its distributed database.

Appendix A

Publications

Publications arising from this work

A number of publications have been generated from this PhD work. Below is the list of all publications in chronological order, including published, accepted or submitted papers for publication.

1. Elisa, N., Yang, L., Chao, F., and Naik, N. (2020). Comparative study of genetic algorithm and particle swarm optimisation for dendritic cell algorithm. In 2020 IEEE Congress on Evolutionary Computation (CEC), pages 1–8. IEEE.

2. Elisa, N., Yang, L., Greensmith, J., and Chao, F. Multi-class dendritic cell algorithm (McDCA). Natural Computing, **under review**.

3. Elisa, N., Chao, F., and Yang, L. (2019). A study of the necessity of signal categorisation in dendritic cell algorithm. In UK Workshop on Computational Intelligence, pages 210–222. Springer.

4. Elisa, N., Yang, L., and Chao, F. (2019). Signal categorisation for dendritic cell algorithm using ga with partial shuffle mutation. In UK Workshop on Computational Intelligence, pages 529–540. Springer.

5. Elisa, N., Yang, L., Fu, X., and Naik, N. (2019). Dendritic cell algorithm enhancement using fuzzy inference system for network intrusion detection. In 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pages 1–6. IEEE.

6. Elisa, N., Yang, L., Li, H., Chao, F., and Naik, N. (2019). Consortium blockchain for security and privacy-preserving in e-government systems. In ICEB 2019 Proceedings, pages 99–107. ICBE.

7. Yang, L., Elisa, N., and Eliot, N. (2019). Privacy and security aspects of e-government in smart cities. In Smart cities cybersecurity and privacy, pages 89–102. Elsevier.

8. Yang, L., Li, J., Elisa, N., Prickett, T., and Chao, F. (2019). Towards big data governance in cybersecurity. Data-Enabled Discovery and Applications, 3(1):10.

9. Elisa, N., Li, J., Zuo, Z., and Yang, L. (2018). Dendritic cell algorithm with fuzzy inference system for input signal generation. In UK workshop on computational intelligence, pages 203–214. Springer. (BestPaperAward)

10. Elisa, N., Yang, L., Chao, F., and Cao, Y. (2018). A framework of blockchain-based secure and privacy-preserving e-government system. Wireless Networks, pages 1–11.

11. Elisa, N., Yang, L., and Naik, N. (2018). Dendritic cell algorithm with optimised parameters using genetic algorithm. In 2018 IEEE Congress on Evolutionary Computation (CEC), pages 1–8. IEEE.

12. Elisa, N., Yang, L., Qu, Y., and Chao, F. (2018). A revised dendritic cell algorithm using k-means clustering. In 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pages 1547–1554. IEEE.

Appendix B

Acronyms

AI	Artificial Intelligence
AIDS	Anomaly-based Intrusion Detection
AIS	Artificial Immune Systems
ANN	Artificial Neural Network
APCs	Antigen Presenting Cells
BA	Byzantine Agreement
CIDS	Collaborative Intrusion Detection Systems
CSM	Costimulatory Molecules
DCA	Dendritic Cell Algorithm
DCs	Dendritic Cells
DDoS	Distributed Denial of Service Attacks
DoS	Denial of Service Attacks

DPoS	Delegated Proof of Stake
DS	Danger Signals
DT	Decision Trees
eIDs	Electronic Identity Management Systems
FIS	Fuzzy Inference Systems
FN	False Negative
FP	False Positive
FRST	Fuzzy Rough Set Theory
G2B	Government to Business
G2C	Government to Citizens
G2E	Government to Employees
G2G	Government to Government
GA	Genetic Algorithm
HIS	Human Immune System
ICTs	Information and Communication Technologies
iDCs	Immature DCs
IDS	Intrusion Detection Systems

IG	Information Gain
IoTs	Internet of Things
MCAV	Mature Context Antigen Value
McDCA	Multi-class DCA
mDC	Mature DCs
MIDS	Misuse-based Intrusion Detection
NB	Naïve Bayes
NIDS	Network Intrusion Detection Systems
NSA	Negative Selection Algorithm
P2P	Peer to Peer
PAMP	Pathogenic Associated Molecular Patterns
PCA	Principal Component Analysis
PKI	Public Key Infrastructure
PoA	Proof of Authority
PoD	Proof of Difficulty
PoS	Proof of Stake
PoW	Proof of Work

PSA	Positive Selection Algorithm
PSM	Partial Shuffle Mutation
PSO	Particle Swarm Optimisation
RFT	Random Forrest Trees
SHA	Secure Hash Algorithm
smDC	Semi-mature DCs
SQL	Structured Query Language
SS	Safe Signals
SVM	Support Vector Machines
TN	True Negative
TP	True Positive
TSK	Takagi–Sugeno–Kang
XSS	Cross-site Scripting

Bibliography

[1] Abdoun, O., Tajani, C., and Abouchabaka, J. (2012). Analyzing the performance of mutation operators to solve the traveling salesman problem. *Int. J. Emerg. Sci*, 2(1):61–77.

[2] Aickelin, U., Bentley, P., Cayzer, S., Kim, J., and McLeod, J. (2003). Danger theory: The link between ais and ids? *Artificial immune systems*, pages 147–155.

[3] Aickelin, U. and Cayzer, S. (2008). The danger theory and its application to artificial immune systems. *arXiv preprint arXiv:0801.3549*.

[4] Alexopoulos, N., Vasilomanolakis, E., Ivánkó, N. R., and Mühlhäuser, M. (2017). Towards blockchain-based collaborative intrusion detection systems. In *International Conference on Critical Information Infrastructures Security*, pages 107–118. Springer.

[5] Alshehri, M. and Drew, S. (2010a). E-government fundamentals. In *IADIS International Conference ICT, Society and Human Beings*, pages 35–42.

[6] Alshehri, M. and Drew, S. (2010b). Implementation of e-government: advantages and challenges. In *International Association for Scientific Knowledge (IASK)*, pages 79–86.

[7] Antonopoulos, A. M. (2014). *Mastering Bitcoin: unlocking digital cryptocurrencies*. " O'Reilly Media, Inc.".

[8] Anttiroiko, A.-V. (2008). *Electronic Government: Concepts, Methodologies, Tools, and Applications*, volume 3. IGI Global.

[9] Arroub, A., Zahi, B., Sabir, E., and Sadik, M. (2016). A literature review on smart cities: Paradigms, opportunities and open problems. In *Wireless Networks and Mobile Communications (WINCOM), 2016 International Conference on*, pages 180–186. IEEE.

[10] Bace, R. G., Mell, P., et al. (2001). Intrusion detection systems. Technical report.

[11] Bai, Z., Huang, G.-B., Wang, D., Wang, H., and Westover, M. B. (2014). Sparse extreme learning machine for classification. *IEEE transactions on cybernetics*, 44(10):1858–1870.

[12] Baker, J. E. (1985). Adaptive selection methods for genetic algorithms. In *Proceedings of an International Conference on Genetic Algorithms and their applications*, pages 101–111. Hillsdale, New Jersey.

[13] Banchereau, J. and Steinman, R. M. (1998). Dendritic cells and the control of immunity. *Nature*, 392(6673):245–252.

[14] Bélanger, F. and Carter, L. (2008). Trust and risk in e-government adoption. *The Journal of Strategic Information Systems*, 17(2):165–176.

[15] Biswas, K. and Muthukkumarasamy, V. (2016). Securing smart cities using blockchain technology. In *2016 IEEE 18th international conference on high performance computing and communications; IEEE 14th international conference on smart city; IEEE 2nd international conference on data science and systems (HPCC/SmartCity/DSS)*, pages 1392–1393. IEEE.

[16] Blockchain in Argentina (2019). Blockchain Project in Argentina. "https://www.bloomberg.com/press-releases/2019-08-26/nec-idb-lab-and-ngo-bitcoin-argentina-to-deploy-a-blockchain-ba/. Accessed: 2020-05-22.

[17] Blockchain in Canada (2018). Blockchain Project in Canada. "https://bitaccess.ca/blog/government-of-canada-ipfs///. Accessed: 2020-03-22.

[18] Blockchain in Mexico (2017). Blockchain Project in Mexico. "https://www.gob.mx/cidge/acciones-y-programas/blockchain-hackmx/. Accessed: 2020-05-22.

[19] Blockchain in USA (2016). Blockchain Project in USA. "https://consensys.net/blog/enterprise-blockchain/which-governments-are-using-blockchain-right-now//. Accessed: 2020-05-27.

[20] Buczak, A. L. and Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2):1153–1176.

[21] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

[22] Burnet, F. M. (1961). Immunological recognition of self. *Science*, 133(3449):307–311.

[23] Buterin, V. et al. (2014). A next-generation smart contract and decentralized application platform. *white paper*, 3:1–37.

[24] Cachin, C. (2016). Architecture of the hyperledger blockchain fabric. In *Workshop on distributed cryptocurrencies and consensus ledgers*, volume 310, pages 1–4.

[25] Carter, L. and Weerakkody, V. (2008). E-government adoption: A cultural comparison. *Information systems frontiers*, 10(4):473–482.

[26] Chelly, Z. and Elouedi, Z. (2013a). A new data pre-processing approach for the dendritic cellalgorithm based on fuzzy rough set theory. In *Proceedings of the 15th annual conference companion on Genetic and evolutionary computation*, pages 163–164.

[27] Chelly, Z. and Elouedi, Z. (2013b). Qr-dca: A new rough data pre-processing approach for the dendritic cell algorithm. In *International Conference on Adaptive and Natural Computing Algorithms*, pages 140–150. Springer.

[28] Chelly, Z. and Elouedi, Z. (2015). Hybridization schemes of the fuzzy dendritic cell immune binary classifier based on different fuzzy clustering techniques. *New Generation Computing*, 33(1):1–31.

[29] Chelly, Z. and Elouedi, Z. (2016). A survey of the dendritic cell algorithm. *Knowledge and Information Systems*, 48(3):505–535.

[30] China Cyber-Attacks (2017). Most cyber attacks on China in 2017. "http://www.xinhuanet.com/english/2019-06/10/c_138131614.htm/. Accessed: 2020-02-14.

[31] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to algorithms*. MIT press.

[32] Crosby, M., Pattanayak, P., Verma, S., Kalyanaraman, V., et al. (2016). Blockchain technology: Beyond bitcoin. *Applied Innovation*, 2(6-10):71.

[33] Curtis, B., Krasner, H., and Iscoe, N. (1988). A field study of the software design process for large systems. *Communications of the ACM*, 31(11):1268–1287.

[34] DasGupta, D. (1993). An overview of artificial immune systems and their applications. In *Artificial immune systems and their applications*, pages 3–21. Springer.

[35] Dasgupta, D., Yu, S., and Nino, F. (2011). Recent advances in artificial immune systems: models and applications. *Applied Soft Computing*, 11(2):1574–1587.

[36] Deshpande, A., Nasirifard, P., and Jacobsen, H.-A. (2018). evibes: Configurable and interactive ethereum blockchain simulation framework. In *Proceedings of the 19th International Middleware Conference (Posters)*, pages 11–12.

[37] Dib, O., Brousmiche, K.-L., Durand, A., Thea, E., and Hamida, E. B. (2018). Consortium blockchains: Overview, applications and challenges. *International Journal On Advances in Telecommunications*, 11(1&2):51–64.

[38] Dib, O., Huyart, C., and Toumi, K. (2020). A novel data exploitation framework based on blockchain. *Pervasive and Mobile Computing*, 61:1–34.

[39] Dorri, A., Kanhere, S. S., Jurdak, R., and Gauravaram, P. (2017). Blockchain for iot security and privacy: The case study of a smart home. In *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*, pages 618–623. IEEE.

[40] Dua, D. and Graff, C. (1998). UCI machine learning repository.

[41] Dubai, S. (2016). Dubai blockchain strategy. *Smart Dubai, Dubai Government, Dec.*

[42] Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

[43] Dudek, G. (2012). An artificial immune system for classification with local feature selection. *IEEE Transactions on Evolutionary Computation*, 16(6):847–860.

[44] Elisa, N., Chao, F., and Yang, L. (2019a). A study of the necessity of signal categorisation in dendritic cell algorithm. In *UK Workshop on Computational Intelligence*, pages 210–222. Springer.

[45] Elisa, N., Li, J., Zuo, Z., and Yang, L. (2018a). Dendritic cell algorithm with fuzzy inference system for input signal generation. In *UK workshop on computational intelligence*, pages 203–214. Springer.

[46] Elisa, N., Yang, L., Cha0, F., and Naik, N. (2020a). Comparative study of genetic algorithm and particle swarm optimisation for dendritic cell algorithm. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE.

[47] Elisa, N., Yang, L., and Chao, F. (2019b). Signal categorisation for dendritic cell algorithm using ga with partial shuffle mutation. In *UK Workshop on Computational Intelligence*, pages 529–540. Springer.

[48] Elisa, N., Yang, L., and Chao, F. (2020b). A decentralised secure and privacy-preserving e-government framework. *Journal of Ambient Intelligence and Smart Environments*, pages 1–13.

[49] Elisa, N., Yang, L., Chao, F., and Cao, Y. (2018b). A framework of blockchain-based secure and privacy-preserving e-government system. *Wireless Networks*, pages 1–11.

[50] Elisa, N., Yang, L., Fu, X., and Naik, N. (2019c). Dendritic cell algorithm enhancement using fuzzy inference system for network intrusion detection. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE.

[51] Elisa, N., Yang, L., Greensmith, J., and Chao, F. (2020c). Multi-class dendritic cell algorithm (mcdca). *Natural Computing*, pages 1–16.

[52] Elisa, N., Yang, L., Li, H., Chao, F., and Naik, N. (2019d). Consortium blockchain for security and privacy-preserving in e-government systems. In *ICEB 2019 Proceedings*, pages 99–107. ICBE.

[53] Elisa, N., Yang, L., and Naik, N. (2018c). Dendritic cell algorithm with optimised parameters using genetic algorithm. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE.

[54] Elisa, N., Yang, L., Qu, Y., and Chao, F. (2018d). A revised dendritic cell algorithm using k-means clustering. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 1547–1554. IEEE.

[55] Ellison, C. and Schneier, B. (2000). Ten risks of pki: What you're not being told about public key infrastructure. *Comput Secur J*, 16(1):1–7.

[56] Elmaghraby, A. S. and Losavio, M. M. (2014). Cyber security challenges in smart cities: Safety, security and privacy. *Journal of advanced research*, 5(4):491–497.

[57] Ethereum Cosmos (2020a). Ethermint Documentation. "https://docs.ethermint.zone//. Accessed: 2020-05-13.

[58] Ethereum Cosmos (2020b). Tendermint. "https://docs.tendermint.com///. Accessed: 2020-05-13.

[59] Evans, M., Maglaras, L. A., He, Y., and Janicke, H. (2016). Human behaviour as an aspect of cybersecurity assurance. *Security and Communication Networks*, 9(17):4667–4679.

[60] Fang, Z. (2002). E-government in digital era: concept, practice, and development. *International journal of the Computer, the Internet and management*, 10(2):1–22.

[61] Forouzan, B. A. (2007). *Cryptography & network security*. McGraw-Hill, Inc.

[62] Forrest, S., Javornik, B., Smith, R. E., and Perelson, A. S. (1993). Using genetic algorithms to explore pattern recognition in the immune system. *Evolutionary computation*, 1(3):191–211.

[63] Forrest, S., Perelson, A. S., Allen, L., and Cherukuri, R. (1994). Self-nonself discrimination in a computer. In *Proceedings of 1994 IEEE computer society symposium on research in security and privacy*, pages 202–212. Ieee.

[64] Gai, K., Wu, Y., Zhu, L., Qiu, M., and Shen, M. (2019). Privacy-preserving energy trading using consortium blockchain in smart grid. *IEEE Transactions on Industrial Informatics*, 15(6):3548–3558.

[65] Ghorai, S., Mukherjee, A., and Dutta, P. K. (2010). Discriminant analysis for fast multiclass data classification through regularized kernel function approximation. *IEEE transactions on neural networks*, 21(6):1020–1029.

[66] Gil-Garcia, J. R., Dawes, S. S., Pardo, T. A., et al. (2018). Digital government and public management research: finding the crossroads. *Public Management Review*, 20(5):633–646.

[67] Glasser, J. and Lindauer, B. (2013). Bridging the gap: A pragmatic approach to generating insider threat data. In *2013 IEEE Security and Privacy Workshops*, pages 98–104. IEEE.

[68] Goldberg, D. E. and Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine Learning*, 3(3):95–99.

[69] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

[70] Greensmith, J. (2007). *The dendritic cell algorithm*. PhD thesis, Citeseer.

[71] Greensmith, J. (2019). Migration threshold tuning in the deterministic dendritic cell algorithm. In Martín-Vide, C., Pond, G., and Vega-Rodríguez, M. A., editors, *Theory and Practice of Natural Computing*, pages 122–133, Cham. Springer International Publishing.

[72] Greensmith, J. and Aickelin, U. (2008). The deterministic dendritic cell algorithm. In *Artificial Immune Systems*, pages 291–302. Springer.

[73] Greensmith, J., Aickelin, U., and Cayzer, S. (2005). Introducing dendritic cells as a novel immune-inspired algorithm for anomaly detection. In *ICARIS*, volume 3627, pages 153–167. Springer.

[74] Greensmith, J., Aickelin, U., and Twycross, J. (2006). Articulation and clarification of the Dendritic Cell Algorithm. In *Proc. of the 5th International Conference on Artificial Immune Systems (ICARIS), LNCS 4163*, pages 404–417.

[75] Gu, F. (2011). *Theoretical and empirical extensions of the dendritic cell algorithm*. PhD thesis, University of Nottingham.

[76] Gu, F., Greensmith, J., and Aickelin, U. (2008). Further exploration of the dendritic cell algorithm: Antigen multiplier and time windows. In *International Conference on Artificial Immune Systems*, pages 142–153. Springer.

[77] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

[78] Holland, J. H. (1992). Genetic algorithms. *Scientific american*, 267(1):66–73.

[79] Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425.

[80] Huang, G., Liu, T., Yang, Y., Lin, Z., Song, S., and Wu, C. (2015). Discriminative clustering via extreme learning machine. *Neural Networks*, 70:1–8.

[81] Huang, G.-B., Zhou, H., Ding, X., and Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(2):513–529.

[82] Huang, Z. and Shen, Q. (2008). Fuzzy interpolation and extrapolation: A practical approach. *Fuzzy Systems, IEEE Transactions on*, 16(1):13–28.

[83] Huh, S., Cho, S., and Kim, S. (2017). Managing iot devices using blockchain platform. In *Advanced Communication Technology (ICACT), 2017 19th International Conference on*, pages 464–467. IEEE.

[84] Israel Cyber-Attacks (2019). Israel-Hamas Cyberwar, when old warfare meets new. "https://limacharlienews.com/mena/israel-hamas-cyberwar//. Accessed: 2020-02-14.

[85] Jain, A., Arora, S., Shukla, Y., Patil, T., and Sawant-Patil, S. (2018). Proof of stake with casper the friendly finality gadget protocol for fair validation consensus in ethereum. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 3(3):291–298.

[86] Jiménez, C. E., Falcone, F., Feng, J., Puyosa, H., Solanas, A., and González, F. (2012). e-government: Security threats. *e-Government*, 11:21–35.

[87] JISC (2017). UK Federation Information Centre. "https://www.ukfederation.org.uk/. Accessed: 2019-09-12.

[88] Joshi, J. B., Ghafoor, A., Aref, W. G., and Spafford, E. H. (2002). Security and privacy challenges of a digital government. In *Advances in Digital Government*, pages 121–136. Springer.

[89] Jun, M. (2018). Blockchain government-a next form of infrastructure for the twenty-first century. *Journal of Open Innovation: Technology, Market, and Complexity*, 4(1):1–12.

[90] Karokola, G. R. (2012). *A framework for securing e-government services: The case of tanzania*. PhD thesis, Department of Computer and Systems Sciences, Stockholm University.

[91] Katz, J. and Lindell, Y. (2014). *Introduction to modern cryptography*. CRC press.

[92] KDD99 dataset (1999). KDD Cup 1999 Data. "http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html/. Accessed: 2018-12-16.

[93] Kim, A., Oh, J., Ryu, J., and Lee, K. (2020). A review of insider threat detection approaches with iot perspective. *IEEE Access*, 8:78847–78867.

[94] Kim, J. and Bentley, P. J. (2001). An evaluation of negative selection in an artificial immune system for network intrusion detection. In *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, pages 1330–1337. Morgan Kaufmann Publishers Inc.

[95] Kim, J., Bentley, P. J., Aickelin, U., Greensmith, J., Tedesco, G., and Twycross, J. (2007). Immune system approaches to intrusion detection–a review. *Natural computing*, 6(4):413–466.

[96] Kóczy, L. and Hirota, K. (1993a). Approximate reasoning by linear rule interpolation and general approximation. *International Journal of Approximate Reasoning*, 9(3):197–225.

[97] Kóczy, L. T. and Hirota, K. (1993b). Interpolative reasoning with insufficient evidence in sparse fuzzy rule bases. *Information Sciences*, 71(1):169–201.

[98] Kosba, A., Miller, A., Shi, E., Wen, Z., and Papamanthou, C. (2016). Hawk: The blockchain model of cryptography and privacy-preserving smart contracts. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 839–858. IEEE.

[99] Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190.

[100] Kuperberg, M., Kemper, S., and Durak, C. (2019). Blockchain usage for government-issued electronic ids: A survey. In *International Conference on Advanced Information Systems Engineering*, pages 155–167. Springer.

[101] Lam, W. (2005). Barriers to e-government integration. *Journal of Enterprise Information Management*, 18(5):511–530.

[102] Lambrinoudakis, C., Gritzalis, S., Dridi, F., and Pernul, G. (2003). Security requirements for e-government services: a methodological approach for developing a common pki-based security policy. *Computer communications*, 26(16):1873–1883.

[103] Lau, H. Y. K. and Lee, N. M. Y. (2018). Danger theory or trained neural network - A comparative study for behavioural detection. In *Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS), Toyama, Japan, December 5-8*, volume 10.1109/SCIS-ISIS.2018.00143, pages 867–874.

[104] Layton, T. P. (2016). *Information Security: Design, implementation, measurement, and compliance*. Auerbach Publications.

[105] Li, J., Qu, Y., Shum, H. P. H., and Yang, L. (2017a). *TSK Inference with Sparse Rule Bases*, pages 107–123. Springer International Publishing, Cham.

[106] Li, J., Yang, L., Fu, X., Chao, F., and Qu, Y. (2017b). Dynamic qos solution for enterprise networks using tsk fuzzy interpolation. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

[107] Li, J., Yang, L., Qu, Y., and Sexton, G. (2018). An extended takagi–sugeno–kang inference system (TSK+) with fuzzy interpolation and its rule base generation. *Soft Computing*, 22(10):3155–3170.

[108] Li, J., Yang, L., Shum, H. P. H., Sexton, G., and Tan, Y. (2015). Intelligent home heating controller using fuzzy rule interpolation. In *UK Workshop on Computational Intelligence*, pages 303–314. Springer.

[109] Li, Z., Kang, J., Yu, R., Ye, D., Deng, Q., and Zhang, Y. (2017c). Consortium blockchain for secure energy trading in industrial internet of things. *IEEE transactions on industrial informatics*, 14(8):3690–3700.

[110] Luiijf, E. (2012). Understanding cyber threats and vulnerabilities. In *Critical Infrastructure Protection*, pages 52–67. Springer.

[111] Luo, J., Vong, C.-M., and Wong, P.-K. (2014). Sparse bayesian extreme learning machine for multi-classification. *IEEE Transactions on Neural Networks and Learning Systems*, 25(4):836–843.

[112] Matzinger, P. (2002). The danger model: a renewed sense of self. *Science*, 296(5566):301–305.

[113] Menezes, A. J., Katz, J., Van Oorschot, P. C., and Vanstone, S. A. (1996). *Handbook of applied cryptography*. CRC press.

[114] Meng, W., Tischhauser, E. W., Wang, Q., Wang, Y., and Han, J. (2018). When intrusion detection meets blockchain technology: a review. *Ieee Access*, 6:10179–10188.

[115] Moen, V., Klingsheim, A. N., Simonsen, K. I. F., and Hole, K. J. (2007). Vulnerabilities in e-governments. *International Journal of Electronic Security and Digital Forensics*, 1(1):89–100.

[116] Moon, M. J. (2002). The evolution of e-government among municipalities: rhetoric or reality? *Public administration review*, 62(4):424–433.

[117] Moustafa, N. and Slay, J. (2015). Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In *Military Communications and Information Systems Conference (MilCIS), 2015*, pages 1–6. IEEE.

[118] Mukhopadhyay, U., Skjellum, A., Hambolu, O., Oakley, J., Yu, L., and Brooks, R. (2016). A brief survey of cryptocurrency systems. In *Privacy, Security and Trust (PST), 2016 14th Annual Conference on*, pages 745–752. IEEE.

[119] Naik, N., Diao, R., and Shen, Q. (2014). Genetic algorithm-aided dynamic fuzzy rule interpolation. In *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*, pages 2198–2205.

[120] Naik, N., Diao, R., and Shen, Q. (2018). Dynamic fuzzy rule interpolation and its application to intrusion detection. *IEEE Transactions on Fuzzy Systems*, 26(4):1878–1892.

[121] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review*, pages 1–9.

[122] Ndou, V. (2004). E–government for developing countries: opportunities and challenges. *The electronic journal of information systems in developing countries*, 18(1):1–24.

[123] Ntulo, G. and Otike, J. (2013). E–government: Its role, importance and challenges. *School of Information Sciences. MoiUniversity*, pages 1–16.

[124] Oates, R., Kendall, G., and Garibaldi, J. M. (2008). Frequency analysis for dendritic cell population tuning. *Evolutionary Intelligence*, 1(2):145–157.

[125] Offensive Security (2019). Kali Linux | Penetration Testing and Ethical Hacking Linux Distribution. "https://www.kali.org/. Accessed: 2019-12-13.

[126] O'Leary, D. E. (2017). Configuring blockchain architectures for transaction information in blockchain consortiums: The case of accounting and supply chain systems. *Intelligent Systems in Accounting, Finance and Management*, 24(4):138–147.

[127] Palanisamy, R. and Mukerji, B. (2012). Security and privacy issues in e-government. In *E-Government Service Maturity and Development: Cultural, Organizational and Technological Perspectives*, pages 236–248. IGI Global.

[128] Pau, L.-F. (2010). Business and social evaluation of denial of service attacks of communications networks in view of scaling economic counter-measures. In *Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP*, pages 126–133. IEEE.

[129] Peterson, K., Deeduvanu, R., Kanjamala, P., and Boles, K. (2016). A blockchain-based approach to health information exchange networks. In *Proc. NIST Workshop Blockchain Healthcare*, volume 1, pages 1–10.

[130] Prentice, S. and Dewnarain, G. (2012). The future of the internet: Fundamental trends, scenarios and implications to heed. *Gartner, available at: http://www. gartner. com/technology/core/products/resea rch/topics/emergingTrendsTechnologies. jsp*, 14:111–125.

[131] Prokopios, D., Dimitris, G., Stefanos, G., Costas, L., and Mitrou, L. (2009). Towards an enhanced authentication framework for e-government services: The greek case. *Trust and Security*, pages 189–196.

[132] Psorakis, I., Damoulas, T., and Girolami, M. A. (2010). Multiclass relevance vector machines: sparsity and accuracy. *IEEE Transactions on neural networks*, 21(10):1588–1598.

[133] Ramya, U., Sindhuja, P., Atsaya, R., Dharani, B. B., and Golla, S. M. V. (2018). Reducing forgery in land registry system using blockchain technology. In *International Conference on Advanced Informatics for Computing Research*, pages 725–734. Springer.

[134] Reece, B. (2006). E-government literature review. *Journal of E-government*, 3(1):69–110.

[135] Santanelli, J. L. and de Lima Neto, F. B. (2016). Network intrusion detection using danger theory and genetic algorithms. In *International Conference on Intelligent Systems Design and Applications*, pages 394–405. Springer.

[136] SEC, S. (2000). *2: Recommended elliptic curve domain parameters*. Standards for Efficient Cryptography Group, Certicom Corp.

[137] Sharma, S. K. and Gupta, J. N. (2003). Building blocks of an e-government: A framework. *Journal of Electronic Commerce in Organizations (JECO)*, 1(4):34–48.

[138] Singapore Cyber-Attacks (2018). A key component for e-government security. "https://www.cryptomathic.com/news-events/blog/key-for-egovernment-security-central-signing-authentication/. Accessed: 2020-02-13.

[139] Singh, P. K., Singh, R., Nandi, S. K., and Nandi, S. (2019). Managing smart home appliances with proof of authority and blockchain. In *International Conference on Innovations for Community Services*, pages 221–232. Springer.

[140] Singh, S. and Karaulia, D. S. (2011). E-governance: Information security issues. In *International Conference on Computer Science and Information Technology (ICCSIT'2011) Pattaya*, pages 120–124.

[141] Spitzner, L. (2003). The honeynet project: Trapping the hackers. *IEEE Security & Privacy*, 99(2):15–23.

[142] Stallings, W. (2006). *Cryptography and network security, 4/E*. Pearson Education India.

[143] Staniford, S., Hoagland, J. A., and McAlerney, J. M. (2002). Practical automated detection of stealthy portscans. *Journal of Computer Security*, 10(1-2):105–136.

[144] Stefanova, M., Stefanov, S., and Asenov, O. (2012). Identity protection accessing e-government through the biometric authentication methods. In *Intelligent Systems (IS), 2012 6th IEEE International Conference*, pages 403–408. IEEE.

[145] Sullivan, C. and Burger, E. (2017). E-residency and blockchain. *computer law & security review*, 33(4):470–481.

[146] Swan, M. (2015). *Blockchain: Blueprint for a new economy.* " O'Reilly Media, Inc.".

[147] Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-15(1):116–132.

[148] Tanzania Cyber-Attacks (2018). How Tanzania lost Tanzanian shillings 187billions to cyber criminals in 2016. "https://www.ippmedia.com/en/business/how-tanzania-lost-187bn-cyber-criminals-2016/. Accessed: 2018-02-14.

[149] Thakur, V., Doja, M., Dwivedi, Y. K., Ahmad, T., and Khadanga, G. (2020). Land records on blockchain for implementation of land titling in india. *International Journal of Information Management*, 52:101940.

[150] Tizard, I. R. (1995). *Immunology: An Introduction.* Saunders College Pub.

[151] Tountopoulos, V., Giannakoudaki, I., Giannakakis, K., Korres, L., and Kallipolitis, L. (2014). Supporting security and trust in complex e-government services. In *Secure and Trustworthy Service Composition*, pages 219–233. Springer.

[152] Turkanović, M., Hölbl, M., Košič, K., Heričko, M., and Kamišalić, A. (2018). Eductx: A blockchain-based higher education credit platform. *IEEE access*, 6:5112–5127.

[153] Twizeyimana, J. D. and Andersson, A. (2019). The public value of e-government–a literature review. *Government information quarterly*, 36(2):167–178.

[154] UK Cyber-Attacks (2019). Cyber Security Breaches Survey 2019. "https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2019/. Accessed: 2020-02-13.

[155] Un, E. (2016). government survey 2014. *E-Government in Support of Sustainable Development/UN Department of Economic and Social Affairs*, 22(10):203–214.

[156] Underwood, S. (2016). Blockchain beyond bitcoin. *Communications of the ACM*, 59(11):15–17.

[157] UNFPA (2018). Urbanization. "https://www.unfpa.org/urbanization/. Accessed: 2020-01-26.

[158] UNPAN, U. (2018). e-government survey 2018. *Leveraging E-government at a Time of Financial and Economic Crisis, New York: UNPAN, Retrieved on*, 2:1–300.

[159] US Cyber-Attacks (2020). personal information of 145 million Americans exposed. "https://www.businessinsider.in/tech/news/us-says-chinas-military-was-behind-2017-equifax-hack-that-left-personal-information-of-145-million-ar
articleshow/74069213.cms/. Accessed: 2020-02-14.

[160] Van Zoonen, L. (2016). Privacy concerns in smart cities. *Government Information Quarterly*, 33(3):472–480.

[161] Verizon Report 2019 (2019). Verizon Insider Threat Report . "https://www.verizon.com/about/news/verizon-refocuses-cyber-investigations-spotlight-world-insider-threats/. Accessed: 2020-03-22.

[162] Vigna, G. and Kemmerer, R. A. (1999). Netstat: A network-based intrusion detection system. *Journal of computer security*, 7(1):37–71.

[163] Vujicic, D., Jagodic, D., and Randjic, S. (2018). Blockchain technology, bitcoin, and ethereum: A brief overview. In *INFOTEH-JAHORINA (INFOTEH), 2018 17th International Symposium*, pages 1–6. IEEE.

[164] Warkentin, M., Gefen, D., Pavlou, P. A., and Rose, G. M. (2002). Encouraging citizen adoption of e-government by building trust. *Electronic markets*, 12(3):157–162.

[165] Wireless Broadband (2016). Wireless Broadband Alliance Launches City Wi-Fi Roaming Project. "https://www.wballiance.com/wireless-broadband-alliance-launches-city-wi-fi-roaming-project///. Accessed: 2020-01-08.

[166] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[167] Wood, G. (2014). Ethereum: A secure decentralised generalised transaction ledger. *Ethereum Project Yellow Paper*, 151:1–32.

[168] Xu, J. J. (2016). Are blockchains immune to all malicious attacks? *Financial Innovation*, 2(1):1–9.

[169] Yang, L., Chen, C., Jin, N., Fu, X., and Shen, Q. (2014). Closed form fuzzy interpolation with interval type-2 fuzzy sets. In *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 2184–2191.

[170] Yang, L., Elisa, N., and Eliot, N. (2019a). Privacy and security aspects of e-government in smart cities. In *Smart cities cybersecurity and privacy*, pages 89–102. Elsevier.

[171] Yang, L., Li, J., Elisa, N., Prickett, T., and Chao, F. (2019b). Towards big data governance in cybersecurity. *Data-Enabled Discovery and Applications*, 3(1):1–12.

[172] Yang, L., Li, J., Fehringer, G., Barraclough, P., Sexton, G., and Cao, Y. (2017). Intrusion detection system by fuzzy interpolation. In *2017 IEEE international conference on fuzzy systems (FUZZ-IEEE)*, pages 1–6. IEEE.

[173] Yang, L. and Shen, Q. (2011a). Adaptive fuzzy interpolation with prioritized component candidates. In *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, pages 428–435.

[174] Yang, L. and Shen, Q. (2011b). Adaptive fuzzy interpolation with uncertain observations and rule base. In *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, pages 471–478.

[175] Yang, L. and Shen, Q. (2013). Closed form fuzzy interpolation. *Fuzzy Sets and Systems*, 225:1–22.

[176] Yang, X.-S. (2020). *Nature-inspired optimization algorithms*. Academic Press.

[177] Zhang, A. and Lin, X. (2018). Towards secure and privacy-preserving data sharing in e-health systems via consortium blockchain. *Journal of medical systems*, 42(8):1–18.

[178] Zheng, Z., Xie, S., Dai, H., Chen, X., and Wang, H. (2017). An overview of blockchain technology: Architecture, consensus, and future trends. In *Big Data (BigData Congress), 2017 IEEE International Congress on*, pages 557–564. IEEE.