ABNORMAL TRANSACTIONS DETECTION IN THE ETHEREUM NETWORK USING SEMI-SUPERVISED GENERATIVE ADVERSARIAL NETWORKS

SALAM RADI MAHMOUD AL-E'MARI

UNIVERSITI SAINS MALAYSIA

2022

ABNORMAL TRANSACTIONS DETECTION IN THE ETHEREUM NETWORK USING SEMI-SUPERVISED GENERATIVE ADVERSARIAL NETWORKS

by

SALAM RADI MAHMOUD AL-EMARI

**Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy**

April 2022

ACKNOWLEDGEMENT

بسم الله الرحمن الرحيم

{نَرْفَعُ دَرَجَاتٍ مِّن نَّشَاء وَفَوْقَ كُلِّ ذِي عِلْمٍ عَلِيمٌ "76"} (سورة يوسف)

First and foremost, All Praises be to " Allah " the Almighty, for delivering me the patience, the strength and the guidance in completing this thesis successfully.

Since the Prophet MOHAMMED "Peace be Upon Him" said: “Whoever does not thank people (for their favours) has not thanked Allah (properly)”, therefore, I would like to express my appreciation to my main supervisor Dr. Mohammed Anbar for his guidance, helpful suggestions, patience and valuable comments during his supervision period at National Advanced IPv6 Centre of Excellence (NAv6), Universiti Sains Malaysia (USM). My appreciation also for my co-supervisor Dr. Selvakumar Manickam for his comments and suggestions.

I am extremely grateful to my filed supervisor and my colleague Dr. Yousef Sanjalawe who spent his efforts for helping, supporting, and advising me in my PhD. I also wish to thank my research committee members at NAv6 for providing constructive comments.

In this situation, I never forget my parents, for their prayers for me to complete my Ph.D. Their prayers were the most important thing for me to be strong and reach to what I’m now. I would like also to thank my brothers, sisters and friends for their great encouragement, patience, support, and love. I dedicate this work to all of these people, and those who support me in the background by their prayers.

TABLE OF CONTENTS

ACKNOWLEDGEMENT ii

TABLE OF CONTENTS iii

LIST OF TABLES viii

LIST OF FIGURES x

LIST OF ABBREVIATIONS xiii

ABSTRAK xv

ABSTRACT xvii

CHAPTER 1 INTRODUCTION 1

1.1 Overview 1

1.2 Background 3

1.2.1 Blockchain Technology 3

1.2.2 Ethereum Network 4

1.2.2(a) Abnormal Transactions 6

1.2.3 Intrusion Detection System 7

1.3 Research Motivation 8

1.4 Research Problem 9

1.5 Research Objectives 11

1.6 Research Scope 12

1.7 Terminologies 13

1.8 Research Contributions 14

1.9 Research Steps 15

1.10 Thesis Organization 16

CHAPTER 2 LITERATURE REVIEW 18

2.1 Introduction 18

2.2 Background 18

2.2.1 Basic Principles of Blockchain Technology 20

2.2.2 Blockchain Benefits, Challenges, and Threats............................... 24

2.2.3 Ethereum Network 27

2.2.3(a) Ethereum Accounts... 27

2.2.3(b) Ethereum Transactions 28

2.2.3(c) Security Issues in Ethereum Network............................ 30

2.2.4 Intrusion Detection System 33

2.2.4(a) IDS Types 33

2.2.4(b) IDS Approaches.. 34

2.2.5 Machine Learning 39

2.2.5(a) K-Nearest Neighbour.. 40

2.2.6 Deep Learning.. 41

2.2.6(a) Convolutional Neural Network..................................... 43

2.2.6(b) Recurrent Neural Network.. 44

2.2.6(c) Generative Adversarial Networks (GAN) 44

2.2.6(d) Semi-Supervised GAN 47

2.2.7 Feature Extraction 49

2.2.8 Feature Selection.. 50

2.2.8(a) Bio-Inspired Feature Selection 50

2.2.8(b) Manta Ray Foraging Optimization (MRFO) 53

2.2.8(c) Practical Swarm Optimization (PSO)............................ 55

2.3 Related works... 56

2.3.1 Supervised Anomaly Detection in Blockchain Network 57

2.3.2 Unsupervised Anomaly Detection in Blockchain Network........... 60

2.3.3 Research Gaps.. 66

2.4 Summary 70

CHAPTER 3 METHODOLOGY 71

3.1 Introduction 71

3.2 Overview of ATD-SGAN 71

3.3 Requirements of ATD-SGAN 73

3.4 Proposed ATD-SGAN 73

3.4.1 Ethereum Dataset Generation (Stage 1) 75

3.4.1(a) The Requirements of Dataset 75

3.4.1(b) Data Gathering 76

3.4.2 Data Pre-processing (Stage 2) 77

3.4.2(a) Cleansing 78

3.4.2(b) Labeling 78

3.4.2(c) Balancing 79

3.4.2(d) Scaling 80

3.4.3 Feature Extraction (Stage 3) 81

3.4.4 Ensemble Bio-Inspired Feature Selection (Stage 4) 82

3.4.4(a) MRFO based Feature Selection 83

3.4.4(b) PSO based Feature Selection 84

3.4.4(c) Proposed Multi-Objective Function 84

3.4.4(d) Mutual Features 85

3.4.5 Abnormal Transactions Detection (Stage 5) 86

3.5 Evaluation Metrics 88

3.6 Summary 90

CHAPTER 4 DESIGN AND IMPLEMENTATION 91

4.1 Introduction 91

4.2 Implementation Environment 91

4.2.1 Programing language 91

4.2.2 Experimental Environment 92

4.2.2(a) Hardware Specifications 93

4.2.2(b) Software Specifications 94

4.3 Design ATD-SGAN Approach 94

4.3.1 Ethereum Dataset Generation 94

4.3.1(a) Acquired Intrusion Ethereum Addresses 97

4.3.1(b) Abnormal Transactions 97

4.3.2 Data Pre-processing 99

4.3.2(a) Data Cleansing 100

4.3.2(b) Data Labelling 100

4.3.2(c) Data Balancing 103

4.3.2(d) Data Scaling 103

4.3.3 Feature extraction 103

4.3.4 Ensemble Feature Selection Mechanism 107

4.3.4(a) MRFO Algorithm 107

4.3.4(b) PSO Algorithm 109

4.3.4(c) Mutual Features 111

4.3.5 Abnormal Transactions Detection 113

4.3.5(a) Hyperparameters of ATD-SGAN 113

4.3.5(b) Detection using SGAN 114

4.4 Summary 116

CHAPTER 5 RESULTS AND DISCUSSION 117

5.1 Introduction 117

5.2 Ensemble Features Results 118

5.3 Evaluation of ATD-SGAN 120

5.3.1 In Terms of Accuracy 120

5.3.2 In Terms of F1-measure 122

5.3.3 In Terms of False Alarm Rate 124

5.4 Comparison with State-of-the-Art Approaches 125

5.5 Significance of Enhancement 126

5.6 Discussion 129

5.6.1 In Terms of accuracy 129

5.6.2 In Terms of False Alarm Rate 130

5.6.3 In Terms of F1-measure 131

5.7 Datasets Comparison 133

5.8 Summary 136

CHAPTER 6 CONCLUSION AND FUTURE WORKS 137

6.1 Introduction 137

6.2 Conclusion 137

6.3 Future Work 139

REFERENCES 140

LIST OF PUBLICATIONS

LIST OF TABLES

Page

Table 1.1 Evolution of Blockchain Technology 4

Table 1.2 Comparison between Bitcoin and Ethereum Blockchain Network......5

Table 1.3 Mapping of the Research Challenges, Objectives and Contributions..15

Table 2.1 Examples of Consensus Algorithms 21

Table 2.2 Fields of Block Structure in Blockchain 23

Table 2.3 Benefits of Blockchain...24

Table 2.4 Challenges of Blockchain 25

Table 2.5 Main Threats of Blockchain...26

Table 2.6 Data Mining Techniques..37

Table 2.7 Learning Methods 39

Table 2.8 Summary of Related Works...63

Table 2.9 Features Extraction in The Existing Approaches to Detect Abnormal Transactions in Blockchain Networks 67

Table 3.1 Data Labeling...79

Table 3.2 Values of the Objectives' Weights...85

Table 3.3 Confusion Matrix 88

Table 4.1 Python Libraries Used to Implement The ATD-SGAN.....................92

Table 4.2 Ethereum Transactions Over Three Years 99

Table 4.3 Fields Description of the BLTE Dataset 101

Table 4.4 Statistics of The Proposed Dataset...103

Table 4.5 Features Code...104

Table 4.6 Extracted Features of Ethereum Transaction Example 106

Table 4.7 Values of MRFO Control Parameters 108

Table 4.8 Subset Features by MRFO 109

Table 4.9 Values of PSO Control Parameters 110

Table 4.10 Subset Features by PSO 111

Table 4.11 Hyperparameters of ATD-SGAN 113

Table 5.1 Sample of Results After Ensemble Feature Selection Stage 119

Table 5.2 Values of Performance Metrics of ATD-SGAN and CNN 123

Table 5.3 Values of Performance Metrics of ATD-SGAN with Different Epochs 123

Table 5.4 Comparison Results of ATD-SGAN with State-of-the-Art Approaches 125

Table 5.5 Significant of Enhancement Results 128

Table 5.6 Qualitative Comparison between Proposed Dataset and Existing Datasets 135

LIST OF FIGURES

Page

Figure 1.1 Total Illicit Values vs. Illicit Share of Total Cryptocurrency Activity, 2017-2020 2

Figure 1.2 Evolution of Blockchain Technology 4

Figure 1.3 Evolution of Ethereum Network 5

Figure 1.4 Ethereum Nodes Running on The Ethereum Network 8

Figure 1.5 Research Scope 12

Figure 1.6 Research Steps 15

Figure 2.1 Blockchain Structure 19

Figure 2.2 Blockchain Over P2P Network 19

Figure 2.3 Externally Owned Accounts vs. Contract Accounts 27

Figure 2.4 The life Cycle of an Ethereum Transaction 29

Figure 2.5 Attacks on Ethereum Blockchain Layers 31

Figure 2.6 IDS Classification-based Location 34

Figure 2.7 Detection Approaches 35

Figure 2.8 The Performance of DL Concerning The Amount of Data 42

Figure 2.9 Structure of Deep Learning 42

Figure 2.10 Schematic Diagram of a Basic CNN Architecture 43

Figure 2.11 RNN Architecture 44

Figure 2.12 GAN Architecture 45

Figure 2.13 Taxonomy of GANs Variants 46

Figure 2.14 SGAN Architecture 47

Figure 2.15 Criteria of Feature Extraction Techniques 49

Figure 2.16 Steps of Evolutionary Algorithms 51

Figure 2.17 Swarm-based Algorithm Stages 52

Figure 2.18 (a) Manta Rays Body, (b) Manta Rays' Physical Construction 53

Figure 3.1 General Structure of ATD-SGAN 72

Figure 3.2 The Main Stages of ATD-SGAN 74

Figure 3.3 Graph Representation of Mutual Information in ATD-SGAN 86

Figure 4.1 Experimental Steps of ATD-SGAN 93

Figure 4.2 Workflow of Ethereum Dataset Generation 95

Figure 4.3 Snapshot of BLTE Dataset Resulting from Generation Stage 96

Figure 4.4 Text Description Format for Ethereum Address 97

Figure 4.5 Percentage of Intrusion Addresses in The BLTE Dataset During 2017-2019 98

Figure 4.6 Comparisons between Normal and Abnormal Transactions During 2017-2019 99

Figure 4.7 Snapshot of the BLTE Dataset after Data Labelling 102

Figure 4.8 Sample of Multi-digraph of Ethereum Nodes 105

Figure 4.9 The General Procedure of Mutual Features 112

Figure 4.10 Flowchart of ATD-SGAN Approach 116

Figure 5.1 Evaluation Strategy 117

Figure 5.2 Mutual Feature Selection Results 118

Figure 5.3 Detection Accuracy of ATD-SGAN with Different Epochs 121

Figure 5.4 F1-meausre of ATD-SGAN with Different Epochs 122

Figure 5.5 False Alarm Rate of ATD-SGAN with Different Epochs 124

Figure 5.6 Enhancement Percentages of ATD-SGAN with Other State-of-the-Art IDSs in Terms of Accuracy 129

Figure 5.7 Enhancement Percentages of ATD-SGAN with Other State-of-the-Art IDSs in Terms of False Alarm 130

Figure 5.8 Enhancement Percentages of ATD-SGAN with Other State-of-the-Art IDSs in Terms of F1-measure 131

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
API	Application Programming Interface
AR	Association Rule
ATD-SGAN	Abnormal Transactions Detection based on Semi-supervised Generative Adversarial Network
BLTE	Benchmark Labelled Transactions Ethereum
BTC	Bitcoin
CIDS	Collaborative Intrusion Detection System
CNN	Convolutional Neural Network
DAO	Decentralized Autonomous Organization
DApps	Decentralized Applications
DDoS	Distributed Denial-of-Service
DL	Deep Learning
DoS	Denial-of-Service
DT	Decision Tree
EOA	Externally Owned Account
ERC	Ethereum Request for Comments
ETC	Ethereum Classic
ETH	Ethereum
EVM	Ethereum Virtual Machine
FE	Feature Engineering
FN	False Negative
FPR	False Positive Rate
FS	Feature Selection
GA	Genetic Algorithm
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
HIDS	Host-based Intrusion Detection System
ICO	Initial Coin Offerings
IDS	Intrusion Detection System
IF	Isolation Forest
IoT	Internet of Things

IPS	Intrusion Protection System
KNN	K-Nearest Neighbors
LDIS	Local Density based Instance Selection
LOF	Local Outlier Factor
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multi-Layer Perceptron
MRFO	Manta Ray Foraging Optimization
NIDS	Network-based Intrusion Detection System
NN	Neural Network
Opcodes	Operation Codes
P2P	Peer-to-Peer
PBFT	Practical Byzantine Fault Tolerance
PCA	Principle Component Analysis
PoET	Proof-of-Elapsed time
PoS	Proof of Stake
PoSp	Proof-of-Space
PoW	Proof of Work
PSO	Particle Swarm Optimization
RF	Random Forest
RNN	Recurrent Neural Network
SGAN	Semi-supervised Generative Adversarial Network
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
Trx	Transaction

PENGESANAN TRANSAKSI ABNORMAL DALAM RANGKAIAN ETHEREUM MENGGUNAKAN RANGKAIAN PERSETERUAN GENERATIF SEPARA-SELIAAN

ABSTRAK

Rangkaian Ethereum adalah suatu platform blockchain yang membolehkan pengguna melakukan transaksi matawang kripto, membuat, dan menggunakan aplikasi terdesentralisasi menggunakan kontrak pintar. Beberapa transaksi abnormal mula muncul akibat serangan sedia ada yang mensasarkan Ethereum, misalnya, serangan Ethereum DAO, dan pengguna berniat jahat dapat mengeksploitasi dan menjejaskan kelemahan dalam kontrak pintar, untuk mencuri sejumlah matawang kripto atau berusaha untuk memenuhi matlamat mereka sendiri melalui transaksi abnormal. Oleh itu, mengesan transaksi tidak normal oleh pengguna berniat jahat ini, yang terlibat dalam aktiviti penipuan dan juga atribusi adalah amat rumit. Walau bagaimanapun, aktiviti jahat menggunakan transaksi matawang kripto, melalui akaun palsu samaran untuk menghantar dan menerima pembayaran tebusan, penyatuan dana yang terkumpul pada pelbagai identiti berbeza; dengan itu, mengawal dan mengesan transaksi abnormal ini adalah prasyarat asas untuk memastikan tahap keselamatan rangkaian Ethereum berada pada tahap yang tinggi. Oleh yang demikian, tesis ini mencadangkan suatu pendekatan untuk mengesan transaksi abnormal dalam rangkaian Ethereum, yang disebut ATD-SGAN, berasaskan Rangkaian Perseteruan Generatif Separa-seliaan (SGAN). Hasil kajian menunjukkan bahawa ATD-SGAN yang dicadangkan berjaya meningkatkan prestasi pendekatan terkini daripada 3.78% kepada 11.05% dari segi ketepatan pengesanan. Sebaliknya, ATD-SGAN Berjaya mengurangkan kadar penggera palsu daripada 42.29% ke 0.15%. Juga, ATD-SGAN telah meningkatkan ukuran F1 daripada 10.39% kepada 3.79%.

ABNORMAL TRANSACTIONS DETECTION IN THE ETHEREUM NETWORK USING SEMI-SUPERVISED GENERATIVE ADVERSARIAL NETWORKS

ABSTRACT

Ethereum network is a blockchain platform that allows users to use cryptocurrency transactions, create, and deploy decentralized applications using smart contracts. Several abnormal transactions came to light due to the existing attacks that targeted Ethereum, for instance, the Ethereum DAO attack, and malicious users might exploit and compromise the vulnerabilities in smart contracts, to steal amount of cryptocurrency or working for their own objectives through abnormal transactions. Therefore, detecting abnormal transactions initiated from these malicious users, implicated in fraudulent activities as well as attribution is excessively complex. However, malicious activities using cryptocurrency transactions, through pseudo-anonymous accounts for sending and receiving ransom payment, consolidation of funds heaped up under diverse identities; thus, controlling and detecting these abnormal transactions is a fundamental pre-requisite to ensure the high level of security in Ethereum network. Therefore, this thesis proposes an approach for detecting abnormal transactions in Ethereum network, called *ATD-SGAN*, which is based on a Semi-supervised Generative Adversarial Network (SGAN). The results show that the proposed *ATD-SGAN* enhances the performance of state-of-art approaches from 3.78% to 11.05% in terms of detection accuracy. On the other hand, *ATD-SGAN* reduces the false alarm rate ranging from 42.29% to 0.15%. Also, *ATD-SGAN* enhances F1-measure ranging from 10.39% to 3.79%.

CHAPTER 1

INTRODUCTION

1.1 Overview

Blockchain is an emerging technology that underlies the infrastructure of Bitcoin. In 2008, Nakamoto discovered blockchain's potential to be used in other domains, thus making Bitcoin the first of blockchain's many implementations. Blockchain technology has been increasingly adapted in different fields, especially in the network security field, which has an important presence in different network environments, such as traditional networks, the Internet of Things (IoT), and cloud computing. Blockchain technology has many features that could enhance the network security. One of the most important features of the blockchain technology is that it works with decentralized and distributed environments; therefore, it does not need a trusted third-party to manage the network. Blockchain technology has been applied to cryptocurrency networks, wherein the blockchain provides cryptocurrency its basic infrastructure.

Currently, there are many different types of digital currencies, such as Bitcoin, Litecoin, Ether, and Ripple, which have been built into a new durable ecosystem and may be integrated into different network types. Recently, cryptocurrency remained appealing for attackers, primarily due to its pseudonymous nature and the ease of use, which allows users to instantly send funds anywhere in the world, despite its transparent and traceable design. Figure 1.1 illustrates the total cryptocurrency value sent and received by cyber-attacks vs. share of all abnormal cryptocurrency transactions from

2017 to 2020 according to Chainalysis[1] report, where roughly $21.4 billion worth of value of abnormal transactions has been transferred in 2019. Wherein the illicit share of all cryptocurrency activity fell to $10.0 billion in transaction volume in 2020. Indeed, overall economic activity nearly tripled between 2019 and 2020 is one reason the percentage of illicit activity decreased.

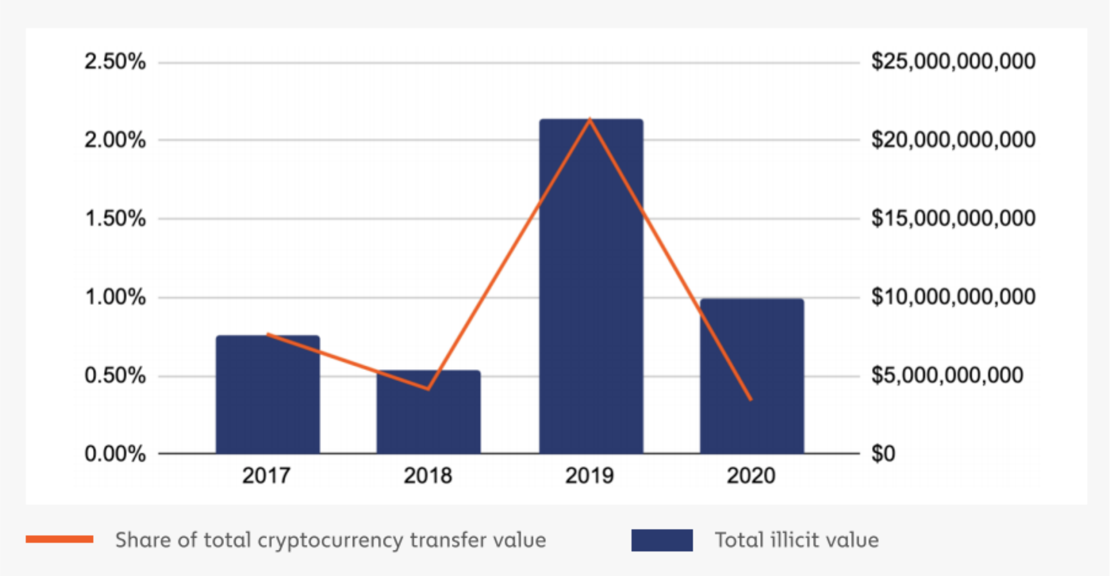

Figure 1.1 Total Illicit Values vs. Illicit Share of Total Cryptocurrency Activity, 2017-2020 (Chainalysis, 2021)

There are many blockchain-based networks technology such as Bitcoin and Ethereum. The adaption of the Ethereum network has been exponentially increased as it is used as a backend for Ether cryptocurrency. Moreover, it has an ability to develop Decentralized Applications (DApps) and smart contracts. However, the Ethereum network has been suffering from a critical challenge in detecting attacks because a variety of cyber-attacks rely on the complexity of this network's environment. The abnormal transactions are common activities among Ethereum-bases attacks such as 51% and Eclipse attacks (Chen *et al.*, 2019). Therefore, detecting abnormal transactions

[1] https://www.chainalysis.com/

could lead to detecting different types of attacks at an early stage. Moreover, traditional Intrusion Detection Systems (IDSs) are incompatible with blockchain networks because of its complex network structure which fails to detect the cyber-attacks related blockchain.

1.2 Background

In this section, an overview of blockchain technology and the difference between Bitcoin and Ethereum blockchain networks are provided. Besides that, discussion on the security issues and abnormal transactions in the Ethereum network. Furthermore, an overview of the IDS to detect cyber-attacks in the Ethereum network is presented.

1.2.1 Blockchain Technology

Blockchain technology was introduced by Nakamoto in 2008 as an underlying technology for Bitcoin to record all transactions of Bitcoin and to create security against potential attacks (Nakamoto, 2008). Figure 1.2 presents blockchain's evolution from 2008 to 2019. Bitcoin's initial infrastructure based on blockchain technology appeared in 2009 over a peer-to-peer (P2P) network, which is called the Bitcoin network. Since then, cryptocurrencies have gained worldwide attention, and the researchers have harnessed and applied blockchain technology to domains, such as smart contracts and supply chain management. This evolution has been the resultant outcome of the blockchain being autonomous, distributive, immutable, and contractual.

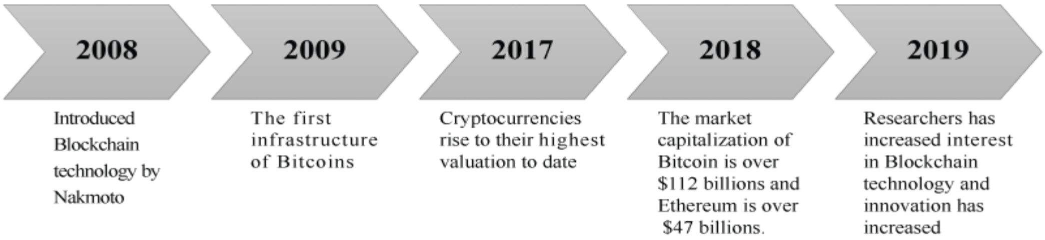
2008
Introduced Blockchain technology by Nakmoto
2009
The first infrastructure of Bitcoins
2017
Cryptocurrencies rise to their highest valuation to date
2018
The market capitalization of Bitcoin is over $112 billions and Ethereum is over $47 billions.
2019
Researchers has increased interest in Blockchain technology and innovation has increased

Figure 1.2 Evolution of Blockchain Technology

However, blockchain technology has continued to evolve since its inception, and Table 1.1 shows the generations of the blockchain technology model from its invention to nowadays (Maesa & Mori, 2020; Mohamed & Al-Jaroodi, 2019).

Table 1.1 Evolution of Blockchain Technology

Generation	Evolution	Description
Blockchain 0.1	Currency	It started with Bitcoin and other cryptocurrencies.
Blockchain 0.2	Smart contracts	It is the first version of Ethereum to run smart contracts.
Blockchain 0.3	DApps	It is the upgrades of the blockchain 0.2 to offer DApps with front-end and back-end interface.
Blockchain 0.4	Industry applications	Satisfying smart manufacturing applications demands through making blockchain usability in real-life business

1.2.2 Ethereum Network

Introduced in 2014 by Vitalik Buterin, Ethereum is the second implementation using the blockchain technology after Bitcoin. Ethereum has been proposed to overcome the challenges in Bitcoin such as block size and time creation (Athina, 2019), more explanations are mentioned in Table 1.2. Besides, the Ethereum defeats the scalability issue and offers cryptocurrency, smart contracts, and DApps based blockchain infrastructure. On the other hand, the Ethereum network is still evolving from 2014 till now (Friebe, 2017; Xie, 2017).

Table 1.2 Comparison between Bitcoin and Ethereum Blockchain Network

Metric	Bitcoin	Ethereum
Released	2009 by S. Nakamoto	2014 by V. Buterin
Concept	Digital Concurrency	Smart Contract & Digital Concurrency
Value	Currency BTC (Bitcoin)	Token ETH (Ethereum)
Release method	Genies Block	Presale
Mining	ASIC (Application-Specific Integrated Circuit)	GPU (Graphics Processing Unit)
Consensus protocol	PoW (Proof of Work)	PoW, planning switch to PoS (Proof of Stake)
Algorithm	SHA256	Ethash
Create block time	10 minutes	12-14 seconds
Block size	1 MB and implement the SegWit mechanism.	Amount of gas (limited by 6.7 million gas limits on each block)
Privacy	Public	Public or Private
Extensibility	Low possibility (Stake-based scripting)	High possibility (Turning-complete)
Scalability	3 Transactions per second	15 Transactions per second
Transaction cost	Block size	Gas

Figure 1.3 illustrates the evolution of the Ethereum network (Athina, 2019; Sheinix, 2019).

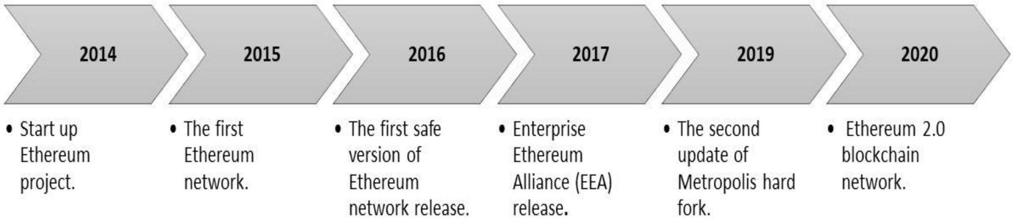
2014
• Start up Ethereum project.
2015
• The first Ethereum network.
2016
• The first safe version of Ethereum network release.
2017
• Enterprise Ethereum Alliance (EEA) release.
2019
• The second update of Metropolis hard fork.
2020
• Ethereum 2.0 blockchain network.

Figure 1.3 Evolution of Ethereum Network

1.2.2(a) Abnormal Transactions

There are two types of Ethereum transactions namely: (i) external transactions to transfer Ether cryptocurrency which is one of the prime activities incidents on Ethereum, and (ii) internal transaction to execute a function of the smart contract. The external transaction includes two fields which are gas and gas price. While the internal transaction includes the name and parameters of the function. In both types, a sender pays a transaction fee to the miner who relies on the consumed gas, and it calculated as Equation 1.1. In addition, the miner is responsible for packing transactions into blocks or executing smart contract instructions (Lin *et al.*, 2020).

$$Transaction\ Fee = consumed\ gas * gas\ price \tag{1.1}$$

Moreover, Ethereum transactions executed by miners where any attack or mistake led to a high cost of failure. Indeed, the main security issue is external transactions because an untrusted account can execute a transaction and leads to unexpected results such as race condition and transaction ordering. Furthermore, the existing abnormal transactions use cryptocurrencies in ransomware, which have seen increased revenue ever since adopting the cryptocurrencies. Another example of abnormal transactions is an increase in investment into Initial Coin Offerings (ICOs), and so on. The existence of abnormal transactions in Ethereum networks are considered as clue for the detecting the Ethereum based attacks. However, the security mechanisms of Ethereum network are still unable to accurately identify or prevent the presence of abnormal transactions in the Ethereum network (Phillips & Wilder, 2020; Rouhani & Deters, 2017). IDS is a security mechanism that is widely used to detect the abnormal traction in Ethereum network. The following subsection provides an overview about IDS.

1.2.3 Intrusion Detection System

An intrusion detection system is a cybersecurity mechanism to find out an attack in the system. There are two main types of IDSs namely: (i) host-based IDS (HIDS) which are installed on one machine and reveal anomalies through unexpected events in a host system, and (ii) Network-based IDS (NIDS) which monitor different network layers to detect attacks (Choudhary & Kesswani, 2019). On the other hand, there are various detection techniques utilized in IDSs, where they are categorized as follows (Garuba *et al.*, 2008; Hodo *et al.*, 2017):

- **Misuse or signature**: Technique has a database or the patterns from prior known attacks, where the IDS needs to update the information to detect a new attack, constantly.
- **Anomaly:** This technique monitors a system's behaviour by constructing a profile through a specific time, this profile has all activities of the system. However, there are different models for creating a profile file for the system, such as time series and threshold models.
- **Hybrid:** This technique combines signature and anomaly detection technique.

Recent research tendencies on IDS leverages, different Deep Learning (DL) algorithms (i.e., Artificial Neural Network (ANN)) have been published from 2017 to protect computer systems and networks from cyber-attacks. The DL techniques have shown outstanding performance in IDS; hence, the DL techniques are widely accepted and deployed (Kim *et al.*, 2020). In addition, the DL can deal with huge data which is appropriate to be adapted in blockchain networks environment to detect abnormal transactions rather than conventional IDS.

1.3 Research Motivation

The motivation of this thesis comes from the spread of cryptocurrencies and smart contracts in several countries. Figure 1.5 demonstrates the Ethereum nodes tracker where the United States has the highest number of Ethereum nodes (Etherscan.io, 2020). Besides, millions of smart contracts have been developed in numerous fields such as IoT, financial, security, etc (Chen *et al.*, 2020). Moreover, Ethereum overcomes the shortcomings of Bitcoin blockchain network, as mentioned above in section 1.2.2.

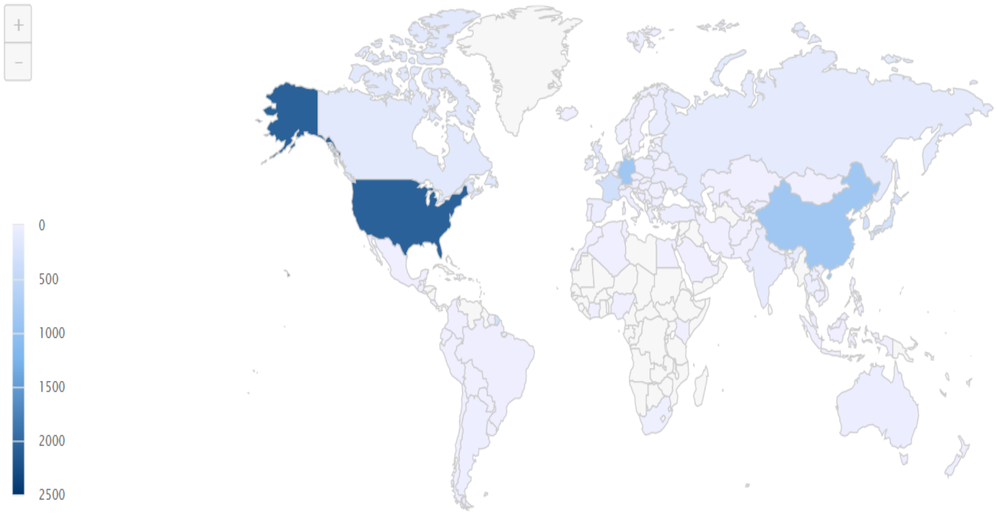

Figure 1.4 Ethereum Nodes Running on The Ethereum Network (Etherscan.io, 2020)

Although the blockchain networks are secure, they are exposed to security vulnerabilities. Consequently, the intruders have emerged in Ethereum networks and made thefts of millions of Ethers. For instance, a DAO attack occurred in 2016 and over $50M were stolen (Brandom, 2016; Chen *et al.*, 2019). In addition, $13M of Ether were stolen by a parity multisig wallet attack in July 2017 and a new version from this attack stole $155M of Ether in November 2017 (Frank *et al.*, 2020). Further, in 2018 integer flow attack stole $2.3 M of Ether (Brent *et al.*, 2018). While $48.7 M of Ether were stolen by an unknown address account in South Korea through cryptocurrency exchange

(Canellis, 2019), and $48.7 M of Ether were stolen by a 51% attack in 2020 (MIT, 2020). Besides that, several attacks attempted to steal cryptocurrencies from the Ethereum network or other malicious actions. All the above-mentioned attacks generate a huge number of abnormal transactions, therefore; detection of these abnormal transactions led to detect the attacks that target Ethereum network. On the other hand, the conventional IDS are unable to detect abnormal transactions because the Ethereum network has a new complex environment and infrastructure. Therefore, it is essential to propose IDS approach mainly to detect abnormal transactions in Ethereum network.

1.4 Research Problem

The dependency on the Ethereum network in different aspects of our life such as cryptocurrencies and decentralized apps has grabbed the attention of the attackers to target the Ethereum network. Furthermore, the technological advent of cryptocurrencies and their respective advantages have been shrouded with several illegal behaviors operating over the blockchain network such as money laundering, phishing, and fraud. Since these technologies exchange huge amounts of sensitive data, and as a result, they are prone to different network attacks and security threats that can affect the provided services network. One of the widespread services is a cryptocurrency exchange such as Bitcoin and Ether that imposed high costs for financial systems because of its exposure to abnormal transactions. For instance, a decentralized autonomous organization (DAO) attack which stole funds $60 million of Ether (Mandloi & Bansal, 2020). Therefore, the security, detection, and protection of the various communication infrastructures using IDSs are of critical importance. However, many challenges have arisen since anomalies are continually changing and involve large amounts demanding a scalable and robust solution.

The existing approaches for detecting abnormal transactions in Ethereum networks are commonly classified into two main categories: (i) supervised, and (ii) unsupervised machine learning-based approaches. Though there are several IDS approaches have been proposed in the literature, such as (Aldwairi & Al-Khamaiseh, 2015; Garcia-Teodoro *et al.*, 2009; Geetha *et al.*, 2018; Sahani *et al.*, 2018), their performance in term of detection accuracy needs to be improved due to several reasons (i) these approaches are evaluated using synthetic datasets where the characteristics of these datasets in term of attack coverage, accuracy, and validity are not revealed. Therefore, these datasets cannot be used to evaluate other future approaches for detecting abnormal transactions (ii) relying on features that are extracted based on simple heuristics to detect abnormal transactions which significantly decreases the detection accuracy.

Meanwhile, deep learning-based approaches have been used in detecting different types of attacks in conventional networks such as IPv4 and IPv6 networks and they show impressive results in terms of detection accuracy, but they are not being commonly used for detecting abnormal transactions in Ethereum network. Although deep learning performs more effectively than conventional techniques, particularly when learning a massive amount of data, but in case of learning from imbalanced data, the performance of deep learning methods decreases significantly. While research on imbalanced data is extensive, many still have challenges, which could cause data loss or overfitting problems.

It is, therefore, peremptory to propose robust IDS that can enhance the detection accuracy, reduce the false alarm rate, and increase the F1-meauer of abnormal transactions in the Ethereum network. The statement of the problem is summarized as follows:

- Lack of availability of reference transaction based Ethereum network dataset.
- Lack of significant features that contribute to detecting abnormal transactions in the Ethereum network.
- Deep learning-based approach still suffers from major challenges including data loss or overfitting problems, which might affect its performance.
- Neglecting the semi-supervised learning method while detecting abnormal transactions in the Ethereum network.

1.5 Research Objectives

The main goal of this thesis is to propose a Semi-supervised Generative Adversarial Network (SGAN) based approach for detecting abnormal transactions in Ethereum network with high performance. To achieve the main goal, the following objectives have been formulated:

1. To propose a benchmark labelled transactions-based dataset of the Ethereum network.
2. To propose an ensemble feature selection mechanism to select the most significant features that contribute to detecting abnormal transactions in the Ethereum network efficiently.
3. To adapt automatic data augmentation mechanism to avoid overfitting and achieve impressive detection performance from few labelled transactions used in training.

1.6 Research Scope

This thesis proposes an approach to detect abnormal transactions in the Ethereum network using SGAN with a high detection performance. The proposed approach extracts a feature-based multi-digraph structure and then uses a semi-supervised learning model to classify Ethereum transactions into normal or abnormal transactions.

Moreover, this thesis relies on using the Generative Adversarial Network (GAN) algorithm, which is one of the most important and popular deep learning models that deals efficiently with both small-sized labelled datasets and large-sized datasets as well (Li *et al.*, 2019; Yang *et al.*, 2019). Figure 1.6 illustrates the scope of this thesis, showing the utilized concepts and their relationships.

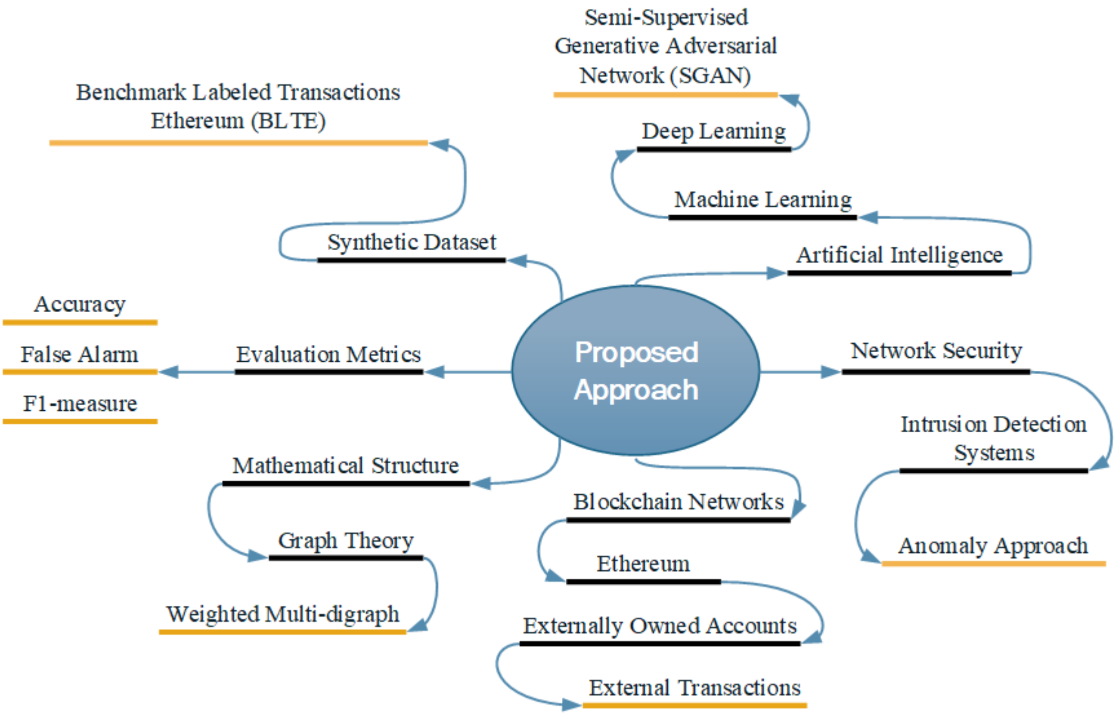

Figure 1.5 Research Scope

1.7 Terminologies

This section provides brief explanation about the terminologies about this research as follows:

Intrusion Detection System (IDS): it is software or hardware that detect system attacks.

Anomaly Approach: it identifies the anomalies in the traffic, and it can detect unknown/new attacks in the network.

Ethereum Network: It is an open and decentralized platform that enables exchange cryptocurrency, development smart contract and decentralized applications that run based on blockchain technology.

External Transactions: the transactions occur between external owned accounts on the Ethereum network to exchange cryptocurrencies. Wherein, the abnormal transactions are resulted from scam and phishing attack.

Internal Transactions: the transactions occur between contract accounts on the Ethereum network to execute smart contracts.

Benchmark Dataset: it is a collection of related set of information that has the basis of fair comparison and validation of computational methods.

Generative Adversarial Networks (GANs): it is artificial intelligence algorithm consists from generative and discriminator neural network model to generate more examples from the estimated probability distribution.

Semi-supervised Generative Adversarial Networks (SGANs): It is kind of GANs that forcing the discriminator network to output class labels and generate more examples from few labels.

Weighted Multi-digraph: A digraph $G = (N, E)$ is directed multigraph (multi-digraph) iff V a set of nodes and E a multiset set of ordered pairs of distinct elements of N called edges. thence, a multigraph allows multiple edges between two nodes.

1.8 Research Contributions

The main contribution this thesis would make is proposing a new approach for detecting abnormal transactions in the Ethereum network based on SGAN with high performance. The achieved contributions of this study can be summarized in the following points:

1. A benchmark dataset of the transactions-based dataset of Ethereum network for the tuning, assessing, and comparing IDSs in Ethereum networks.
2. A set of features that are used in detecting abnormal transactions efficiently.
3. An ensemble feature selection mechanism. This mechanism is based on bio-inspired feature selection algorithms and a multi-objective feature that will be used to reduce the dimensionality of a dataset and will improve detection performance.
4. Semi-supervised feature selection with deep learning, which is suitable for the detection of abnormal transactions in the Ethereum network compared to state-of-the-art approaches.

The mapping between research objectives (RO), and research contributions (RC) of this thesis are summarized in Table 1.3.

Table 1.3 Mapping of the Research Challenges, Objectives and Contributions

Challenge(s)	RO	RC
Lack of availability of reference transaction-based blockchain network dataset.	RO1	RC1
Lack of significant features that contribute to detecting abnormal transactions in the Ethereum network.	RO2 RO3	RC2 RC3
Deep learning-based approach still suffers from major challenges including data loss or overfitting problems, which might affect its result in.	RO3 RO4	RC3 RC4
Neglecting the semi-supervised learning method while detecting abnormal transactions in the Ethereum network. The need for huge number of labelled transactions in training stage to train the classifier efficiently.	RO4	RC4

1.9 Research Steps

This thesis proposes a new approach based deep learning method to detect intrusion addresses' Ethereum based on abnormal transactions. Figure 1.7 illustrates the research steps that will be followed to achieve the objectives of this thesis.

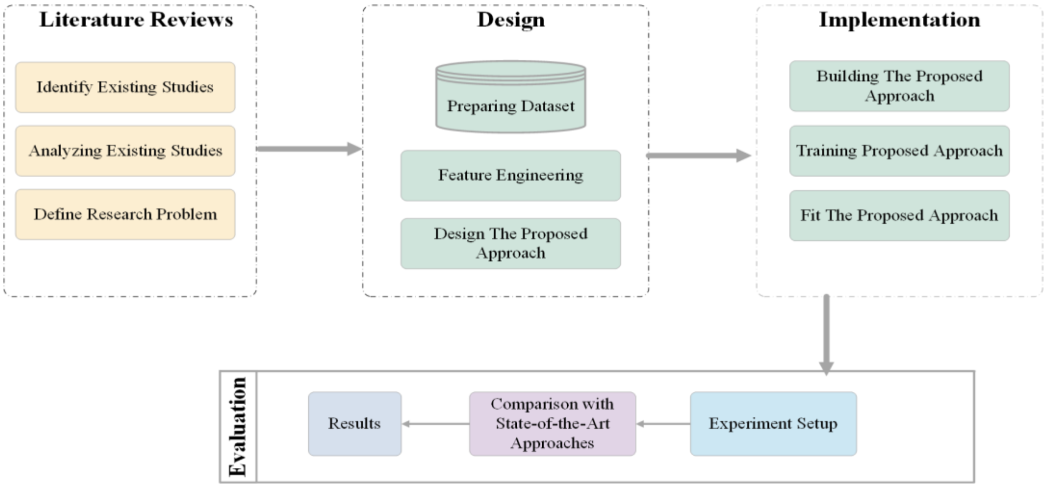
Literature Reviews
Identify Existing Studies
Analyzing Existing Studies
Define Research Problem
Design
Preparing Dataset
Feature Engineering
Design The Proposed Approach
Implementation
Building The Proposed Approach
Training Proposed Approach
Fit The Proposed Approach
Evaluation
Results
Comparison with State-of-the-Art Approaches
Experiment Setup

Figure 1.6 Research Steps

Step one is to study and analyse the related works. In this step, an analysis of the existing IDS used in blockchain networks is conducted and the challenges and research gaps of these models are highlighted to identify the research problem.

Step two is proposing the solution by identifying the main stages and requirements needed to design and implement it.

Step Three is designing and implementing the proposed solution by clarifying the design of each stage of the proposed approach, benchmark dataset, and evaluation metrics. Besides, the implementation and configurations of the proposed approach are presented in this step.

Step four is to evaluate the proposed solution. The proposed solution is evaluated based on the evaluation strategy and comparison with the state-of-the-art approaches based on predefined evaluation metrics.

1.10 Thesis Organization

This thesis comprises the following chapters:

CHAPTER 2 discusses the research background and related studies. This chapter critically reviews the existing solutions for detecting anomalies in the Ethereum network. Furthermore, this chapter comprehensively discusses blockchain technology, intrusion detection system, and deep learning approaches.

CHAPTER 3 presents the research methodology in detail. It shows how the methodology phases have been integrated to achieve the research objectives.

CHAPTER 4 discusses the design and implementation of the proposed approach. This chapter discusses the design and implementation of each phase of the proposed policy in details. Besides, evaluation metrics of the proposed approach is also discussed in detail.

CHAPTER 5 analyzes the experiments and their findings. In addition, it presents a comprehensive analysis of the results achieved using the proposed approach. Moreover, the performance of the proposed approach has been evaluated in comparison with existing state-of-the-art approaches in this chapter.

CHAPTER 6 presents the conclusions and discusses future research directions.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents the reader with the necessary background on the main concepts and components that have been used throughout this thesis. The chapter starts by giving a comprehensive review of blockchain technology and Ethereum network, followed by the basic concepts of IDSs, machine learning methods, deep learning methods, feature extraction, and feature selection in Section 2.2. Next, the chapter provides a brief review of the related studies based on existing approaches that adopt IDSs in two blockchain networks are the Ethereum and Bitcoin network to detection abnormal transactions. In addition, analysis of the limitations of the current approaches whereas these restrictions motivated to develop the proposed approach in this thesis in Section 2.3. Finally, the chapter is summarized in Section 2.4.

2.2 Background

A blockchain is a linked-data structure wherein each block has two main sections: a header and body. The header section consists of a nonce, a previous hash, a Merkle root hash, a timestamp, and a difficulty target. The body section contains a list of transactions. Figure 2.1 presents the structure of a blockchain. The first block is always called a genesis, all blocks are linked together via cryptography, and blocks are distributed between nodes over a network (Gao *et al.*, 2018; Udemy, 2019).

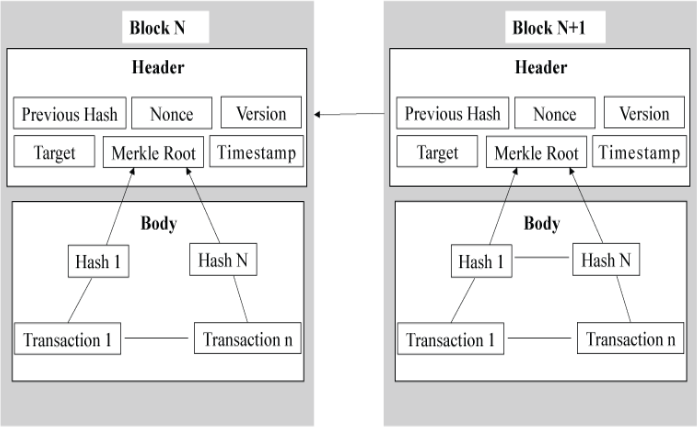

Figure 2.1 Blockchain Structure

Furthermore, to adhere to the rules of blockchain technology, all nodes in the blockchain network must have the same block list, which is presented in Figure 2.2. When a new block is added, it broadcasts to all nodes in the network. Each node verifies the new block through a consensus mechanism that confirms a transaction in the block. There are various consensus algorithms to ensure that all nodes have the same blockchain list, such as proof of work and proof of stake (Liang *et al.*, 2017; Muzammal *et al.*, 2019).

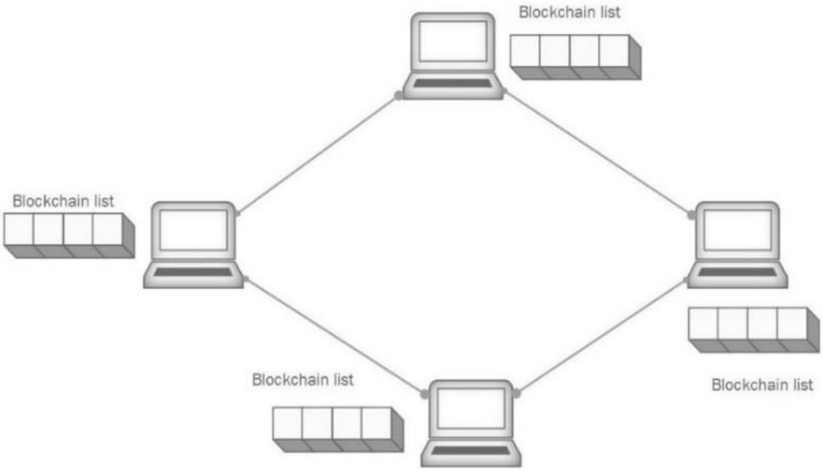

Figure 2.2 Blockchain Over P2P Network

2.2.1 Basic Principles of Blockchain Technology

There are many principles of blockchain technology that are applied to three main layers: the network, data, and application layers. First, the network layer is compatible with the P2P network architecture, which supports decentralized connections and distributed network mechanisms. The network layer is responsible for forwarding and verifying data between nodes. In addition, blockchain technology stores the same chain in all nodes over a network; thus, all nodes are synchronized. Therefore, when a new block is generated, it is then verified by a consensus algorithm. If the new block is valid, then it broadcasts to all other nodes. Otherwise, it is discarded. In addition, there are several types of consensus algorithms that all operate on two principles: (i) the freshness principle achieves fair competition through fresh resources for each new block that is added, and (ii) the unpredictability principle prevents any participant from predicting which node will create a new block. Table 2.1 illustrates description, advantages and disadvantages for some of the consensus algorithms that are used in blockchain networks such as Proof-of-Work (PoW), Proof-of-Stake (PoS), Proof-of-Elapsed time (PoET), and Practical Byzantine Fault Tolerance (PBFT) (Yuan & Wang, 2018; Zheng *et al.*, 2017).

Table 2.1 Examples of Consensus Algorithms

Algorithm	**Description**	**Advantages**	**Disadvantages**
Proof-of-Work (PoW)	PoW is widely used in blockchain verification to validate data in complex mathematical computation. The first node solves the crypto puzzle, then it adds a new block that will be verified later, by using existing-verified nodes in the network.	Verification technique for PoW is extremely efficient	High power consumption
Proof-of-Stake (PoS)	PoS selects participants based on their stake cryptocurrency	It reduces energy consumption in PoW, and it is efficient for large-scale networks.	It suffers from a DoS attack, and there is a lack of synchronization between participants.
Proof-of-Elapsed time (PoET)	Randomly, it generates waiting time slots for each participant, while a user who has less waiting time will be added into a new block.	It consumes less energy than PoW. Also, it ensures freshness and unpredictability principles.	It does not indicate how the algorithm can solve the conflict. Also, its voting approach is very complicated.
Proof-of-Space (PoSp)	A verifier requests from the prover to reserves a disk space to store necessary information, then a prover	It reduces power consumption, which makes it more difficult	Producing a new block is difficult; therefore, it is challenging in solving the

	sends to the verifier to ensure reserving that disk space.	for malicious participants to join the network.	distributed consensus problem.
Practical Byzantine Fault Tolerance (PBFT)	There are three sequential steps required to add a new block to the chain successfully, namely: (i) new round, (ii) prepare, and (iii) commit, where each step is executed after getting two-thirds voting from nodes in the network.	It can handle a third pernicious network. No need for the miner; thus, it reduces energy consumption efficiently.	The node can not join the network before verifying it by the whole network.

Second, the data layer presents the data structure of the block. Blocks contain data or transactions that do not exceed several megabytes in size. Each block is linked together by a previous hash field through a miner. When a block solves a cryptographic puzzle and obtains the previous hash, a new block is appended to the end of the chain. Furthermore, each block has several fields are version, timestamp, previous hash, target, nonce, Merkle root, and hash which are described in Table 2.2. The data layer also concerns user authentication and transaction encryption. Each user has a public key to validate authentications, and this key is visible to anyone in the blockchain network. Digital signatures are used to verify miners' transactions, and all validated transactions are kept in a public ledger (Gao *et al.*, 2018; Ismail *et al.*, 2019).

Table 2.2 Fields of Block Structure in Blockchain

Field	**Description**
Version	It is the identification rules used by the protocol.
Timestamp	It records the time required for creating a block, and it is used for ensuring traceability.
Previous Hash	It indicates the previous block used for linking the current block with the chain.
Target (nBit)	It is used by consensus algorithms to define the difficulty level of their mechanism.
Nonce	It is calculated by the miner to generate a hash block, while it should be a unique number and leading by zeros.
Merkle Root	It includes all hashes values of legitimate transactions.
Hash	Hashing transaction occurs by Merkel tree, where each node is related with its parent node; therefore, if the transaction is modified, then it will affect all hash tree from the leaf node to the Merkle root, respectively.

Finally, the application layer is responsible for interacting with users, whether they are programmers or end-users. The application layer can be classified into two different layers. The first layer is meant for developers to build and test the application's code and is called the fabric layer. The second layer is the application layer, which allows end-users who use applications as a black box to perform specific tasks without knowing the details of the code (Glaser, 2017). The next section 2.2.2 presents some of blockchain's benefits, challenges, and threats.

2.2.2 Blockchain Benefits, Challenges, and Threats

Blockchain technology provides several benefits to its users. Some of these benefits are summarized below in Table 2.3. The main advantage of blockchain is its decentralization feature. Decentralization means that there is no need for third parties and that all participants make decisions about the information contained in a network (Niranjanamurthy *et al.*, 2018).

Table 2.3 Benefits of Blockchain

Field	**Description**
Decentralization	The nodes might share transactions between themselves without the need for a central point.
Empowered Users	Users have full privilege and permission to manage their transactions before adding them into a blockchain list, while those users can only read their transactions after adding them into a blockchain list.
High-quality data	Data in a blockchain is available over different nodes consistently. It is characterized by its accurateness and freshness.
Reliability and Robustness	Blockchain's nodes resist against any malicious attack since it is a decentralized network.