Improved Blockchain - Proof-of-Work Consensus Protocol - Performance using Machine Learning

Mujistapha Ahmed Safana

A thesis submitted in partial fulfilment of the requirements of the University of Greenwich for the Degree of Doctor of Philosophy

December 2021

DECLARATION

I certify that the work contained in this thesis, or any part of it, has not been accepted in substance for any previous degree awarded to me or any other person, and is not concurrently being submitted for any other degree other than that of Doctor of Philosophy which has been studied at the University of Greenwich, London, UK.

I also declare that the work contained in this thesis is the result of my own investigations except where otherwise identified and acknowledged by references. I further declare that no aspects of the contents of this thesis are the outcome of any form of research misconduct.

I declare any personal, sensitive or confidential information/data has been removed or participants have been anonymised. I further declare that where any questionnaires, survey answers or other qualitative responses of participants are recorded/included in the appendices, all personal information has been removed or anonymised. Where University forms (such as those from the Research Ethics Committee) have been included in appendices, all handwritten/scanned signatures have been removed.

Student Name: Mujistapha Ahmed Safana.
Student Signature:
Date: 01/01/2022

First Supervisor's Name: Yasmine Arafa
First Supervisor's Signature:
Date: 10/10/2022

Second Supervisor's Name: Jixin Ma.
Second Supervisor's Signature:
Date: 10/10/2022

ACKNOWLEDGEMENT

Profound gratitude to Almighty Allah, the ever-living, master, and maker of all that exists, it's in his infinite mercy and blessings that I have come to the end of this research. May Allah make the research beneficial to me and humanity. I will never do justice in finding words to describe my deepest gratitude to my supervisor Dr Yasmine Arafa whose experience and supervision steered me through this research. I am extremely grateful for her support over the years. A special thanks to my second supervisor Prof Jixin Ma, for his support throughout the research. Finally I thank my parents, family, friends and colleagues for their unconditional support.

ABSTRACT

Blockchain technology has proven to be a secured and reliable technology by bringing security, trust and data integrity to a distributed system. It is a new paradigm that helps in the existence of cryptocurrency and eliminates the third party in a financial transaction. It has the potential to optimise, enhance and streamline many processes outside the cryptocurrency and financial sector but the adoption of the technology is limited by the hindering performance issue. Unfortunately, the current blockchain suffers a performance degrade with the increasing size because of the complexity of its consensus protocol known as Proof-of-Work (PoW). Many industries, researchers and organisation have been working on providing a solution to the performance issues of the technology but most of the proposed solutions has so far ended in proposing a newly designed protocol which ends up facing another issue referred to as the scalability issue; having to trade off one of security or decentralisation to get speed. To address the performance issue, the research has carried out experiments to clear pathways in identifying the specific problem and the outcome has identified the mining process, block size and scalability as the main factors affecting the performance of the technology. The research further investigated these factors and identified the time taken to generate a block as the most time-consuming task within the consensus process, regardless of the traffic, size or number of connected nodes. The research has also explored alternative ways of speeding the nonce finding process and identified machine learning as the best technique because of its ability to learn and predict. Using the quantitative approach of the research, different machine learning models were analysed and compared, and linear regression was identified as the best fit model for the research problem. The research used linear regression model Machine Learning technology to reduce the block generation time without sacrificing security or decentralisation of the proof-of-work consensus protocol. The model has achieved a 58 percent accuracy improvement on the traditional mining process. The model reduces the block generation time when tested on the blockchain simulation by an average of 4 seconds on the Ethereum network and a more significant reduction for the Bitcoin network depending on the computer hardware. In this thesis, blockchain is referred to as the blockchain that uses the PoW consensus protocol.

Contents

1 Introduction **11**

1.1 Motivation 13

1.2 Aims and Objectives 14

1.2.1 Objectives 14

1.3 Research Methodology 15

1.4 Research Questions 16

1.5 Research Hypothesis 17

1.6 Research Contributions 18

1.7 Research Impact 19

1.8 Road Map 19

1.9 Publications 20

2 Literature Review **21**

2.1 Blockchain Review 21

2.1.1 Issues and Challenges 22

2.1.2 Current Solutions/SoA Addressing Performance in Blockchain . 32

2.1.3 Consensus Protocols: 37

2.1.4 Conclusion 44

2.2 Machine Learning Review 46

2.2.1 Review of machine learning techniques 46

2.2.2 Machine Learning and Blockchain 52

2.2.3 Blockchain security attack detection 53

2.2.4 Cryptocurrency and Mining 54

2.2.5 Transaction entity classification 55

2.2.6 Blockchain-enabled machine learning model 56
2.2.7 Blockchain for data security 56
2.2.8 Blockchain for system improvement 57
2.2.9 Transportation 59
2.2.10 Healthcare 59
2.2.11 Supply Chain Systems 60
2.2.12 Blockchain and Machine Learning Issues and Challenges 60
2.2.13 Other Applications and results 61
2.2.14 Conclusion 62
2.3 Conclusion 63

3 Dataset and Performance Analysis 65
3.1 Definition of variables 66
3.2 Data collection 68
3.2.1 Data Downloads 68
3.2.2 Simulation 69
3.2.3 Conclusion 73
3.3 Performance analysis 73
3.3.1 Performance Growth Analyses. 79
3.3.2 Conclusion 80
3.4 Conclusion 80

4 Experiment and Result 82
4.1 Nonce booster Model 82
4.1.1 Evaluation metrics 83
4.1.2 Performance metrics 84
4.2 Machine learning models 85
4.2.1 Linear Regression Model 86
4.2.2 Support Vector Regression (SVR) 92
4.2.3 Random Forest Regression 93
4.2.4 Conclusion 94
4.3 Model Optimisation 95
4.4 Performance Evaluation 99

4.4.1 Model Integrated into Ethereum 100
4.4.2 Model Integrated into Ethereum Bitcoin 103
4.4.3 Performance Evaluation Conclusion 108
4.5 Conclusion 108

5 Conclusion and discussion **111**
5.1 Investigations and findings 111
5.2 Research Novelty and Results 112
5.3 Further Discussions 113
5.4 Limitations and Future Work 115

A Appendix **134**
A.1 Blockchain Architecture 134
A.1.1 Block structure 135
A.1.2 The core component and concept of blockchain 138
A.2 Conclusion 140

List of Figures

1.1 The research flow 15

2.1 Blockchain size from 2010 to 2018 29
2.2 Plasma description 35
2.3 Flow of the mining process 38
2.4 Proof of Stake 40
2.5 Byzantine Fault Tolerance 41
2.6 Proof of Learning Flow 44

3.1 How the Bitcoin dataset downloaded 67
3.2 How Ethereum dataset extracted from the simulation 67
3.3 Sample of the downloaded bitcoin dataset 69
3.4 Sample of the downloaded Ethereum dataset 71
3.5 Larger sample of the downloaded Ethereum dataset 72
3.6 Sample dataset of block collection time 74
3.7 Sample dataset of nonce search time 75
3.8 Sample dataset of block insertion time 76
3.9 Mining distance to the nonce value 77
3.10 Mining distance to the nonce value 78
3.11 The behaviour of the dataset 79

4.1 Ethereum prediction range description 84
4.2 Bitcoin prediction ranges description 85
4.3 The distance of the prediction compared to the mining distance to the nonce value 90

4.4 The distance of the prediction compared to the mining distance to the nonce value 91
4.5 The distance of the prediction compared to the mining distance to the nonce value after optimisation 97
4.6 The distance of the prediction compared to the mining distance to the nonce value after optimisation 98
4.7 How the nonce booster model was implemented in the simulation . . . 100
4.8 Performance improvement as a result of model implementation 101
4.9 The accuracy of the final implementation 102
4.10 Performance improvement as a result of model implementation 103
4.11 How the nonce booster model was implemented in Bitcoin mining . . . 104
4.12 The accuracy of the Bitcoin implementation 105
4.13 Results of the comparison between the bitcoin dataset and model prediction 106
4.14 Results of the comparison between the bitcoin dataset and model prediction 107

A.1 An example of blockchain with a continues growth 135
A.2 Bitcoin block header 136
A.3 Ethereum block header 137
A.4 The main component of blockchain technology 138

List of Tables

2.1 Source: (Dangeti, 2017) 52

3.1 Dataset comparison 73

4.1 Model comparison 94

4.2 Sample data from the Bitcoin implementation 105

GLOSSARY OF TERMS

Term	Definition
Blockchain	In this thesis, blockchain is referred to as the blockchain that uses the PoW consensus protocol.
Performance	Refers to the speed of process within the consensus protocol
Scalability	Scalabilty is the ability for one of decentralisation, security or speed to change without affecting the other
Simulation	A locally running blockchain network that allow source code modification
Difficulty or Difficulty target	Is the value that set how difficult it is to find a block hash. A high difficulty means that it will take more computing power to mine the same number of blocks, making the network more secure against attacks.

Chapter 1

Introduction

Blockchain technology's success with bitcoin boosted the technology into the limelight and brought the rise of other cryptocurrencies and attracted high interest across the different sectors because of its ability to make transactions immutable, secure and transparent. It brought a new paradigm for decentralised systems that are secure and trustworthy in an untrusted ecosystem but its wide adoption has been hindered by some challenges that are primarily security and performance issues (Gao et al., 2018). The concern was raised and intensified by the increased adoption of the technology which leads to an increase in the number of transactions and the added restriction on the block size (Zheng et al., 2017c).

In July 2010, Satoshi Nakamoto added a 1MB limit to the block size of the blockchain and that limits the rate at which information is added to the Blockchain. Thus, it constrains the need to have a finite number of transactions (Puthal et al., 2018). Therefore, the block in the blockchain used by some cryptocurrencies such as Bitcoin and Ethereum is only capable of processing only 7 and 20 transactions per second respectively (Kim et al., 2018) which is not enough to compete with the likes of Visa and PayPal which process approximately about 24,000 and 193 transactions (Bez et al., 2019).

Beyond the cryptocurrency and financial sector, blockchain technology incorporate the right techniques to support a broad range of application across many sectors such as healthcare, manufacturing, distribution and governance. It can enable a controlled

sharing of electronic health records among healthcare providers and allow patients to have full and secure control of their health data (Krawiec et al., 2016). It can help reduce management costs by reducing human error, delays and add efficiency to the process of having agreements between logistic companies (Mendling et al., 2018). It can also improve traceability, transparency and add efficiency to contract management in the telecommunication sector (Al-Jaroodi and Mohamed, 2019). Aside from the cryptocurrency, the technology can also be utilised in other aspects of the financial sector for example; it can remove the third party in the stock exchange and financial settlement, enhance insurance policy and reduce the cost of financial activities in the traditional banking system (Fanning and Centers, 2016).

The current performance of blockchain technology is not efficient for cryptocurrencies because it gets slower with an increase in the network size and the number of transactions (Lu, 2019). More industrial applications or wider adoption of blockchain technology will grow the network and generate more transactions, therefore, making the network performance slower. Thus, there is a need to optimize the techniques and find a lasting solution that will enable adoption from a wider horizon of sectors (Aste et al., 2017).

Many industries, researchers and organisations have been working on providing a solution to the performance issues of the technology but all efforts end in another issue referred to as the scalability issue. The scalability issue means having to trade security or decentralisation in order to get speed (Yu et al., 2018). Sometimes, the solution ends up satisfying a particular use case/application instead of addressing the issue for the whole technology as discussed in section 3.2. Therefore, the research aims at finding a solution to improve the performance of the protocol without facing the scalability issue by reducing the time taken to find the nonce value in the mining process.

The research carried out a study using a series of simulations and experiments to identify alternative techniques that can be used to optimise the performance of the protocol. The study paved way for the idea to optimise the performance using a technique with the ability to learn and predict such as machine learning. Machine learning is the driving force of artificial intelligence today, which has proven capable of speeding up research

and enabling systems across different sectors as seen in image recognition, product recommendations, fraud detection, self-driving cars and many more. The research aims to leverage machine learning ability to learn from experience and make an accurate prediction when fed with data. Again, in this thesis, blockchain is referred to as the blockchain that uses the Proof-of-Work consensus protocol.

1.1 Motivation

Blockchain incorporates various techniques that bring security, trust and data integrity to a distributed system. As a distributed ledger, it requires multiple entities to agree before an action is taken in the system; it is a complex process that takes a lot of time and energy yet effectively achieves the goal on a small-scale blockchain (Mechkaroska et al., 2018a). Unfortunately, the current blockchain suffers a performance degradation with the increase in the transaction rate and the overall size of the blockchain - the more transactions the network processes, the slower it becomes. (Tasatanattakool and Techapanupreeda, 2018). The rapid growth in size and entities can lead to the capability rate of how the network synchronises data not satisfying the transaction throughput. Thus, a transaction can take an unacceptable amount of time to be completed and this can create some problems with the application of the technology both in and outside the financial sector.

Blockchain has enabled the existence of both a smart contract and cryptocurrency. It has the potential to enhance, optimise and streamline many processes in the industrial world but the adoption of the technology is limited by the hindering performance issue that increases with any growth in size (Al-Jaroodi and Mohamed, 2019). In addition, so far, any attempt to address the issue ends up in the scalability trilemma of sacrificing one of the important parts of the blockchain (security, decentralisation or speed). Security and decentralisation bring reliability and trust to the technology, they are part of the pillar attributes of the system.

Therefore, there can't be wider adoption without solving the performance issue otherwise it will only further hinder the performance by increasing the size of the blockchain. Thus, to utilise blockchain technology across sectors such as healthcare, manufacturing,

the internet of things (IoT) (Dorri et al., 2016) and others, it is important to find a lasting solution that is scalable, reliable, secure and efficient in addressing the issues from multiple perspectives (Yaga et al., 2018).

1.2 Aims and Objectives

The research aim is to improve the performance of the proof-of-work consensus protocol without trading off security or decentralisation using machine learning. Preliminary research supports the idea that using machine learning to find the nonce value will reduce the transaction throughput and allow applications to process more transactions per second and accelerate adoption within application for which the slow and decreasing performance is a bottleneck.

1.2.1 Objectives

1. Based on literature and theoretical evidence, identify the main factor affecting the performance of the consensus protocol.

2. Practically examine these theoretical findings and establish facts of the actual problem.

3. Explore alternative novel solution to improve the performance of the protocol that advances the current state of the art.

4. Implement and test the identified solution in a blockchain simulation with real time effect.

5. Validation of the research solution approach by peers and experts in the research area via paper publication.

6. Evaluate the performance of the improved consensus protocol and complete thesis.

1.3 Research Methodology

The research used the quantitative research methodology approach to study and find a solution to the research question. By quantitative, the research experiment statistically compared the result of different machine learning models within the criteria of their prediction accuracy and prediction time as the matrix of choosing the best fit model for the research implementation. The approach was adopted in the research evaluation stage, and the performance of the research implementation was statistically compared with the current performance of the protocol to critically examine the impact of using machine learning techniques to improve the performance of the consensus protocol. The flowchart in figure 1.1 describes the research approach in stages.

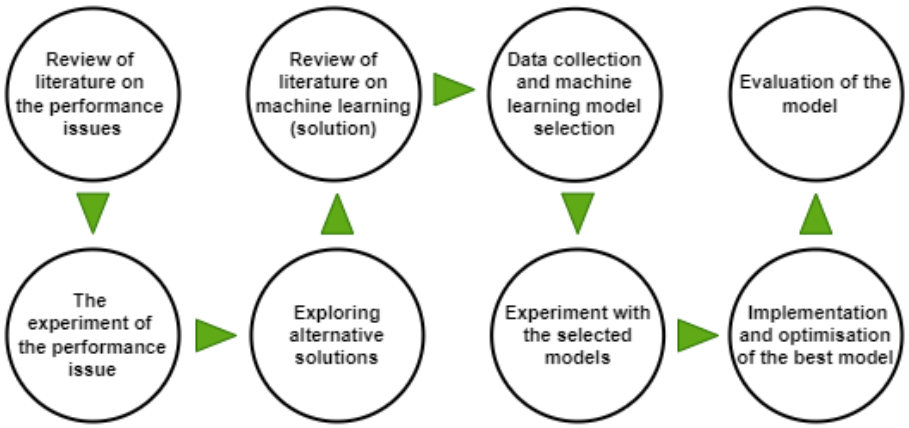

Figure 1.1: The research flow

As shown in figure 1.1, the research started with an empirical investigation of the research area which was carried out through literature review and experiments to create a base for knowledge and provides insights into the general idea of blockchain technology. It provided insight into the topic and cleared a pathway to understand the scope of the problem that the research is trying to address. The research investigation was narrowed down to focus on the issues and challenges around the performance of the technology and finally investigated the work done by others (other consensus protocols and solutions) to address the issues. The process laid a solid foundation of knowledge and develop ideas that put the research closer to its goal.

In the second stage of the research, the Blockchain network was simulated to enable smooth experiments executions away from the busy and complex real public blockchain network. In doing so, a working blockchain network that can allow playing around with the parameters is required. Therefore, a set of platforms that use PoW were analysed and Ethereum was chosen because it is open source and provides the required complexity and flexibility to conduct the research experiments. Ethereum network was simulated and used in testing, evaluating and generating facts on the research finding which helps in building research contribution ideas.

In the third stage, ideas were explored as an alternative to the problematic current approach that hinders performance and the machine learning technique was found as the perfect fit technique to solve the research problem. Therefore, in the next stage (fourth), a review of machine learning was conducted to enable a solid understanding of the technique and enable the successful implementation of the research idea.

In the fifth stage, data were collected and analysed as a requirement for a successful implementation of machine learning. The result was able to suggest the right distribution for the research problem. All relevant models were implemented in the sixth stage as elaborated on later in the experiment chapter and the results were quantitatively compared to enable the selection of the best fit model.

The seventh stage includes implementing the research solution into the simulation environment and the larger dataset collected until this stage was used to optimise and validate the model. Finally, the research was evaluated in the final stage and the research was successful in optimising the performance of the protocol. All stages were elaborated on in the subsequent section

1.4 Research Questions

The research was carried out in the path that will answer the following questions:

- The preliminary research question for the research asks what are the issues affecting the performance and scalability of blackchin technology? After investigating the question and identifying the issues and their factors, the research successfully

discussed and answered this question in detail later in the literature review and experiment chapters.

- What are the right methods of improving performance without altering or modifying any of the protocols?
  - Could speeding up the nonce search improve the performance without facing the scalability issue? To answer this question, there is a need to identify the right technique that can be used.
  - What is the right technique that can be used to speed up the nonce search process? The proposed idea suggested using prediction - Machine Learning.
- Could applying machine learning techniques improve the performance of the proof-of-work consensus protocol without sacrificing security or decentralisation?
  - What is the most appropriate machine learning model for the research use case? The research found Linear regression as the best fit model as discussed later in both the performance analysis and experiment chapters.

The research finds the sub-questions necessary and helpful in unpacking the complexity of the work and setting direction and focus for the research investigation.

1.5 Research Hypothesis

To understand the rationale that supports the research hypothesis it is important to understand the following points. For the first hypothesis, there was no concern about the performance of the technology until the increased adoption of the technology that led to an increase in the size of the network - increasing the complexity of the mining protocol (Zheng et al., 2017c). The research investigated the complexity and identified that complexity only affects the block generation time. The complexity introducing the concerns around performance which later led to the scalability issue that validates the first hypothesis. The second hypothesis can be supported by the machine learning

[1]Hypothesis is tested by observation, as opposed to via statistical analysis.

track record of speeding up processes using its ability to predict a value, for example, Bitcoin price Demir et al. (2019).

The research has the following alternative hypothesises:

- Reducing the block generation time improves the transaction throughput and the overall performance of the protocol.

- Reducing the nonce search space improves the block generation time without effect on the security or decentralisation

1.6 Research Contributions

By addressing the research questions, the work produced the following main contributions:

- A novel approach that used regression and reduced the block generation time without sacrificing security or decentralisation and ultimately improves the overall performance of the proof-of-work consensus protocol.

- A novel approach that introduced Machine Learning technology into the blockchain block generation process of blockchain technology.

- Scaled solution that is decentralised, secured and faster - Improves the block generation time without altering any other part of the protocol allowing improved performance without affecting the current level of security or decentralisation.

- Improved transaction confirmation time of the blockchain technology: Faster block generation time means that transactions/data will be offloaded from the transaction pool to the blockchain faster and the faster transactions are added into the blockchain the faster the confirmation time

- Save the cost of transaction fees: Improved confirmation time will reduce and eliminate for some people the need to pay extra for the transaction to be given priority. A reduced waiting time reduces the amount to pay for those that still

need their transactions to be confirmed faster.

- Reduced the power consumption/waste of the mining algorithm: Taking the miners closer to the nonce value has reduced the unused computations that take a high amount of processing power.

- Reduce adoption concern in other sectors: The long transaction confirmation time is one of the main concerns hindering the adoption of the technology, especially in those areas where prompt data processing is key such as healthcare emergencies, immigration and others.

1.7 Research Impact

This research work is of theoretical and empirical significance to the scientific community. The theoretical importance of this study lies in demonstrating that blockchain technology possesses the ability to keep anonymous and immutable data across sectors beyond the financial industry without the need for central monitoring. This study has demonstrated the importance and the right approach to scale blockchain technology and enable firm performance that will suit a wider horizon of sectors.

Empirically, the study has safely enhanced the performance of blockchain technology by reducing the block generation time without sacrificing security or decentralisation. By scaling the block generation time, the study has reduced the task of the mining process which is time and power-consuming. According to a report published by the International Energy Agency, the overall power consumption of the Bitcoin network is higher than in many countries (Monrat et al., 2019). Thus, the research has practically optimised the performance and cut down the power consumption.

1.8 Road Map

Chapter 1 is the introductory chapter, it briefly highlights the research motivation, research questions and contribution and discussed the methodology followed in conducting the research. Chapter 2 is the literature review chapter, it states reviews on

articles, journals and books that discussed the research area in general and later narrowed down to focus specifically on the concerns around the performance issues. It also covers the literature review on machine learning as the technique the research is using to find the solution to the problem. Chapter 3 is data collection and processing, it explains how data was collected and analysed to support research decisions and enable the implementation of the machine learning model. Chapter 4 is the experiment chapter, it discusses the research experiment, compared results and explains how the model was optimised. The chapter also covers the performance evaluation of the model. Chapter 5 restates the research question and how they were addressed.

1.9 Publications

Ahmed, M, S., Arafa, Y. and Ma, J., 2020, November. Improving the performance of the Proof-of-Work Consensus Protocol Using Machine learning. In Proceedings of the 2nd International Conference on Blockchain Computing and Applications (BCCA). IEEE. The paper is an extract of this thesis that presents the research as a novel approach and the initial results as a proof of concept.

Ahmed, M, S., Arafa, Y. and Ma, J., (in writing). Improved proof of work consensus performance using machine learning linear regression model. IEEE Access,

Chapter 2

Literature Review

2.1 Blockchain Review

In 1991 Stornetta and Haber (1991) brought the idea of certifying the time a document was generated in a way infeasible for a user to alter using any tool or service. The proposal came with two schemes that use hash functions. The first one is called the linking solution; It requires chaining together the hash values of all documents sent to the timestamp service in a way nothing can feasibly tamper with. The second one is called a random witness solution. It requires a certain number of users to date and sign the document and use the composite of their signatures as the certificate (Stornetta and Haber, 1991). These ideas provide a way to keep an immutable date of documents, enhancing the credibility of the documents. Bayer et al. (1992) later worked to develop the idea and were able to add a Merkle tree to the design by the year 1992 which provides the ability to collect sets of data into one block, adds efficiency and provides a secured way of data verification.

The idea was enhanced by a presumed pseudonymous called Satoshi Nakamoto to build the blockchain that serves as a peer-to-peer P2P protocol for Bitcoin (Nakamoto, 2008). Bitcoin is a decentralised digital currency referred to as cryptocurrency: a digital banking system that doesn't require any central monitoring of individuals or organisations but rather a set of nodes (computers) located in different locations across the world.

It uses a consensus approach between all connected nodes to perform, verify and store transactions in an efficient, transparent, reliable and secure manner. The storage capability of the blockchain technology is used to store sensitive information that does not in any way tolerate alteration (e.g financial transactions) because the blocks are arranged in an immutable growing sequel structure that cannot practically be altered.

Blockchain technology enhances data integrity through distributed structure: A decentralised structure that requires all records of transactions to be replicated across all participating nodes and all nodes have to come to a consensus before any data is added to the blockchain. This helps build trust between the participant of the network and enable the elimination of central monitory. A transaction cannot be approved by a single authority or a central monitory nor can any of them set a specific rule of accepting transactions instead they get processed, verified, approved and recorded through the consensus process (Yli-Huumo et al., 2016).

For a secure and effective blockchain or any other decentralised system, it is key to consider an effective and suitable way of securing the flow of data across its different repositories. Blockchain consensus is the process used in verifying the state of the network, it checks and agrees all blocks keep the same information with correct values across all nodes. Bitcoin blockchain uses the consensus algorithm called proof of work (PoW) that requires at least 51 per cent control of the whole network to enable manipulating the network. That is practically not feasible depending on the number of connected nodes and because the attacker will have to attack all nodes simultaneously.

2.1.1 Issues and Challenges

This section will discuss the numerous challenges encountered through blockchain technology by many organizations within a few numbers of varied contexts. The discovery made by Al-Saqaf and Seidler (2017) asserts that the lack of principles and interoperability within the open domain is a major factor that deters the widespread acceptance of the blockchain initiative. Alketbi et al. (2018) claimed that even though blockchain proposes to outrun many security hurdles that include data integrity and secure data

sharing, in more ways it also offers new security challenges to be exclusively analyzed and confronted. Atlam et al. (2018) buttressed a number of blockchain integration challenges that include scalability, legal and lack of skills: Scalability is the trade of security or decentralisation to get speed, the legal issue that comes in many forms including compliance with financial services regulation or jurisdiction boundaries as nodes can be located anywhere in the world. Boulos et al. (2018) claimed that blockchain encounters normal challenges just like any similar technology threatening to destabilize existing processes and highlighted many of its challenges inclusive of interoperability, privacy and security and also the urge to sustain suitable and adaptable business models for execution. Dorri et al. (2017) argued that the adoption of blockchain within the Internet of Things (IoT) schemes and discussed main challenges in the likes of computational overhead and time duration in the mining of blocks, inadequate scaling of nodes and important traffic load in case there is an increase in the number of nodes in the network.

Lacity (2018) illustrated the challenges of blockchain technology in the aspects of stability, interoperability and performance with similar systems. Additionally, this literature will also highlight the challenges concerning the management of blockchain applications that include standards, shared governance, regulation and building a reliable atmosphere that encourages progress. Mendling et al. (2018) illustrated some series of technological challenges blockchain constantly encounters. The illustrated challenges include throughput, size and bandwidth, latency, limited usability, security and misused resources. Also, Salman et al. (2019) discussed some challenges of blockchain adoption that include the capacity of storage and scalability, anonymity and data privacy, and security issues.

In addition, Mingxiao et al. (2017) claimed that blockchain is still fresh as it is just emerging and it is prone to face several issues that involve throughput (i.e., transactions are optimal numerically by seven theoretically per second), size and bandwidth, latency in the terms of duration of the period of time to access the blockchain. Zheng et al. (2017b) insisted that blockchain is going through multifaceted challenges and recapitulated three common challenges of this technology that includes privacy leakage, scalability, and self-centred mining. Self-centred mining is known as an approach for

mining bitcoin in which some miners organize to increase their return by building their personal branch of the blockchain.

2.1.1.1 Performance

Blockchain technology isn't fast enough for the cryptocurrency to satisfy the need for the mainstream payment method such as VISA (Vasin, 2014). Gao and his colleagues considered performance and security as the wider application of blockchain technology and concluded by indicating the need to improve the performance of the blockchain for the technology to be competitive with the traditional software implementation (Gao et al., 2018). In their discussion of the emerging concerns in blockchain technology, they identified scalability and availability as part of the contributing factors in the performance issue where they mentioned the block size and transaction throughput to be the key players.

Throughput

In his literature Swan (2015) had expressed seven technical future challenges and boundaries of the Bitcoin technology. A prominent part of the issue is Throughput. The issue with Bitcoin as of 2022, it handles an average of 4 transactions per second (tps) (Xu et al., 2016), with a maximum possible theoretical throughput of 7 tps. In contrast, the VISA transaction network is believed to process up to 24,000 tps (Xu et al., 2016).

A block is generated in the public blockchain at the rate of 1 in every 10 minutes and a transaction can only be termed confirmed if it is added to a block that is included in the blockchain (Yli-Huumo et al., 2016). Blocks are where data is stored in the blockchain, so a limited block size (1MB) confines the amount/capacity of data that can be added to the blockchain. Therefore, Bitcoin can only append to the block about 7 transactions with a block creation time of 1 in 10 minutes (Kim et al., 2018). Looking at things from the smart contract perception, the scalability issues seemed to have been downsized. The transaction speed depends partially on your gas usage instead of factors such as consensus, transaction validation and others. One Ethereum block using a smart contract facility takes about 15 seconds. Averagely for such a transaction, the total wait time of a transaction takes about 15/2 seconds for a complete block to be mined. Mostly the amount of time a transaction takes to be completed in a smart contract

Scherer (2017).

In a typical smart contract blockchain analogy, to both compensate nodes for a transaction executed and to reduce computation, a fee is charged in relation to the proportion of computation used in the blockchain which is known as gas (Wood, 2017). The more gas the more it costs to run. The gas limit and block time determine the basic throughput speed as mentioned by Scherer (2017). Even though with such a mechanism the slow speed of transactions implies that smart contracts aren't completely suitable for e-commerce applications now. It's either the high cost to pay of computation power or the slow nature of the transaction, all the mentioned constraints seem to render the scalability of blockchain to be questioned (Tonelli et al., 2019).

Mrs Chauhan and his colleagues stated in their article, there is a daily average of 130,000 bitcoin transactions in 2017, thus, the transaction waiting time has increased to 29 minutes (Chauhan et al., 2018). While the transaction increase looks good indicating the acceptance of the system, the waiting time increase is not looking good for a technology/system that is to compete with the likes of Visa and PayPal. Expect more increases in the waiting time as transaction increases, this is not because they are proportional, they are not. It is because Bitcoin can only process 7 transactions per second (Conti et al., 2018). This is far from satisfactory compared to the likes of Visa and PayPal which are capable of processing 24,000 and 193 transactions per second respectively (Chauhan et al., 2018). Though the numbers are slightly better with Ethereum as it can process 20 transactions per second as of 2021, that is still not enough to compete with the likes of Visa and PayPal (Bez et al., 2019). According to Chauhan et al. (2018), Ethereum is supposed to process about 1000 transactions per second in theory. However, unfortunately, it can only practically process about 20 because of the enforced gas limit which is the price for running a transaction that is based on the computational effort.

Block Size

A block is only capable of taking a limited amount of data after Nakamoto appended a 1MB block limit which according to Mechkaroska et al. (2018b) is believed to be a security measure. The block size has attracted so much argument standing for or

against it, most of them against believing that increase in the block size will lead to centralisation which will make mining more expensive and leads to having fewer miners running full nodes and having more power on the network (Mechkaroska et al., 2018b). According to ETH Zurich et al. (2016) research findings, increasing the current bitcoin size up to 4 MB does not have a significant impact on security especially selfish mining and double-spending resilience. Chaudhry and Yousaf (2018) mentioned the security of the blockchain is not penalised with the current block size of 1 MB and a block generation interval of 1 minute. The point indicates there will not be a significant issue if the time taken by the miners to generate a new block will be optimised to 1 minute from the current situation of 10 minutes.

According to Monrat et al. (2019) the block size slows the transaction process. Kim et al. (2018) believes the increase in the number of transactions affects the technology by increasing the number of transaction waiting times: the time taken before a transaction is added to the blockchain. According to Gao et al. (2018) and Mechkaroska et al. (2018b) the public blockchain Bitcoin has a restriction on the block size in order to leverage the security of the blockchain consensus but Zheng et al. (2017a) indicated larger blocks will lead to slower propagation in the blockchain network, which will gradually violate the decentralisation principle of the technology.

Lin and Qiang identified the transaction throughput issue and how it is affected by the block size, yet increasing the block size arbitrarily cannot be the best solution because of how the larger blocks can impact the performance of the blockchain (Lin and Qiang, 2019). They made a point that increasing the block size will not increase the complexity of the system and it will be easy to implement. However, it increases the risk of forking because the change cannot be effective on the old nodes and the process of verifying and synchronising blocks will take longer.

2.1.1.2 Scalability Issue

A major part of the challenging problems concerning Blockchain implementations is scalability. In order to gain access to the theoretically proven security, the Blockchain adaptation must have a reasonable number of full nodes. Otherwise, its implementation might result in a less decentralised system (Beck et al., 2016). The limits of scalability

of the Blockchain are related to the size of the data on the Blockchain, the rate of processing transactions, and the latency of the transmission of data. According to Xin et al. (2017), the discussion on the blockchain scalability challenge has been vigorous and at times acrimonious with no clear method to be deployed in addressing the issue. Yli-Huumo et al. (2016) a systematic review on blockchain technology and the authors stated there will be a direct impact on the scalability when the size of the Blockchain increases. Thus, for the technology to be ready for pervasive use, scalability issues such as performance and latency have to be addressed.

According to Chaudhry and Yousaf (2018), scalability is the fundamental requirement in blockchain technology to deal with big data in today's environment and PoW is not scalable enough. In the paper titled "Blockchain scalability" written by Chauhan et al. (2018), the authors mentioned there is a propositional relationship between a fall in scalability and an increase in the network size. They referred to the miners as the bottleneck of the scalability issue because they are given the task of processing and verification of every transaction that occurs in the system, the faster the process the faster the finality of the blocks. Also briefly highlighted the need to optimise the speed of performing a transaction in blockchain technology, showing a big difference compared to its counterparts such as VISA. Lin and Qiang (2019) paper has indicated that the issue on the blockchain is not only on scalability but also on the technical standardisation that is dynamic enough to fit different sectors and their requirements.

Many interesting solutions are coming up in order to address the scalability issue. This issue includes Lightning Network, which involves the addition of an additional layer to the standing blockchain network just to hasten faster transactions, in addition, the interesting solution is Sharding which relates subsets of nodes into their limited networks or 'shards', which are then solely responsible for the transactions that are related to their shard. We will elaborate on the solutions in the state-of-the-art section.

2.1.1.3 Blockchain Storage

The total size of the Bitcoin network hits 197GB at the end of 2018. This is a growing size of data whose copy is expected to be stored on all connected nodes (Palai et al., 2018). Excluding the few that might have been foreseen, most of the identified or

worrying issues of the technology a raised after the technology has stored and dealt with a large amount of data. With less data, the technology has been satisfying. Therefore, some researchers believe reducing the storage alone can be enough to scale the technology (Zheng et al., 2017c).

Zheng et al. (2017c) discussed storage optimisation and redesigning the blockchain as the best solution to the scalability issue of the technology. Storage optimisation: Was proposed by J.D. Bruce (2017) as a scheme that requires the old transactions to be forgotten by the network and a consensus algorithm called Proof-Chain to provide the security loss. Bruce argues to have achieved scalability at the expense of some security trade-offs that can be dealt with.

2.1.1.4 Other Issues

Organisations such as IBM, Microsoft and Amazon have been exploring possible methods to cater to the cost and complexity issues of blockchain technology through the introduction of cloud technology (Karame, 2016). The adoption of a blockchain in cloud-based technology as proposed by these organizations is going to adopt a blockchain-as-a-service. This service will allow organizations to set up and run their blockchain network making the cost, as well as the complexity of blockchain service responsibilities, will depend entirely on the host organization. According to the diagram in figure 2.1 from Statista, blockchain was 185GB in size at the end of September 2018, the figure shows how the size of the technology has been increasing since the early days of the technology. The y-axis shows the blockchain size in megabytes while the x-axis shows the time from 2010 to 2018 in three quarters.

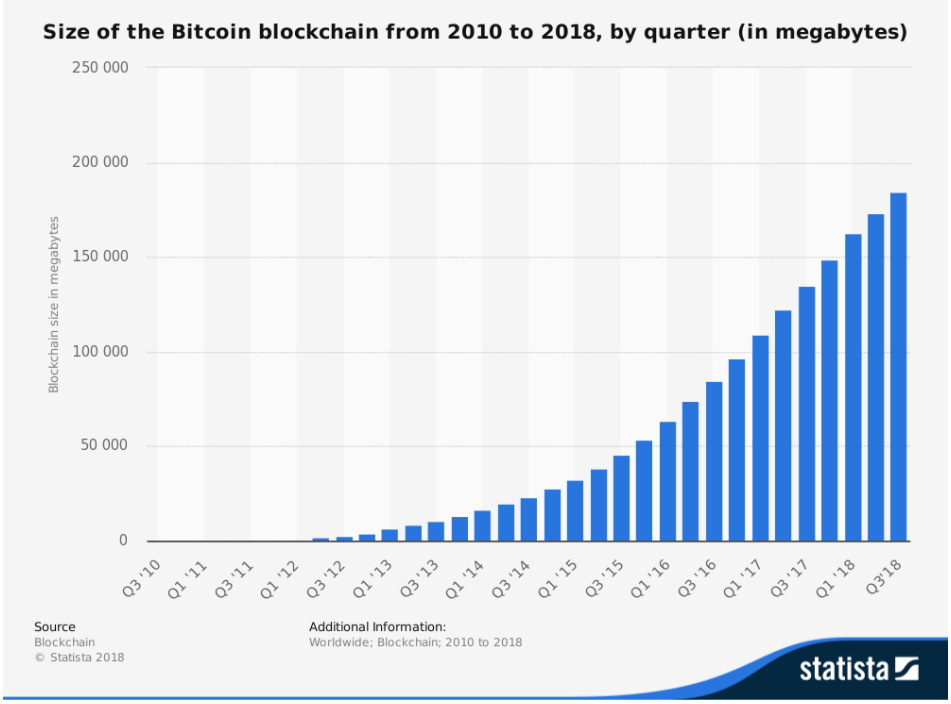

Figure 2.1: Blockchain size from 2010 to 2018

(statista, 2018)

Latency issues, the time factor is a determinant of the critical issues in Blockchain executions. The bitcoin requests are being processed online almost immediately, and that in a way stands as an obstacle in regard to universal technological acceptance Beck et al. (2016). For security's sake concerning bitcoin, any transaction related to its medium is advised to be completed in a space of about 10 minutes. For an enormous transfer amount, the cost of a double-spent attack can be up to an hour. The completion of the VISA transaction process only takes a few seconds (Swan, 2015) (Yli-Huumo et al., 2016). With the decentralization and distribution of public ledgers, there seems to be a massive number of nodes that needs to attain a consensus can be achieved. Every node needs access to the whole blockchain. This would amount to a great deal of database with time. Getting access to the entire blockchain to a large number of nodes will also increase the security threat. The resultant effect is that there are thousands more transactions waiting in the queue than the ones that are being verified every minute. Many ecosystems have made efforts to mitigate this issue by only partially decongesting their distributed ledger, where only a sizable number of prominent nodes reach consensus.

Cost The price tag attached to Bitcoin is a drawback to its decentralization and blockchain technology. The users of Bitcoin have to pay for their transactions and computational power. The fact that users have the option to choose centralized solutions is that they have to be constantly reminded that their actions are attached to a fee, but the prices are not openly displayed (Beck et al., 2016).

Integration with Legacy Systems There is the issue of corporations on how to integrate blockchain with their institutional system(s). Most likely, if they decide to make use of blockchain, establishments are needed to wholly restructure their former system, or organize a pattern for the successful integration of the two technologies (Atlam et al., 2018). Another major issue is that due to the lack of enlightened developers, establishments do not have access to the needed pool of blockchain ideologies to infuse in this process. Reliance on a foreign party can lessen the hurdles (Dorri et al., 2017). Most present solutions on the market need the organization to invest a tangible amount of time and materials to accomplish the transition. There are more incidences of data loss and breaches that are disengaging most establishments from transitioning to the blockchain. All enterprise is reserved and unwilling to adjust to their database, and for reasonable reasons, data loss or data dilapidation possess a major risk (Lin and Liao, 2017). A few years ago, innovations came up that enabled legacy systems to integrate with a blockchain backend. One such solution is the Modex Blockchain Database, an innovation structured to assist people without a background in technology, gaining access to the benefits of blockchain technology and eliminating the challenges presented by losing sensitive data (Janowski, 2015).

Security and Privacy Challenges Blockchain is the enabler of crypto and it is mainly perceived as an arena for bad actors, frauds, hackers, and speculators. More significantly are issues such as immaturity (still slow and tedious), lack of scalability, the inefficiency of interoperability, stand-alone projects, difficulty in integration with legacy systems, complications and lack of blockchain initiatives. An issue with the open distributed ledgers is the highly speculative character with a trade-off between the elements of the network and decentralization (Atzei et al., 2017). The possibility of attacks by malicious actors on Bitcoin is they have to control 51 percent of the network. In such a case, a miner can have total control of the majority of the network which

seems to be a big problem (Lin and Liao, 2017). It was highlighted that several security breaches that happened in Bitcoin include distributed denial-of-service (DDoS) attacks, hacking of accounts by Trojan horses and viruses contacted through ads. According to Mingxiao et al. (2017), a total of 11 million USD had been forfeited to scams by about 13,000 Bitcoin victims between September 2013 to September 2014.

Privacy issues: A major issue that is connected to Bitcoin privacy is the issue of multiple addresses. For instance, users of the Bitcoin system have the opportunity to create numerous addresses and researchers try to collect all the addresses that relate to a single user (Boulos et al., 2018). Address collection is done to trace the economic transactions of the same user. The aim is to discover all the possible addresses involved in the transaction that is linked to the targeted user (Sullivan and Burger, 2017). The researchers discovered that some of the Bitcoin addresses could be traced through the IP addresses by examining the traffic of the transaction.

Double-spending attacks: Blockchain operations are susceptible to double-spending attacks. Concerning the case of Bitcoin, a double-spending attack may occur when an intruder keeps his/her Bitcoin while getting services that can be expended again. This happens when the intruder credits an account, receives goods or services through the account holder and later rearranges the ledger by reverting the transaction that pays into the account. Nakamoto (2008) insisted that the Bitcoin system is hampering double-spending attacks by showcasing the attacker and the group of honest players as competing users that are performing a random activity moving towards a single direction with probabilistic steps. Though, it is not certain that in Bitcoin's decentralized atmosphere the intruder may try to introduce disagreement between the sincere miners Raju et al. (2017).

2.1.1.5 Conclusion

The section discussed many concerns about blockchain technology from different works of literature. These concerns are mostly around the amount of data the technology can take at a given time referred to as the throughput and the block size issues. This shows that there is a gap in this technology that calls for a solution and is required to allow the utilisation of the full potential of the technology both in the financial and other

sectors looking to adopt this tremendous technology. The review also shows that the concerns highlighted in the research problem and motivation are shared among other researchers, experts and other parties interested in the technology.

The research finds the performance issue as a key area of concern in the challenges discussed because it affects the efficiency of the technology and happens to be a direct or indirect factor in most of the other issues excluding the blockchain size. Therefore, it serves as a solid ground that supports and validates the need and importance to improve the performance of the technology. The review in the next section is narrowed down to focus on other proposed solutions to address the performance issue and the impact of the scalability issue.

2.1.2 Current Solutions/SoA Addressing Performance in Blockchain

In this section, we will discuss the work done by others to address the issues mentioned in the previous section and how the solutions are affected by the scalability challenge. Blockchain technology remains the most decentralised and secured information processing system with its great attributes that can be utilised beyond Bitcoin and the financial sector. The adoption is still however facing issues such as performance and scalability. Researchers have taken different approaches to solving the challenge for different use cases. The approaches can be categorised into On-chain (first layer), Off-chain (second layer), Side-chain and Child-chain (Chauhan et al., 2018).

On-chain or first layer solution: require making a change to the actual blockchain without altering any of the main features or characteristics of the technology. Off-chain or second-layer solution: is an additional protocol built on top of the main chain to enable performing transactions without adding much congestion to the network. Side-chain: approach is to enable communication between blockchains. Child-chain: uses the parent-child approach, where a transaction is recorded in the parent chain after being processed in the child chain (Kim et al., 2018).

Blockchains using PoW are still facing hindering performance issues and any attempt to improve that value ends in trading off security or decentralisation (Khan and Miˇ, 2018). The block size is still limited to 1MB which leaves the transaction pool getting

flooded with unconfirmed transactions, and increasing the size will require all participating miners to reach a consensus before they can adopt the new size (Mechkaroska et al., 2018a). The section will discuss work done by the financial industries and other interesting parties trying to address the scalability and performance issues around the Proof of Work consensus protocol. The discussion will critically analyse the approach based on how it satisfies security, decentralisation and speed.

2.1.2.1 On-chain/first-layer solution:

Sharding (ETHEREUM): Sharding simply means partitioning, it is a solution suggested by Ethereum developers in their quest for a scalable blockchain. The idea suggests dividing the blockchain into a partition of shards with each shard having a different state, therefore, different history. Nodes will be grouped into shards and each shard is capable of processing a set of transactions and at the same time updating the state of the blockchain. The mining process is divided into the partition, and each node will focus on mining its partition instead of all nodes mining the whole data (Luu et al., 2016).

One of the backbone reasons for the PoW algorithm succeeding security-wise is the ability to maintain a single state across all nodes on the network, and a successful attack requires at least 51 percent of the whole network. Dividing the workload on the network will improve the transaction processing time because there will be fewer amounts of data to verify and validate when processing and confirming transactions (Dang et al., 2019). It also makes the network less secure because a successful shard attack can give an attacker the monopoly to control data if the attacker succeeds in taking over the majority of the collators in the shard (Chauhan et al., 2018).

Sregated Witness (SEGWIT): The idea was proposed by the bitcoin developer Dr Pieter Wuille. He suggests increasing the number of transactions in the Bitcoin block by removing and storing the signature data outside the transaction block, allowing more transactions to fit in a block. The signature data takes more than 60 percentage space in a single transaction and removing it has enabled processing 1.7 to 4 times more transactions than before (Kim et al., 2018). SEGWIT has been added to the Bitcoin protocol in August 2017 Mechkaroska et al. (2018b). The solution has improved the

transaction throughput without any scalability issues but the rate is still not enough to satisfy the performance requirement.

Hard Fork: Hard fork is an update in the network protocol that validate the previous invalid blocks and transactions, it makes the previous version of the software incompatible. There is a new coin called Bitcoin Cash (BCH) that has the same structure as the normal Bitcoin but with a larger block size of about 8MB. The coin is expected to have a throughput of 60+ transactions per second (Kwon et al., 2019). Another example is Litecoin, designed to reduce block generation time. Its block generation time is 4 times better than bitcoin thus, has a faster transaction speed of about 56 transactions per second (Clarke et al., 2018). It is still not enough to satisfy the performance requirement.

2.1.2.2 Off-chain/Second-layer solution:

There are dozen of new application solutions that are faster than the bitcoin network as of the time of writing, especially with the rapid growth of the technology world and the potential of blockchain technology. There are the likes of oracle blockchains that provide blockchain as a service to ease and enable blockchain adoption. This section will concentrate on the solutions that are believed to be motivated by the proof-of-work challenges and relevant in the context of solving the scalability issue.

1. *Lighting Network:*

   The lighting network solution proposed addressing the scalability issue by adding an extra layer to the blockchain and introducing channels to be used in performing a transaction. It enables users to create a payment channel that requires only the participating parties when validating a transaction (Chauhan et al., 2018). The required number of nodes to participate in the consensus on the traditional blockchain is equal to the number of participants in the network which can be any number depending on the connected nodes. On the lightning network, the number of participating nodes equals the number of participants. It uses native smart-contract scripting to store transaction information on the main chain after it is conducted off-chain. Thus, it is very fast promising a transaction speed of

millions of transactions per section surpassing Visa's ability of 45,000 transactions per second (Poon and Dryja, 2015). Should it be achieved, it will be the fastest financial transaction on the globe. The solution did not technically solve the blockchain's challenge rather it solved the Bitcoin's or cryptocurrency challenge by proving a new way of performing transactions away from the blockchain complexity and issues.

2. *Plasma:*

Plasma is an off-chain solution for the Ethereum blockchain network that focuses on a scalable autonomous smart contract which proposed pushing the computation process off-chain. Like the lightning network approach addressing the issue from the bitcoin angle, Plasma is doing similar to the smart contract perspective. It uses a child chain on top of the main blockchain refers to as a root chain that allows having multiple side chains with each having its business logic. The plasma does the data computing and process before passing only the block header to the root chain where the state of the blockchain is stored (Bez et al., 2019). Figure 2.2 depicts a small architecture of how it works. The Root Chain is the main blockchain that all block headers are sent to after processing in other shown chains.

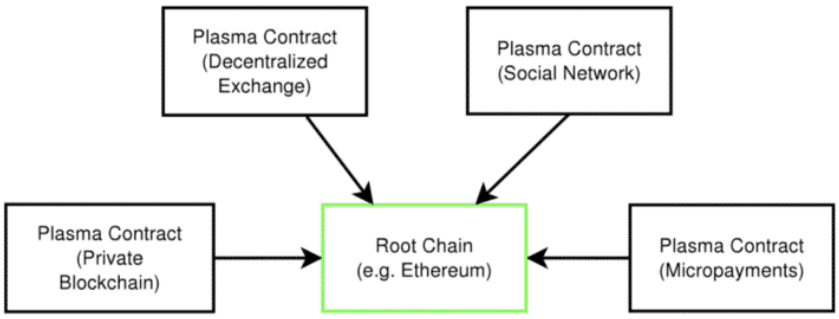

Figure 2.2: Plasma description

Source (Rosic, 2020)

A plasma chain can be custom created by anybody for a variety of reasons, but the state transition is enforced by fraud-proof. If fraud is identified by proof of fraud when submitting a block header to the blockchain that block will be rolled back

(Poon and Buterin, 2017). Security-wise, it is safe to rely on the root chain which is just relying on the current blockchain to keep the smart contract. It is another solution that is addressing the challenges for a particular use case (Ethereum) instead of solving it for the whole technology.

3. **Raiden Network:**

   The Raiden solution is referred to as the etheruem version of the lightning network. It works just like the lighting network, it also includes the two-way channel that is a big factor in the lighting network. The only difference comes in the Bitcoin and Etheruem way of function which is, Bitcoin is strictly for financial transactions (Cryptocurrency) and Etheruem can be used for smart contracts (Khalil and Gervais). Nevertheless, the solution cannot be generalised because it provides an avenue to perform a transaction away from the blockchain in a way that suits a particular cryptocurrency (Ethereum) and might not apply to another cryptocurrency not to mention sectors.

4. **Trinity:**

   Trinity is a second layer solution built on the NEO blockchain, it aims to enable real-time transactions with a lower fee and reduce the traffic on the main chain as explained in (David Yiling Li, 2018). It uses the state channel approach and adopts zero-knowledge proof to protect data and improve privacy. All settlement amounts are broadcast to the entire network to ensure trust and decentralisation. Trinity does for NEO assets what the Lighting Network is doing for Bitcoin and Raiden Network is doing for Ethereum Ramachandran et al. (2018). It also solves the challenges for its use case, not the technology.

2.1.2.3 Conclusion

The section discussed other solutions proposed and implemented to address the concerns around the technology. On-chain solution sharding proposed by the Ethereum developers speeds up transaction processing time but lessens the security because an attacker has higher chances of successfully attacking a partition than attacking the whole blockchain. Other solutions have been successfully implemented but do not cer-

tify the performance requirement especially when future network growth is considered. Off-chain solutions review came across many prospect ideas and solutions that will very well certify the blockchain performance requirement for some particular use cases or types of use cases but not the blockchain as a technology. Sidechain is a solution that uses the normal blockchain in its current form in the background and its proposed side-chains do not conform with the decentralised idea of the blockchain that enforces the need to keep the same state across all connected nodes. DPoS used as the consensus protocol also has scalability issues as discussed in the next section.

2.1.3 Consensus Protocols:

The solution to the performance issue of blockchain technology may be addressed by looking into the various protocols. This allows gaining insight and quantitative analysis of other protocols proposed as a solution to the performance and scalability issue. The research interest and comparison parameters are the decentralisation, security and performance of the protocols.

2.1.3.1 Proof of Work Protocol

The proof of Work algorithm is a consensus protocol used in Bitcoin and Ethereum blockchain in producing new blocks in the chain and bringing all nodes in agreement, validating and verifying transactions while securing the entire network against malicious attacks (Yaga et al., 2018). It solves a complex mathematical puzzle in order to validate transactions, generate a new block and link it to the blockchain. One of the main attributes that made the protocol succeed is the fact that it is hard to find a solution for a given problem but easy to verify if a solution is correct. There are other consensus protocols such as Proof-of-Stake, Proof-of-Vote but the research is trying to address the performance issue of the blockchain technology from the PoW algorithm perspective because the protocol remains the most secure and decentralised, trustworthy and reliable protocol than any other one because it allows all connected nodes to participate in the mining process without any form of selection.

The part of the protocol used in generating a new block is called the mining process, the process that incurs a large amount of processing power. Mining is a key concept that is

part of PoW protocol contrary to some misunderstanding the research came across that portrays it as being the whole protocol. The mining is a task within the protocol that is used in finding the signature or in other words the right hash value of a block before a block can be added to the long chain (Aste et al., 2017). The task is computationally expensive; therefore the first miner to find the right signature is rewarded with some Bitcoins (Ghimire and Selvaraj, 2019). The flows in figure 2.3 represent the flow of the mining process. The states of the flow chart are self-explanatory.

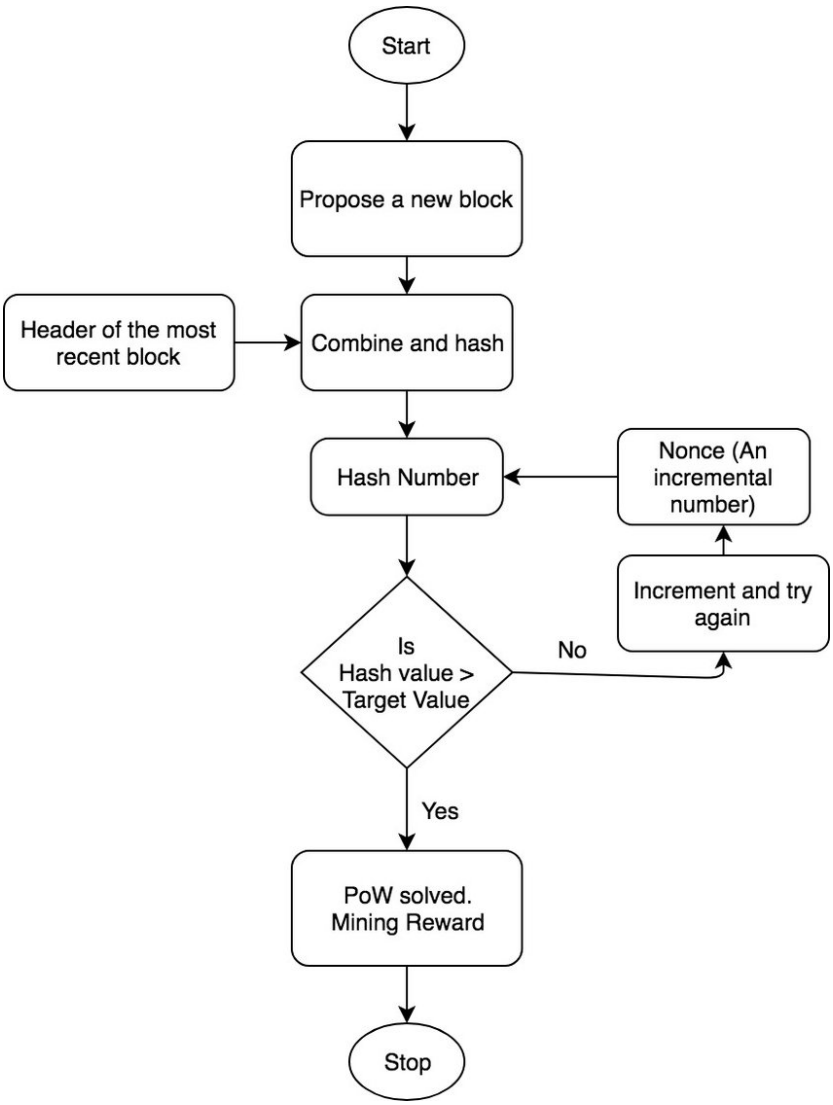

Figure 2.3: Flow of the mining process

Source (Ghimire, 2018)

The idea is for miners to search for a value called nonce value, an integer number between 0 to 2^{224} that enables the resulting hash (SHA-256) of the block header to

start with a certain number of zero (O'Dwyert and Malone, 2014). The rate at which the value can be found is controlled by a set difficulty target value but is still influenced by the speed of the mining nodes (Ghimire and Selvaraj, 2019).

- **Difficulty** The relationship between the target and the difficulty can be represented in the equation below.

$$D = \frac{T_{max}}{t} \tag{2.1}$$

Difficulty (D), Target (t), Largest possible target value T max.

- **Nonce** The miners are expected to find a value that satisfies the following equation.

$$H(B.N) < T \tag{2.2}$$

Hash function value (H), other block data (B), Nonce value (N), the target value (T).

- **Probability** The probability of finding a nonce value.

$$p = \frac{T}{2^{256}} = \frac{T_{max}}{D2^{256}} = \frac{1}{D2^{32}} \tag{2.3}$$

True Positives (TP), True Negatives (TN), Total of positives (P), and Total of Negatives (N).

2.1.3.2 Proof of Stake (PoS):

Is some worth different than the PoW in the methods of validation of transactions? The Proof of Stake shares the same purpose as PoW but the procedures for achieving the purpose differ (Thin et al., 2018). It is energy efficient unlike PoW, which doesn't reward miners for creating a new currency, it has all its currency created at stake (Decker et al., 2016). Saving the cost and the processing time. PoS is faster than PoW in processing transactions with block mine time at 15 sec which is about 40 times faster (Thin et al., 2018). Therefore, with a limited block size of 1MB, this approach can accommodate approximately 7 transactions in 15 sec, which is much faster than PoW but still insufficient to compete with the likes of Visa and not as secure as PoW.

There are chances of a small group having the monopoly of the system by having the majority of the tokens which is a big concern for the security of the blockchain system. Figure 2.4 is a good illustration of the miner, block and a statement on how the miner is chosen.

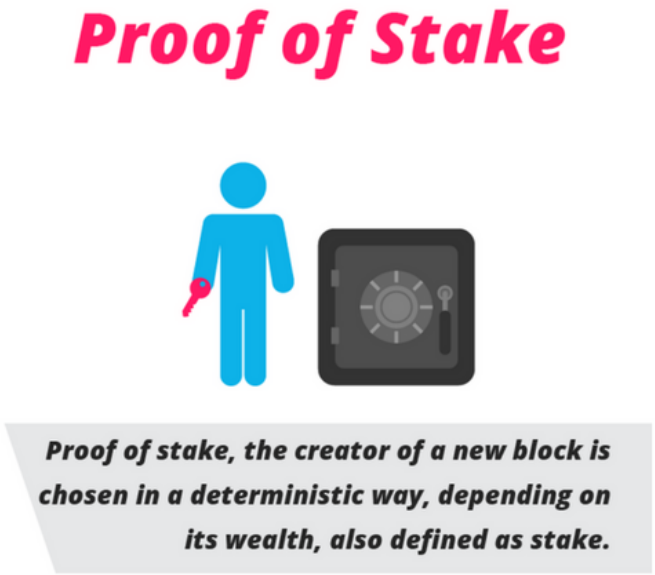

Figure 2.4: Proof of Stake
Source (Sharma, 2018)

2.1.3.3 Delegated Proof of Stake (DPoS)

Uses a consensus algorithm that maintains an irrefutable agreement on the trust that exists across the network. DPoS serves as a validation mechanism for transactions but in this case, through acting as a digital democracy mediator which uses a real-time system and entails a voting system (Chaumont et al., 2019). The procedure is effective with its democratic approach enabling producing a block at the speed of 10 sec per block, 60 times faster than PoW and surpassing PoS by 5 sec. It saves energy by reducing the complexity of the mathematical puzzle and specifying a time for adding a block. DPoS is deemed to be efficient and effective in the method of validating transaction (Yang et al., 2019). The solution provides higher transaction throughput but loses security and decentralisation compared to PoW because it requires fewer nodes in keeping the network alive which makes the ”51 percent” attack easier and the nodes with more tokens can influence the network.

2.1.3.4 Byzantine Fault Tolerance (BFT)

To understand BFT better it is important to understand Byzantine General Problem: Byzantine army are camped in divisions a night before a battle, and each division has a general that command all its activities. There are messengers used in passing messages between the generals and the fear is one or more of either the messengers or the general is a traitor. Thus, the traitor messenger can pass wrong information and the traitor general can sabotage the plan (Lamport et al., 1982).

Byzantine Fault Tolerance defines a system that tolerates the Byzantine General Problem in safeguarding the system against faulty nodes (Castro and Liskov, 1999). It is based on the assumption that having a reliable network requires at least two-thirds of the consensus participating nodes to be reliable and honest (Driscoll et al., 2003). Reducing the influence of faulty nodes by enabling cryptocurrencies to reach consensus based on PoW where there are maliciously acting nodes on the network (Sankar et al., 2017). Technically, the incoming message must be repeated by all connected nodes, if successful, the malicious issue will be ruled out and otherwise, the node is considered malicious or faulty (Jalalzai et al., 2019). Figure 2.5 illustrate the byzantine general problem. The blue arrow shows a well-functioning node while the red shows faulty nodes. the left side shows a successful attack while the right side shows an uncoordinated attack.

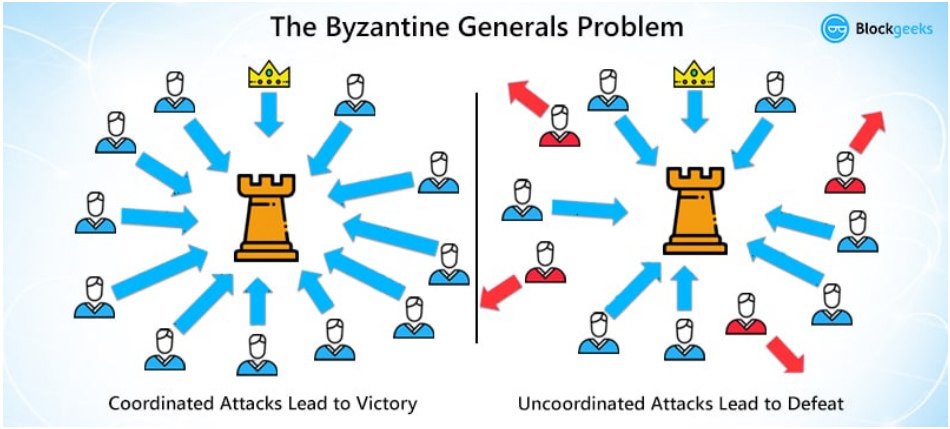

Figure 2.5: Byzantine Fault Tolerance
Source (Mitra, 2019)

2.1.3.5 Practical Byzantine Fault Tolerance (PBFT)

Practical byzantine fault tolerance is a replication of the tolerate Byzantine faults (BFT) designed to address many of the problems associated with BFT (Castro and Liskov, 1999). PBFT achieves consensus by solving a mathematical puzzle that is less complex and transactions are processed and finalised faster than the PoW because it does not require multiple confirmations. The algorithm has a finite or known number of entire nodes, therefore, it is hard to use the protocol in public systems. One of the limitations of this protocol is being susceptible to a Sybil attack: an attack where an entity can have many identities. It also has a scalability issue because an increase in the number of nodes increases the response time(Castro and Liskov, 1999) . PBF is currently used in Ripple and Stellar.

The protocol uses a voting approach in approving blocks and the validators communicate with a peer-to-peer gossip protocol. The protocol proceeds in rounds trying to reach a consensus on the next block. In each round, a proposer will suggest what should be the next block and the nodes will validate and respond with a message called VoteMessage. The block will be committed and later added to the chain if enough nodes vote for the same block. According to Cachin and Vukolić (2017), the protocol suffered a live lock bug pertaining to the locking and unlocking votes, there is a need for thorough correctness analyses of the protocol.

2.1.3.6 Ripple:

The protocol is usually known as Ripple Transaction Protocol (RTXP), it consists of two types of nodes, one participates only in transferring funds and the others are involved in the consensus process. The consensus protocol validates transactions through a network of servers instead of the blockchain mining concept, the approach saves energy, enables almost instant transaction confirmation and costs less. Ripple does not have a limit on the block size but rather on the bandwidth limiting the amount of data that can be sent within the confirmation window. Unlike the PoW consensus protocol, ripple doesn't handle failure in the case of an attack, it notifies the validators for the bad actor to be removed and allows the consensus process to resume (Chase and MacBrough, 2018).

2.1.3.7 Stellar Consensus Protocol (SCP):

The protocol was proposed by David Mazieres introducing the concept of quorum slices- A subset of quorum that helps nodes in the processing agreement. A quorum is a set of nodes working together to achieve consensus. It consists of nomination protocol and ballot protocol. The nomination protocol is used in producing a new set of candidates for approval, the values will be circulated to each node to vote for a single value and eventually result in having a collectively agreed value for that slot. Slot: is used to identify updates. The ballot protocol is used in deciding the faith of the result of the nomination, either to commit or to abort. The protocol does not entail dynamic or complex security within the process like the way PoW there are not enough encryption and verifications thus, it does not always guarantee safety (Mazieres, 2015).

2.1.3.8 Proof of Vote

Li et al. (2017) proposed proof-of-vote as a new consensus algorithm targeting the consortium blockchain. The protocol's approach to reaching consensus is through voting. The protocol has different roles and security identities that regulate the behaviour of the protocol. They include the Commissioner that performs the duty of verifying and forwarding blocks transaction and the Butler (in other words node) that specialised in producing a block. The model can ensure a low transaction delay with a high performance of the blockchain. POV is not as decentralised as the PoW it has controllable security.

2.1.3.9 Proof of Learning

Bravo-Marquez et al. (2019a) proposed a new cryptocurrency called wekaCoin that uses proof-of-learning as its consensus protocol. The protocol's approach is fully captured in figure 2.6. The protocol approach to validating blocks and transactions is inspired by and uses machine learning competition instead of the hashing-base puzzles. The process has three types of participants: A supplier that hosts the machine learning competition, a trainer that trains and submits transactions and a validator that evaluates and verifies the models, reaches consensus and proposes a new block to be added to the chain.

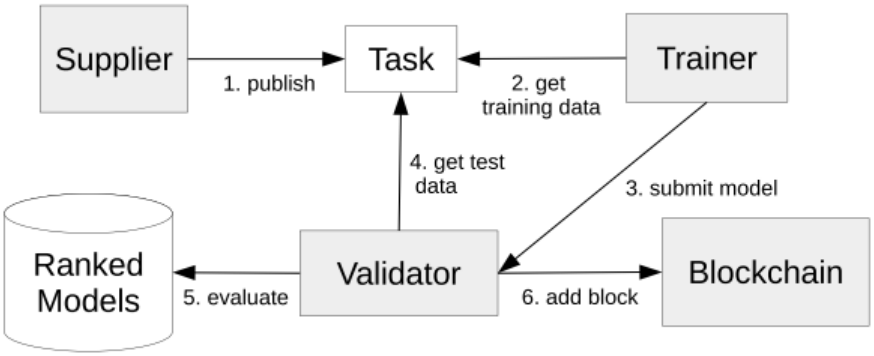

Figure 2.6: Proof of Learning Flow

Source (Bravo-Marquez et al., 2019b)

2.1.3.10 Conclusion

The section discussed other solutions proposed as a new consensus protocol. The main concern of these solutions is the scalability issue where they have to trade off decentralisation or security. The main assets of blockchain technology include the decentralisation and security it brought through its mechanism, enabling the technology to eliminate the need for central monitoring in financial transactions (Chauhan et al., 2018). The reliability and trust that comes with the technology when Bitcoin started were a result of the decentralisation and security of the protocol. Therefore, it is important to find a solution that is secured and decentralised to ensure more efficiency of the technology. Thus, the research focuses on improving the PoW consensus protocol that is the most decentralised and secured.

2.1.4 Conclusion

The first part of this section discussed many concerns about the performance of blockchain technology from different works of literature that there is a gap in this technology that calls for a solution to improve the efficiency of the technology across sectors. The second part of the section discussed other solutions proposed and implemented to address the concerns around the technology and the review came across many prospective ideas and solutions that will very well certify the blockchain performance requirement for some particular use cases or types of use cases. The new proposed protocols are struggling with the trading-off decentralisation or security which are the pillar part of the

blockchain architecture (Chauhan et al., 2018).

Only the second-layer solutions have achieved the required speed but the argument with these solutions is that they are not addressing the issues for the technology, they provide another layer where millions of transactions can be processed and confirmed before they will be passed on to be stored on the blockchain. The second layer solutions are also not applicable to other use cases and blockchain as a technology is a solution that is used across many sectors therefore, they are not considered a solution to the technology which is why other researchers continue to propose new protocols.

The review serves as a solid ground that supports and validates the research problem statement and motivation by highlighting the importance to improve the performance of the technology and how the concern is shared among other researchers, experts and other parties interested in the technology. Since all new solutions are considered less secured or decentralised if compared with the PoW, then, the research will focus on solving the issues on the protocol (PoW) in line with its research questions and hypothesis.

At this stage, the research carried out a series of analyses as discussed in chapter 5. The essence of these experiments is to have a practical understanding of the problem and identify the exact factor affecting the performance of the technology as the answer to the first research question. The experiments identified the nonce searching process when generating a new block as the main factor affecting the performance of the protocol. To answer the question, there is a need to answer the subquestions by first finding the right technique that can be used to speed up the nonce searching process.

After another series of experiments trying many ideas and techniques, the research proposed using prediction to take the miners searching for the nonce value closer to the value. This way, there is no modification to the protocol that will affect or lead to losing the decentralisation or security of the protocol. Therefore, the research identified using the machine learning technique as the best approach to achieving the goal of the research because of its ability to learn and predict or classify. Machine learning is new to the research, therefore, there is a need for a detailed understanding of the

technique to ensure successful implementation. From here research was introduced the second research question and the next chapter introduces machine learning as a technology.

2.2 Machine Learning Review

In this section, we are going to review Machine learning because it is the technology we are going to use to optimise the block generation time. Machine learning is an important part of artificial intelligence that refines the automotive learning process of computers using techniques from a wide range of algorithms applicable in different domains based on data models such as regression or classification (Ray, 2019). It deals with teaching a computer without explicitly programming it (Mitchell, 1997). It comprises principles from various combinations of disciplines such as information theory, philosophy, statistics and probability, neurobiology and psychology, control theory, computational complexity and artificial intelligence (Kim et al., 2016). The algorithms of machine learning are used in various applications and provide various benefits. Performance of machine learning strongly depends on the amount and quality of given data, which shows the strong connection between data integration and machine learning as emphasised by (Dong and Rekatsinas, 2019). Over the years, machine learning techniques have been used to automate processes in a wide range of fields including cryptocurrency Demir et al. (2019) used some regression models to predict Bitcoin price. It has been used to solve many problems for systems such as recognition systems, informatics, data mining, and autonomous control systems.

2.2.1 Review of machine learning techniques

The generosity of machine learning algorithms or techniques illustrates data analysis and problem-solving approaches to resolving complexities (Bernardi et al., 2019). The machine learning technique is an important tool used in data mining to determine the relationship between data in large datasets (Sharma et al., 2013). The techniques empower data insights and tend to forecast future trends, changes and opportunities (Kotsiantis et al., 2006). To harness the machine learning key techniques, we employ computational methods to learn information directly from available data (Shrivastava

and Kumar, 2019). The available machine learning techniques count on supervised, unsupervised, neural or deep learning, reinforcement and transfer learning methods to accommodate data efficiency and effectiveness in Blockchain technology.

2.2.1.1 Supervised Learning

Supervised machine learning makes use of statistical models to predict output in numerical data and have the correct label classified (Saravanan and Sujatha, 2018). Supervised ML techniques count on the available dataset to make reasonable predictions (Bzdok et al., 2018). The famous supervised techniques are categorized as classification and regression. The outcome of supervised learning is based on past experiences, optimising performance criteria and resolving real-time computational problems (Singh et al., 2016). The strength of supervised machine learning resides in known training and labelling, class identification, specification of data, and predicting the numerical target value of given data and labels. The supervised learning is applicable to marketing and sales, lifetime analysis, churn rate and people analytics (Winter, 2019).

Classification Classification tends to group label input through binary or multi-class classification, F.Y et al. (2017) categorized supervised learning algorithms as linear (logistic regression, and support vector machines), non-linear models (KNN, SVM, Naïve Bayes, decision tree and random forecast classification). The method facilitates the specification of class for determining class elements and usability of discrete value output (Kotsiantis et al., 2006). Common examples of classification include the classification of email as spam or harm, customer segmentation, bank loan grants, and positive or negative outcomes. Using the probabilistic interpretation, classification aims to group or classify similar tasks and things by agreeing on considerable satisfied conditions. The implications of classifications are extended to real-life problem solutions (Binkhonain and Zhao, 2019).

Regression process The regression analysis techniques predict values by taking training or input data and emphasising continuous numerical values (Cui and Gong, 2018). The regression models are applied for determining the relationship between two variables to ascertain forecasting. The regression takes on quantity while on the other hand, classification methods account for discrete class levels. The implications

of machine learning and regression analysis feature relationship effect, change in target variable and predict future trends. The categorization of regression features linear, multiple, logistic and polynomial regression (Binkhonain and Zhao, 2019). Oracle attributed regression as a data mining function for predicting numbers but the scope of regression is extended to financial forecasting, trend analysis, marketing and time series prediction. Some common types of regression algorithms include:

- Logistic regression: The algorithm identifies the dependent variable as binary or dichotomous where two class values determine the categorical dependent variable. The logistic regression method estimates the occurrence of an event by interpreting input values and the probabilistic nature outline the result in 0 and 1 and 1 is ascribed to complete certainty (Bonaccorso, 2017). The real-life implications of logistic regression surround fraud detection, credit card scoring, clinical trials, customer insurance and real estate. The distinguishing element of logistic regression is openness to dependent variables and projecting quantified value of the relationship between variables.

- Linear regression: The methods interpret the relationship between quantitative target variable Y and input variable X. The linear regression predicts output with the constant slope in a continuous range in making predictions and estimations. The regression algorithm corrects weights and overcomes biases for approximating the best fit. The regression slope depicts a positive or negative relationship between two variables while multiple regression features relationship complexity. Abdelsamea et al. (2017) emphasized regression models in machine learning and minimizing the variance between actual and predicted values and it incorporates cost function and gradient descent. The model performance or goodness of fit for predicting the best fit line can be obtained through the R-Square method (Cui and Gong, 2018). The model can be used in financial portfolio prediction, salary forecasting, and traffic trends accompanied by multiple regression exposure to social sciences.

Challenges of supervised learning Supervised learning encounter challenges to justify the adoption of supervised machine learning techniques. Collection and organi-

zation of data are critical for extracting relevant information and predictions and it's a big hurdle to find relevant data that is enough to allow good performance of the model. In addition to that, probability accuracy is also subjected to deviations and standardized approaches are less relevant to data mining and extraction (Kathole and Chaudhari, 2019).

2.2.1.2 Unsupervised Learning

Unsupervised machine learning techniques lean on intrinsic or hidden structures by drawing inferences from input data without labelled responses (Celebi and Aydin, 2016). Unsupervised machine learning does not possess label data. In regards to this, data samples are stored in cluster groups depending on how similar or dissimilar they are (Vats et al., 2018) making use of different approaches. For instance, association rules algorithms and K means clustering. The rationale for unsupervised learning is to maximize the utility of exploratory analysis, finding unknown patterns, real-time processing and ease of extracting unlabeled data. The unsupervised learnings categorize clustering, visualization algorithms, anomaly detection and association. The unsupervised learning techniques don't require laborious data labelling and take on unlabeled data for the training process to comprehend output or the decision making process. Some common techniques of unsupervised learning include:

- Clustering: It's a concept that deals with the structure or patterns of unlabeled data and process data in clusters or groups by determining class methods and class models. The clustering process categorizes or groups data of similar characteristics where the algorithm defines output. The famous clustering techniques are K-means, K-Medians, Expectation Maximization (EM) and Hierarchical clustering for accommodating machine learning (Celebi and Aydin, 2016).

- Visualization: The visualization algorithms absorb unlabeled data and display it in 2D or 3D format where visualization of cluster provides room for interpretation. The anomaly detection technique is applied prior to training and is applicable for detecting criminal activities (Abdelsamea et al., 2017).

- Association: The outcome of association learning methods create a relationship

among data objects and variables in large databases. The examples of associations illustrate new home buyers who are more likely to buy new furniture

- Anomaly detection: The algorithm detects anomalies or errors in data without any prior training or input data for example suspicious credit card transaction detection to undermine criminal activities. The anomaly detection technique is used to detect the different or distinguished things and spot outliers with more specifications (Hasan et al., 2019). For example, anomaly detection complements the detection of suspicious data and taking corrective measures for corrections

Challenges of Unsupervised learning The freedom to use raw data questions relevancy and algorithm usefulness as data lacks label information. The uncertainty and lack of precision are quite evident in unsupervised learning and manual inspection becomes paramount in unsupervised learning. Unsupervised learning defines blurred problems due to a lack of labels and it becomes ambiguous for AI to interpret data and performance measures (Soni, 2020).

2.2.1.3 Reinforced Learning

The area of machine learning maximizes the likelihood of cumulative reward by focusing on the balance between explorations (unchartered) and exploitation (current knowledge). Reinforcement learning prioritizes penalties or rewards for actions for correct and wrong moves. The rationale for using reinforcement learning is yielding worthy actions and rewarding them by configuring which action needs large rewards (Montague, 1999). Examples of reinforcement learning are autonomous cars, deepsense.ai or learning to run projects, industry transformation through robotics, trading and finance and natural learning processing (NLP).

The reinforcement learning methods emboss decision making by presuming game-like situation and use trial and error methods for providing solutions with the typical reinforcement scenario features agent, environment (e), reward (R) state (s), policy (π), Value (V) and Q value or Action value (Q). The operational scenario of reinforcement learning emanates value-based, policy-based, and model-based algorithms. In reinforcement learning, there is no supervisor for sequential decision making and delayed

feedback for positive and negative whereby Q learning and Markov decision process (Montague, 1999).

The reinforcement learning process enables an agent to learn through the environment by using feedback from their own actions and experiences. The exploration vs exploitation trade-off is the dilemma for exploring new states and maximizing rewards. Q learning and SARS are frequently applicable methods of reinforcement learning. The challenge with reinforcement learning is the availability of a simulation environment which is task-dependent. The complexity of reinforced learning is attributed to uncertainty in an environment which restricts collision. Moreover, reinforcement learning also encounters scaling and tweaking of agents and reaching local optimum are also enlisted as challenges to reinforcement learning. The limitation of a specific environment where enough data availability is ensured makes reinforcement a complex time-consuming method. The reinforcement learning method also faces challenges of reward/feature design, parameters, realistic environment, and information overload (Zihan and Dong, 2020).

2.2.1.4 Comparison of Supervised, Unsupervised and Reinforcement techniques

The thin line between supervised and unsupervised learning accentuates labelled datasets, available with algorithms to interpret data accuracy on training data. The availability of both input and output variables distinguishes supervised learning as algorithms are trained to use labelled data. Supervised learning ponders over input and output data links for the training models which authenticate the accuracy and trustworthiness of output with the real-time learning process while the other hand, while unsupervised learning doesn't use output data where a number of classes are not known. The classification of supervised learning is a challenge whereby unsupervised learning is labelled and not known (Sathya and Abraham, 2013).

Supervised learning requires a large amount of data inputs followed by annotation for categorization. On the other hand, unsupervised learning is free from data classification and the clustering process counts on machine learning interpretation. The algorithms used in supervised learning are enlisted as support vector machines, neural networks,

classification and regression. On the contrary, unsupervised learning uses different algorithms namely clustering, K-means, and hierarchical clustering which becomes complex for computation (Berry, Mohamed and yap, 2019). The comparative analysis of supervised, unsupervised and reinforcement learning methods is tabulated in the table 2.1

Attributes	**Supervised**	**Unsupervised**	**Reinforcement**
Specification	Data labels are used for training and machine learning	The training process counts on unlabeled data without any prior guidance	Agent interacts with environment and learns from errors or rewards
Categories	Classification and Regression	Clustering and Association	Reward-based
Data type	Labeled	Unlabeled	Spontaneous data or feedback
Supervision	Strict	No	No
Perspective	Mapping labelled data inputs to known outputs	Patterns understanding and discovering outputs	Trial and error
Algorithm	Linear regression, logistic regression, SVM and KNN	K-Means, C-Means and Apriori	Q Learning and SARSA
Objective	Outcome based	Discovering patterns	Learning through feedback
Application	Risk Evaluation and forecasting	Anomaly detection	Self-driving and automation

Table 2.1: Source: (Dangeti, 2017)

2.2.2 Machine Learning and Blockchain

In recent times, machine learning and blockchain have been two major focal areas in emerging research. Machine learning is the practice of developing learning models on computers that will parse data and provide decisions or predictions for some real-

world problems. On the other hand, blockchain can process and store data, preserve the integrity of data and govern the accessibility of peers without needing any centralized administration. Those researches are significant data-driven and each has its bottlenecks and advantages. This section highlights the strides of other researchers and company experts to use machine learning in blockchain technology. The discussion will focus on discussing what has been achieved, techniques and results to provide a base for the possible success of the approach. The discussion highlighted all ideas the research could find in no particular order of selection because they are not many.

2.2.3 Blockchain security attack detection

A major concern of the public in regard to blockchain technology is its performance in terms of security. Although blockchain makes use of consensus and cryptography to enforce its network privacy and security, it is not entirely immune to attacks. Bitcoin researchers Conti et al. (2018) identified that bitcoin is vulnerable to some specific attacks, even though it has been able to run continuously for years. Some vulnerabilities of Ethereum were exposed in 2019 and it was identified that the network has experienced different forms of attacks like 51 percent attacks and data breaching attacks (Chen et al., 2019a). One of the tools to improve blockchain security is machine learning. The work Scicchitano et al. (2020) proposed an unsupervised machine learning approach in identifying various activities of the network on blockchain. The proposed detection system of anomaly constructs a neural model of encoder-decoder which possess the capability to summarise the status of the ledger in a ledger sequence-by-sequence manner. The system possesses the ability to identify the difference in the status between anomalous situations and have the alert triggered accordingly.

Dey (2019) intended to improve the consensus mechanism of blockchain through making use of game theory and algorithm of supervised machine learning. An improved Proof-of-Work consensus was adopted for preventing any attacks that can be quantified. Through the analysis of the attacker's activities and rewards, a payoff/utility function can be achieved and fed into the supervised machine learning model. This model of machine learning possesses the ability to detect whether an attack can be expected to happen or not based on community/service value. If the attack has a high likelihood of

happening, the machine learning agent possesses the ability to prevent confirmation of blockchain until a new block is produced again. The work iteHou2019SquirRLAA introduced the SquirRL framework a deep reinforcement framework for learning that can be used to evaluate the rewards in the blockchain. Although the squirrel framework is used in detecting the activities of adversaries in the network, it can automate the detection of a vulnerability in the incentive mechanism of the blockchain. In the instance that the theoretical analysis is infeasible, SquirRL stands as an essential tool for blockchain engineers in verifying the design of protocols during their phases of development.

2.2.4 Cryptocurrency and Mining

Thanks to cryptography and blockchain, the advent of cryptocurrency has drawn significant attention. Unlike stocks and fiat money, cryptocurrencies have experienced major unstable fluctuations and have been a major disruption in the investment industries. Researchers have carried out steady and continual progress on how the profitability of cryptocurrencies can be improved through applying machine learning models to assess network data and market performance. The work McNally et al. (2018) proposed a method that can help in the prediction of fluctuations in the price of cryptocurrencies. The method proposed collects the online posts of users and comments that are related to activities of the cryptocurrency market and carries out an association analysis between the data collected and price fluctuations of cryptocurrencies. The model finally drawn identified that approximately 74 percent weighted average precision in the Ethereum and Bitcoin markets. The work Madan (2014) focused on automating the trading of Bitcoin through supervised machine learning algorithms by adopting binomial logistic regression and random forest in supporting the vector machine. Their method of learning was trained with the price index of Bitcoin and achieves more than 55 percent precision. Jang and Lee (2018) adopted the Bayesian Neural Network algorithm for the training of the supervised learning model. The data for training the empirical study comprise the cryptocurrency market volumes and prices, financial stock market information and global currency ratio. A promising result was presented in their research in regards to anticipating the price time series of Bitcoin and providing an explanation for the volatility of the bitcoin market.

The work McNally et al. (2018) combined two different forms of deep learning models in forecasting the price of Bitcoin with the Long-Short Term Memory (LSTM) and Recurrent Neural Network (RNN) algorithm. The two models were identified to achieve approximately 50 percent of accuracy in simulations but the LSTM model possesses the capability to acknowledge the dependencies of the market in the long term period. Jourdan et al. (2019) formulated a few dependencies that are conditionally induced by the block design of the protocol of Bitcoin and proposed a probabilistic graphical representation for the prediction of the value of UTXOs, which records the number of Bitcoin that is used in each transaction. The work Wang et al. (2021) adopted the Reinforcement Learning algorithm to analyze the profit attached to different forms of mining strategies and discovered the optimal strategies to mine over time-varying networks of blockchains. Some researchers showed that the mining of Bitcoin cannot be quantified as a Markov Decision Process (MDP), and different reinforcement learning algorithms can be applied in constructing the model of MDP (Eyal and Sirer, 2013). Aside from that, the work iteNguyen2020 introduced a reinforcement learning-based offloading scheme that provides assistance for mobile miners to determine the optimal decisions to offload, reduce consumption of energy and avoid latency in the network.

2.2.5 Transaction entity classification

Bitcoin has continually been accepted as an alternative medium for exchanging value. Some users have taken advantage of the Bitcoin network for illegal transactions and purposes. With CoinJoin mix services, Bitcoin has been identified as a currency that is safe in the darknet markets and can be used to launder money. As a result, there is an urgent need for developing address and transaction tracking systems. Machine learning has generally been identified as an essential tool to carry out cryptocurrency address labelling and clustering for detecting illegal activities in the year 2017, the work Sun Yin and Vatrapu (2017) developed different forms of classifiers making use of supervised machine learning models for identifying Bitcoin addresses that relate to illegal or criminal activities. The following year Harlev et al. (2018) also proposed a supervised learning model with an algorithm for gradient boosting. All those models of learning can achieve 75 percent accuracy in the simulation of address clustering. Aside from that, Akcora et al. (2019) introduced a traceable and efficient framework

identified as the BitcoinHeist. Through the application of topological data analysis into the record of transactions, the BitcoinHeist can automate the prediction of new transactions of ransomware in a cluster of addresses and identify the ransomware that does not have any record.

2.2.6 Blockchain-enabled machine learning model

s

While the systems of machine learning have become essentially powerful tools for resolving problems in the real world, people have continued to question the level of its trustworthiness. In the first instance, machine learning might be susceptible to data poisoning attacks. Hackers might endeavour to manipulate the performance of the system by altering the data collected or inserting constructed poisoning instances. On the other hand, it is difficult for humans to understand the decision that machine learning systems make if there is no system logs that can be traced or specific training history. Thirdly, the centralized servers are still significantly required for the completion of the model training processes. Lastly, the stages of model construction are not automated and the involvement of humans may bring about biases in the final system. A great potential has been identified in Blockchain and smart contracts in potentially solving such challenges.

2.2.7 Blockchain for data security

Blockchain has been well known for keeping data safe and secure. With traceable and reliable data stored on the blockchain, researchers can ensure that machine learning algorithms will produce the most credible and trusted result. The work iteshayan2019biscotti identified a federal learning system identified as Biscotti which makes use of cryptographic and blockchain primitives in coordinating a privacy-preserving federated process of learning between peering clients. While all the iterations of training have been stored in the blockchain, only the updates that are peer-verified are committed to the final model. Training data are stored locally with the data providers. This system possesses the ability to protect the privacy of the data of an individual client and also defend against attacks on data poisoning. The work Mugunthan et al. (2020) offered a

BlockFlow which is a privacy-preserving federated learning system.

The system adopts various differential privacy models, introducing a novel mechanism for auditing model contributions and making use of smart contracts to incentivize positive behaviours. Nevertheless, the system does not possess the ability to detect any form of an anomaly during the process of learning. For that issue to be addressed Desai et al. (2021) developed another blockchain-based federated framework for learning which was identified as BlockFLA. After deploying of learning algorithm, the BlockFLA framework makes use of smart contracts to automatically detect and discourage any form of backdoor attacks by having the responsible parties held accountably. Both of the frameworks ensure that the algorithm of machine learning tends to be resilient to malicious attacks.

As of the year 2018, Chen et al. (2018) introduced a secure system of supervised machine learning identified as LearningChain. In regards to this, they developed a differential mechanism of privacy for the process of local gradient computing to protect individual providers of data and an l-nearest scheme of aggregation for defending against attacks of Byzantine in the process of global aggregation gradient. Afterwards, Kim et al. (2019) identified that the LearningChain system has various limitations such as low efficiency in computation, zero support for non-deterministic computations of function and weak privacy preservation. To have the issues revolved systematically, they build an improved distributed model of machine learning for permission blockchains. With an error-based aggregation rule and differentially private stochastic gradient descent method as core primitives, their model provides better defences against byzantine attacks and possesses the capability of handling the learning algorithm with defined non-deterministic functions. Aside from that, Zhou et al. (2020) also proposed a system that is similarly called the PIRATE to provide the distributed algorithms of machine learning with byzantine-resiliency but this was designed for the 5G network.

2.2.8 Blockchain for system improvement

Smart contracts and blockchain can also be adapted to provide improvement for the processes of machine learning and eliminate the involvement of humans. The work

Ouyang et al. (2020) adopted a novel framework for federated learning collaboration, the learning markets. Within the learning markets, the blockchain creates a trustless environment for transactions and collaboration. The providers of learning tasks simply need to publish initial tasks towards the market and deposit rewards within the network. The trainers and the data providers participate in the process of learning through depositing an entrance fee, having data uploaded/downloaded on the IPFS network and having their computation power controlled. Various predefined smart contracts serve as network agents in the maintenance of collaborative relationships and market mechanisms. The work iteKim2020 introduced on-device architecture for blockchain-based Federated learning identified as BlockFL. The data on the device of users are locally processed and the local updates are accumulated on the blockchain. The updates of the global model are calculated based on the user updates that are recorded on each of the blocks. Primarily their architecture mainly focuses on the minimization of latency and scalability of the system. They also identify that systems may not have the ability to retrieve the updates of the local model on time as a result of network delay or intermittent problems availability.

The work ur Rehman et al. (2020) provided a complete requirement list for a federated learning framework that is blockchain-enabled including decentralization, penalization, fine-grained federated learning, trust, incentive mechanism, heterogeneity, activity monitoring and contest awareness, communication, model synchronization and bandwidth efficiency. They also proposed the term reputation and describe how the attribute works in their proposed framework. Aside from this, some researchers in this regard work on the development of new mechanisms of blockchain for the distributed task of machine learning. The work iteBravoMarquez2019ProofofLearningAB invented a new protocol identified as Proof-of-Learning, which achieves distributed consensus through ranking systems of machine learning for a given task of machine learning. The objective of such a protocol is to help mitigate the computational consumption in solving puzzles that are hashing-based while still ensuring data integrity. Toyoda and Zhang (2019) improved the common mechanism of incentive in the current network of blockchain and make it more applicable to the blockchain network when the task of machine learning is involved.

2.2.9 Transportation

The work ite9079513 introduced a mathematical framework that adapts the design of blockchain-based federated learning into the autonomous vehicle sector. They adopt the mechanism consensus and a renewal reward approach to be enabled on-vehicle machine learning training in the network of distribution. The global models and on-vehicle updates are maintained in the blockchain, which is visible to and verifiable by all vehicles. Rewards are distributed to the owners of the vehicle based on the size of their updates that are accepted into the global model. They also have the limitations discussed in regard to the designs and the performance of the system based on numerical analysis and simulations. Hua et al. (2020) endeavoured to apply federated learning algorithms into heavily haul railway management. In their research, the train controls are quantified into various multiple classes and the data on an individual train applies to the SVM-based mixed kernel. The smart contract carries out the global model. This research resolves the issues relating to Data Island in this sector and the algorithm of asynchronous collaborative learning designed without the involvement of a central server.

2.2.10 Healthcare

The healthcare sector has long been an early adopter of technological advancement and greatly benefited from it. Chen et al. (2019a) introduced a blockchain-based disease classification framework identified as Health-Chain. In the system of Health-Chain, multiple institutes can have their model trained with their patient records, asynchronously collaborate in the blockchain network and contribute to the global model with preserved privacy. The researchers implement the systems in two tasks for disease recognition, ECG arrhythmia classification and, breast cancer diagnosis and both simulations demonstrate promising results. Kumar et al. (2020) proposed a more elaborate but similar supervised machine learning framework for the detection of COVID-19 in patients. The framework proposed can utilize data that are up to date which can improve the recognition of ICT images. Both the above researchers focus on developing the machine learning models and the blockchain is adopted for enforcing the consensus across various research institutes and aggregating the models of training. The work

Rahman et al. (2020) gave a complete picture of the way blockchain can be adopted from the Internet of Health Things (IoHT) perspective. From the framework they proposed, n smart contracts are used to manage the trust management, train plan, participant authentication and data encryption of the device. The framework design has high scalability and security level in the health management area of IoHT.

2.2.11 Supply Chain Systems

Kamble et al. (2021) developed a model for prediction using machine learning techniques to evaluate the probability of an organization of successful adoption of blockchain within the supply chain sector. The researcher focuses on explaining the extent of blockchain adoption through the use of psychological constructs from the literature regarding technology adoption. The model of learning can help managers in predicting their organizational readiness. The work Mao et al. (2018) introduced a credit evaluation system that is blockchain-based for strengthening the efficiency of management and supervision of the food supply chain. The system collects the evaluation of credit from blockchain traders, directly analyzes the evaluation through a deep learning network and provides the credit results for the management and supervision of regulators. The work Yong et al. (2020) proposed a "vaccine blockchain" system in detail based on machine learning technologies and blockchain. The vaccine system is designed to support the prevention of supply record fraud and tracing vaccine inventory.

2.2.12 Blockchain and Machine Learning Issues and Challenges

These technologies are being expected enthusiastically across the globe, but yet various obstacles resist the integration of blockchain technology and machine teaching (Bravo-Marquez et al., 2019b). The integration of both is still at its infancy stage and many open challenges and issues are yet to be identified or expressed. The most relevant challenge of utilising machine learning in the application of blockchain technology is data privacy - data generated through devices to be stored on the blockchain is available to the entire nodes of the blockchain (Zheng et al., 2017c). This results in a potential concern for privacy for data that needs to be stored either confidential or private and this imposes barriers on Machine learning models for analytics and prediction.

2.2.13 Other Applications and results

Machine learning is not limited to supervised, unsupervised learning and reinforcement learning rather its scope is extended to semi-supervised learning, self-learning, feature learning, sparse learning and Robotics. The categorization of machine learning is spread to hybrid learning, statistical inferences, and learning modules. Semi-supervised learning is the type of supervised learning which counts on training data whereby few data labels and unlabeled data are embedded for making effective use of available data. Taking inspiration from unsupervised learning for interpreting unlabeled data for extracting clustering or pattern identification. For example, labelling the dataset of photographs (van Engelen and Hoos, 2020). Self-supervised learning techniques count on a corpus of unlabeled images and training for supervised learning. For example, making images grayscale or colourization and auto-encoders. Ensemble methods take the idea by combining multiple predictive models for determining high-quality predictions. For example, Random Forest algorithms combine multiple decision trees to reduce the variance and bias of the model.

Neural networks and deep learning captures non-linear patterns in data by embedding layers of parameters. The simple neural network is flexible enough to build linear and logistic regression. Deep learning collaborates with neural networks by holding a set of multiple hidden layers and large data and better computational power for the best performance. Deep learning is a growing field for image analysis and face recognition where graphical processing units are required. Neural networks are expensive and time-consuming and require large databases for accuracy and reliability. The neural networks are reflected as black boxes where data scientists are unaware of processing and the network tends to overfit and becomes hard to interpret (Ghasemi et al., 2018).

Transfer learning is a machine learning technique which re-trains neural networks or re-uses part of data to adopt new tasks. Once the neural network is trained, transferring the fractions to new tasks can maximize the adoption and learning process. The advantageous factor of transfer learning is related to less consumption of data and it becomes difficult to find labelled data for training. Natural language processing prepares text for machine learning and is indirectly related to machine learning and serves as a

support function to machine learning. Word embedding through TFM and TFIDF are numerical representations of text documents which consider frequencies for quantifying the text documents.

To complement data analysis, supervised, unsupervised and reinforcement learning techniques provide comprehensive details. By delving into challenges and benefits, the modern evolving methods of deep learning and neural processing are complex and time-consuming. The complex nature of deep learning and the time-consuming process makes it difficult to propel and require investment and research to gauge the viability of machine learning in blockchain technology. Combining machine learning with blockchain technology is considered beneficial for the interpretation of data where accuracy and reliability complement transactions. The enhancement of security will complement data analysis and machine learning is likely to facilitate blockchain technology features for data sharing (Chen et al., 2019b). The evolution of data collaborates with blockchain technology and considers secure and accurate reliable data sharing.

2.2.14 Conclusion

The first part of the section discussed in-depth the different techniques of machine learning and cleared the pathway of understanding how and when the technique can be used. This provided enough insight to allow the selection of the right technique of machine learning to be applied in the research. The second part of the chapter reviewed how the technique was successfully used in Blockchain applications such as predicting bitcoin prices. Although that does not imply it will also be a success in this research, it serves as a strong base of support for the research hypothesis that says Machine learning techniques can be used to narrow down the block generation time and improve transaction throughput. The research is going to use the supervised learning approach in optimising the block generation time because it uses past data to learn from experience and there are many publicly available blockchain data to work with, and more data can be generated. The research will use the regression approach of supervised learning instead of the classification because we are trying to predict a value not classify one.

2.3 Conclusion

In summary, this first section covered the relevant reviews of literature that enlightens the research about the issues that hinder the performance of the Proof of Work consensus protocol and the strides by other people to solve the challenges. The findings can be summarised to:

- Throughput – the number of transactions that can be processed per second. The current tps rate is very low compared to the rate achieved by other counterparts and the mainstream payment systems.

- Block size – the amount of transaction a block can take. Understandably, It is important to have a limit to how many transactions there can be in a block but such a limit is, unfortunately, a big constraint to the performance because of the time it takes to generate a block.

- Scalability – trading of security or decentralisation to get speed. Efforts by others to solve the performance issue have so far all resulted in having to trade off security or decentralisation to get speed.

The findings in the first part of the chapter have answered the first research question that asked about the factors affecting the performance of blockchain technology. The second part of the chapter has informed the research about important factors to the success of machine learning implementation. Also, discussed the success and scope of machine learning integration into the blockchain to further pave the way and guide the research direction to novelty. The findings can be summarised to:

- Selecting the right technique is important in ensuring the model's accuracy and the selection depends on the dataset's behaviour. In this case, the analyses discussed in chapter 5 has identified the data behaviour to be linear therefore, a linear regression model has been selected for the research.

- Machine learning has already been used in blockchain applications to make predictions that support or inform users but it has never been used within the protocol

to support any technical process. Thus, the research idea to use ML in improving the mining process is novel.

- There is a need for enough, clean and accurate training datasets to ensure model accuracy. Enough – large enough for the model to train, learn and identify patterns within the dataset. The more complex the dataset, the more data is required.

The findings in the second part of the chapter have partially answered the sub-question of the third research question that asked for the appropriate machine learning technique for the research use case. The research will need to explore the algorithms that support linear regression and find the appropriate one. To do so, there is important to collect accurate data that will allow such experiments. Therefore, the next chapter will discuss the data collection process.

Chapter 3

Dataset and Performance Analysis

The review in chapter 4.1 informed the research of the critical importance of extracting relevant information when collecting and organising data to ensure accurate prediction of a machine learning model. Also, it is sometimes difficult to find the amount of data needed to train a model and ensure good performance. The amount of data needed is dependent on the complexity of the problem that needs to be solved. The data needs to be cleaned and transformed where necessary into a state Machine Learning can learn from. Therefore, this chapter discussed how data used in training the research's machine learning model was collected and processed and how the performance was analysed.

The research used two different types of datasets, the Ethereum dataset was used for the simulation and model development and the Bitcoin dataset was used for the evaluation of the model. The rationale behind choosing Ethereum and Bitcoin is because they both use PoW consensus protocol and Ethereum as an open-source solution allows the research to run a simulation, experiment and play around with parameters. The research datasets were collected through simulation and downloads from publicly available application interface API.

3.1 Definition of variables

The two datasets include attributes included in the block header sent to miners to find the nonce value. It is important to understand that the dataset here does not include transactions stored in the block because the research focuses on block generation and that only requires the header attributes. But the transactions are also represented on the header attribute as the Merkle root attribute in the Bitcoin block header and as transaction root in the Ethereum dataset. Our dataset is expected to have a uniform data type, therefore, only some attributes of the block header are required. The Ethereum dataset includes seed, Difficulty, GasLimit, GasUsed, Time and nonce. Figure 3.2 shows how the dataset was collected from the simulation. The Bitcoin dataset includes confirmations, size, stripped size, weight, height, version, time, median time, difficulty, nTx and nonce. Figure 3.1 shows how the dataset was collected from the blockchain API. Detailed information for all variables of the two datasets has been listed in the appendix. After the first experimentation and its poor accuracy, the researcher studied the process of sending the header attributes to the miners and identified a variable called seed in the Ethereum mining process that is important to the nonce searching process but missing in the block header or record. The missing seed value holds the total target value of the nonce which is the value the search starts from.

```
18  ff = open("dataset1.txt", "w")
19  counter = 702340
20  while counter <= rpc_connection.getblockcount():
21      block = rpc_connection.getblock(rpc_connection.getblockhash(counter))
22      ff.write(str(block["confirmations"]) +
23              "," + str(block["size"]) +
24              "," + str(block["strippedsize"]) +
25              "," + str(block["weight"]) +
26              "," + str(block["height"]) +
27              "," + str(block["version"]) +
28              "," + str(block["time"]) +
29              "," + str(block["mediantime"]) +
30              "," + str(block["difficulty"]) +
31              "," + str(block["nTx"]) +
32              "," + str(block["nonce"]) +
33              "\n")
34      counter += 1
35  f.close()
```

Figure 3.1: How the Bitcoin dataset downloaded

Figure 3.1 shows the python implementation of how the experiment connects and collects data from the blockchain through the API. Dataset1.txt on line 18 is the name of the file where all collected data is stored and the words in the quotation and green colour from lines 22 – 32 are the names of header attributes that were collected. Figure 3.2 shows the code added to the Ethereum open source code to collect the dataset. "Myexperiment" is a class added to keep all addition or custom methods that are required in the experiment for example "WriteToFile" which allows writing collected data to a text file. "strconv" is a library used to convert the header attributes into a string to allow writing to the giving text file – dataset.txt.

```
221                 myexperiment.WriteToFIle("dataset.txt",
222                         strconv.FormatUint(seed, 10)+","+
223                         header.Difficulty.String()+","+
224                         strconv.FormatUint(header.GasLimit, 10)+","+
225                         strconv.FormatUint(header.GasUsed, 10)+","+
226                         strconv.FormatUint(header.Time, 10)+","+
227                         strconv.FormatUint(nonce, 10))
```

Figure 3.2: How Ethereum dataset extracted from the simulation

3.2 Data collection

The data collection process initially focused on the available data from the public blockchain downloaded through blockchain.info's blockchain data API and BitcoinRPC. After the first experiment with the initial downloaded data, the seed value missing in the actual dataset was identified as an important value that is required to enable the machine learning model to identify the right pattern within the dataset. The seed value is not included in the block header nor is it stored on any ledger, it is a value only used within the consensus protocol. Therefore, it called for the need to simulate the Ethereum network and generate the dataset within the consensus protocol to ensure model accuracy.

3.2.1 Data Downloads

The blockchain.com API provides all data stored in the blockchain record and the dataset requires only variables included in the block header. Therefore, To download the Ethereum dataset, the python script used to download the data was implemented to download only the records of the block header attributes. The Bitcoin dataset was downloaded by running a full node of the network and using the BitcoinRPC API of the Bitcoin core software to read all the block records as seen in figure 3.1. The python script used in Bitcoin differs from Ethereum because, in Bitcoin, the mining script uses much of the data received to generate the block header.

Thus, the first data to be collected for this research was at least a hundred thousand (100,000) out of six hundred thousand plus (600,000+) blocks in the bitcoin network. But the implementation to generate the header out of the record was a struggle and it was decided to use the Ethereum dataset instead. Another one hundred thousand (100,000) Ethereum block records were downloaded. After a thorough study of how the Bitcoin miner is implemented and a good understanding of how to generate the block header out of the block record sent to miners, over seven hundred thousand (700,000) records of the Bitcoin blockchain were downloaded for the evaluation of the research model. A sample of the collected dataset can be found in figure 3.3 and 3.4

```
699985  2356,1803785,729753,3993044,699984,536887300,1631327441,1631325129,18415156832118.24,347,1125819244
699986  2355,1707805,763493,3998284,699985,671080452,1631328007,1631325924,18415156832118.24,1055,1519120135
699987  2354,1771749,740407,3992970,699986,551550980,1631328462,1631326114,18415156832118.24,978,3295980925
699988  2353,1783956,736344,3992988,699987,536928260,1631328832,1631327171,18415156832118.24,630,813521595
699989  2352,1837635,718659,3993612,699988,536887300,1631328892,1631327283,18415156832118.24,159,1135650033
699990  2351,485614,288579,1351351,699989,536870916,1631329010,1631327441,18415156832118.24,509,2696245030
699991  2350,1249689,700466,3351087,699990,536895492,1631329670,1631328007,18415156832118.24,1314,4128078627
699992  2349,735564,417754,1988826,699991,1073676292,1631330259,1631328462,18415156832118.24,1086,3435846502
699993  2348,142889,85986,400847,699992,543162372,1631330419,1631328832,18415156832118.24,292,1518875222
699994  2347,173569,100581,475312,699993,536870916,1631330524,1631328892,18415156832118.24,202,2392312324
699995  2346,283467,174198,806061,699994,536879108,1631330826,1631329010,18415156832118.24,568,2475332722
699996  2345,448046,242806,1176464,699995,547356676,1631331088,1631329670,18415156832118.24,500,3674620545
699997  2344,686040,356633,1755939,699996,545259524,1631331729,1631330259,18415156832118.24,1014,1027539897
699998  2343,866328,471012,2279364,699997,549453828,1631332460,1631330419,18415156832118.24,1407,1111837075
699999  2342,98524,60994,281506,699998,536928260,1631332591,1631330524,18415156832118.24,213,974345258
700000  2341,141595,77043,372724,699999,832872452,1631332753,1631330826,18415156832118.24,292,648207636
700001  2340,1276422,907224,3998094,700000,1073733636,1631333672,1631331088,18415156832118.24,1276,2881644503
700002  2339,474051,320315,1434996,700001,536895488,1631333702,1631331729,18415156832118.24,496,2789376717
700003  2338,159985,101177,463516,700002,1073733636,1631333827,1631332460,18415156832118.24,255,3598498317
700004  2337,129039,82126,375417,700003,536870916,1631333877,1631332591,18415156832118.24,132,758642864
700005  2336,265800,118399,620997,700004,1073725444,1631334045,1631332753,18415156832118.24,320,67719332
700006  2335,408080,244339,1141097,700005,536870916,1631334474,1631333672,18415156832118.24,739,2729231674
700007  2334,141310,90876,413938,700006,536870916,1631334600,1631333702,18415156832118.24,208,1187060718
700008  2333,969802,554117,2632153,700007,671080452,1631335351,1631333827,18415156832118.24,1341,849197457
700009  2332,1385203,871487,3999664,700008,541065220,1631337023,1631333877,18415156832118.24,2464,418138790
700010  2331,496755,317213,1448394,700009,541065220,1631337205,1631334045,18415156832118.24,706,2599266873
700011  2330,76889,59953,256748,700010,545259524,1631337277,1631334474,18415156832118.24,129,2136474629
700012  2329,597090,321726,1562268,700011,545259524,1631337730,1631334600,18415156832118.24,873,3086641226
700013  2328,557834,320329,1518821,700012,549453828,1631338256,1631335351,18415156832118.24,965,573147392
700014  2327,123025,89777,392356,700013,545259524,1631338350,1631337023,18415156832118.24,221,747064107
700015  2326,574174,341443,1598503,700014,671080452,1631338971,1631337205,18415156832118.24,1103,2042562361
700016  2325,239090,128266,623888,700015,536870916,1631339219,1631337277,18415156832118.24,424,1858279548
700017  2324,1337733,885277,3993564,700016,691879940,1631341242,1631337730,18415156832118.24,2238,3429036652
700018  2323,1136955,952227,3993636,700017,536870916,1631341362,1631338256,18415156832118.24,1184,2027005034
700019  2322,1021426,629772,2910742,700018,536879108,1631341835,1631338350,18415156832118.24,1805,1176919477
700020  2321,1236433,918942,3993259,700019,536870916,1631342561,1631338971,18415156832118.24,995,3084275147
700021  2320,1362906,876793,3993285,700020,536870916,1631344275,1631339219,18415156832118.24,2217,1937397588
700022  2319,1392324,868681,3998367,700021,1073733636,1631346135,1631341242,18415156832118.24,2438,799453391
700023  2318,1390578,867694,3993660,700022,543162372,1631346167,1631341362,18415156832118.24,3120,2517992592
700024  2317,1397864,865206,3993482,700023,538968068,1631350038,1631341835,18415156832118.24,2894,2583383081
700025  2316,1617414,794041,3999537,700024,536870916,1631351026,1631342561,18415156832118.24,2762,838999041
700026  2315,1352185,882494,3999667,700025,1073733636,1631351172,1631344275,18415156832118.24,1927,1771543862
700027  2314,1385529,869218,3993183,700026,607584260,1631351224,1631346135,18415156832118.24,3153,1890021889
700028  2313,1443369,849904,3993081,700027,545259524,1631352489,1631346167,18415156832118.24,2655,4147666759
700029  2312,1135475,605062,2950661,700028,545259524,1631352574,1631350038,18415156832118.24,1781,1791017089
700030  2311,691048,384099,1843345,700029,536928260,1631353112,1631351026,18415156832118.24,1426,3397175967
```

Figure 3.3: Sample of the downloaded bitcoin dataset

Figure 3.3, give a glimpse of the cleanly collected dataset from bitcoin. All header attributes collected can be found in 3.1 and the comma was used to separate the attributes.

3.2.2 Simulation

Using the publicly available open-source code of the Ethereum network written in go-lang as the official implementation, a simulation of the Ethereum network was used to generate the required Ethereum dataset for this research. The go-lang implementation is the official implementation and the publicly available one currently. Simulation is an important factor of this research but at this stage, the simulation focuses only on generating block data that will be used in the training of the machine learning model. Nevertheless, the complexity of the network is still required to reflect the nature of the

real network. Therefore, the network started with a single node to allow understanding of the processes before the complexity. Web3j API was used with a python script to generate transactional traffic. The transactions were financial transactions generated randomly because such transactions are the easiest to generate and reflect the type of most of the records stored in the public blockchain. The simulation started with a single node and subsequently added more nodes. The data was generated by modifying the open-sources code to write the required block header attributes to a text file as seen in figure 3.2. The figure below shows a sample of the data generated.

```
602937  4716659596579711996,5041068,8000000,0,1603948675,4716659596581630388
602938  1147250099976107832,5038623,8000000,0,1603948703,1147250099976280267
602939  3396652198537476716,5041099,8000000,0,1603948705,3396652198537691019
602940  7406030387091791202,5043576,8000000,0,1603948709,7406030387092140001
602941  3498802345100610692,5046054,8000000,0,1603948714,3498802345101229740
602942  3663993849260569612,5048533,8000000,0,1603948723,3663993849261107409
602943  1214021333334961774,5051014,8000000,0,1603948731,1214021333334971555
602944  404852817264098196,5053496,8000000,0,1603948732,404852817264147861
602945  1519206260935441392,5055979,8000000,0,1603948733,1519206260938923490
602946  4836470214776637298,5053527,8000000,0,1603948781,4836470214779034672
602947  8547323945740734790,5051076,8000000,0,1603948816,8547323945742008018
602948  3041850903839757456,5048626,8000000,0,1603948834,3041850903840168217
602949  4509296254499889261,5051107,8000000,0,1603948840,4509296254500813158
602950  8862661320939686959,5048657,8000000,0,1603948853,8862661320939818772
602951  2235741718775876275,5051138,8000000,0,1603948855,2235741718778540956
602952  377821540114286886,5048688,8000000,0,1603948894,377821540114465061
602953  3880046987308171945,5051169,8000000,0,1603948896,3880046987309345684
602954  4842696350204914566,5048719,8000000,0,1603948913,4842696350205181045
602955  5417471655281982484,5051200,8000000,0,1603948917,5417471655282535666
602956  4238687465539585059,5053682,8000000,0,1603948925,4238687465543438238
602957  4163302098017157334,5051231,8000000,0,1603948980,4163302098018308522
602958  8143481318345967568,5048781,8000000,0,1603948997,8143481318347824261
602959  21430165211365543,5046332,8000000,0,1603949023,21430165213678879
602960  8355585598635243435,5043884,8000000,0,1603949056,8355585598635244767
602961  2567730038352917356,5046362,8000000,0,1603949057,2567730038354622484
602962  9213404410830402793,5043914,8000000,0,1603949081,9213404410830535530
602963  2127868787234289639,5046392,8000000,0,1603949083,2127868787234724990
602964  3444989568950526817,5048872,8000000,0,1603949089,3444989568951571576
602965  3790593296443158310,5046423,8000000,0,1603949104,3790593296446145750
602966  4212922621154594467,5043975,8000000,0,1603949147,4212922621156095740
602967  2455174193588499928,5041529,8000000,0,1603949169,2455174193589289100
602968  7227930096712148403,5044006,8000000,0,1603949180,7227930096712175419
602969  3193154456049604093,5046484,8000000,0,1603949181,3193154456049965600
602970  3796779804029569238,5048964,8000000,0,1603949186,3796779804031644476
602971  1652205446355869407,5046515,8000000,0,1603949216,1652205446357434893
602972  1895670413400208601,5044067,8000000,0,1603949238,1895670413400302234
602973  4096989564682535323,5046545,8000000,0,1603949240,4096989564682930692
602974  8366350349635284832,5049025,8000000,0,1603949245,8366350349640517589
602975  8849903241460004616,5046576,8000000,0,1603949320,8849903241461193481
602976  8813307446268972121,5044128,8000000,0,1603949337,8813307446269930147
602977  4606024502783159185,5041682,8000000,0,1603949351,4606024502785681879
602978  6768356602961549744,5039237,8000000,0,1603949387,6768356602962941900
602979  4314294025677544475,5036793,8000000,0,1603949407,4314294025678612667
602980  2574588327030548277,5034350,8000000,0,1603949423,2574588327032813681
602981  7475271463613402438,5031908,8000000,0,1603949455,7475271463615264431
```

Figure 3.4: Sample of the downloaded Ethereum dataset

Figure 3.4, give a glimpse of the cleanly collected dataset from the simulation. All header attributes collected can be found in 3.2 and the comma was used to separate the attributes.

The evaluation of the model also required a larger dataset therefore, the experiment ran for at least 9 months continuously until the evaluation stage was complete. Thus

the continued simulation generated more datasets used to evaluate the research model. The simulation has in total generated over 1.6 million block data as seen in the figure 3.5.

```
1637812  3008647073486401616,40506138,8000000,0,1640612978,3008647073506459526
1637813  2237444647758353143,40502744,8000000,0,1640613274,2237444647766035326
1637814  2503019350644359101,40499352,8000000,0,1640613387,2503019350645956246
1637815  6036497889929686992,40495961,8000000,0,1640613411,6036497889932692078
1637816  7659674361743304511,40492572,8000000,0,1640613455,7659674361749536798
1637817  5086142655569102141,40489185,8000000,0,1640613548,5086142655577177175
1637818  1089067082622345380,40485799,8000000,0,1640613667,1089067082666601504
1637819  208657179193946093,40482415,8000000,0,1640614319,208657179229846157
1637820  3781511949387351340,40479033,8000000,0,1640614849,3781511949394971008
1637821  3504787548856651579,40475652,8000000,0,1640614961,3504787548873835753
1637822  7984335338111931525,40472273,8000000,0,1640615215,7984335338115992232
1637823  6390682248298010866,40468896,8000000,0,1640615275,6390682248308303533
1637824  3471680329924680383,40465520,8000000,0,1640615427,3471680329931642529
1637825  7646163776136523229,40462146,8000000,0,1640615530,7646163776139761106
1637826  653250364234460458,40458774,8000000,0,1640615578,653250364236150320
1637827  5767888400327451741,40455403,8000000,0,1640615603,5767888400330608385
1637828  3593878728795347061,40452034,8000000,0,1640615650,3593878728802656014
1637829  4836726553519442243,40448667,8000000,0,1640615758,4836726553520781417
1637830  8662611844108163008,40445301,8000000,0,1640615778,8662611844119218998
1637831  5363846093841844769,40441937,8000000,0,1640615941,5363846093854607036
1637832  8668610564562176892,40438574,8000000,0,1640616129,8668610564569879522
1637833  7739626873383505228,40435213,8000000,0,1640616243,7739626873390640000
1637834  2476108952928986177,40431854,8000000,0,1640616348,2476108952988169800
1637835  2714681893857628893,40428496,8000000,0,1640617221,2714681893867393724
1637836  916988642242918958,40425140,8000000,0,1640617365,916988642244267004
1637837  380614314498853698,40421786,8000000,0,1640617385,380614314513670995
1637838  8979563560870030271,40418433,8000000,0,1640617604,8979563560878775943
1637839  1972767521423520561,40415082,8000000,0,1640617733,1972767521423564784
1637840  1391973600248874066,40451199,8000000,0,1640617734,1391973600253391697
1637841  3949646735633183740,40447832,8000000,0,1640617801,3949646735645833307
1637842  4738198860956196650,40444467,8000000,0,1640617987,4738198861010487703
1637843  3735792069372351855,40441103,8000000,0,1640618788,3735792069375873026
1637844  735691646754823554,40437741,8000000,0,1640618840,735691646757248349
1637845  2265375312629087665,40434381,8000000,0,1640618876,2265375312630031153
1637846  3933245567299036739,40431022,8000000,0,1640618890,3933245567299047575
1637847  2028862471941257501,40467147,8000000,0,1640618891,2028862471961008660
1637848  7159919065252170771,40463772,8000000,0,1640619181,7159919065258670953
1637849  361585508116873336,40460399,8000000,0,1640619277,361585508118685415
1637850  328819652654686932,40457027,8000000,0,1640619304,328819652695758726
1637851  6897503077919310296,40453657,8000000,0,1640619910,6897503077924518614
1637852  9222594944775940340,40450289,8000000,0,1640619986,9222594944780103852
1637853  2956301800737112829,40446922,8000000,0,1640620048,2956301800772541751
1637854  1363480070010503223,40443557,8000000,0,1640620570,1363480070028409338
1637855  4049019996548733963,40440194,8000000,0,1640620834,4049019996556882583
1637856  3220238801386879872,40436832,8000000,0,1640620955,3220238801396772685
1637857  8677360063645016115,40433472,8000000,0,1640621101,8677360063667798679
1637858  1351353718330746642,40430114,8000000,0,1640621437,1351353718342481700
```

Figure 3.5: Larger sample of the downloaded Ethereum dataset

Figure 3.5 shows the same details as 3.4, the only difference is that 3.5 indicate a higher line number to show how large the dataset is.

Another simulation with a slightly different approach was run, the aim was to generate data that will be used to analyse the behaviour of the model. This time, several simulations were run with 3 hours average run time and different manually set target difficulty values. The rationale behind this is to collect data that captures the behaviour of the network when the difficulty increases. The collected data was used in the data analyses next section.

3.2.3 Conclusion

Dataset	Size	Maximum Difficulty	Source
Ethereum Downloaded	100,000+	13,785,148	Etherscan
Bitcoin Downloaded	700,000+	15,011,455,676	BitcoinRPC
Ethereum Simulated	1.6 Million	306,579,150,222,389	Open Source Code

Table 3.1: Dataset comparison

Table 3.1 visualised all data collected to provide a clear quantitative view of the dataset. The size represents the total size collected, maximum difficulty represents the maximum difficulty target value within the dataset and the source tells the medium through which the data was collected.

Multiple sources were used to collect the required data for the study. A part of the collected data was used for the performance analyses to investigate the performance issues identified in the literature review. Another part of the data was used for the research experiment and the final dataset will be used to evaluate the research implementation. There was no particular limit to the amount of data needed, therefore data was continuously generated through the simulation.

3.3 Performance analysis

The analyses used both the collected data and the simulation process to identify the main contributing factor to the slow performance of the mining process. The analysis aims to identify the relationship between each attribute of the collected data and how they contribute to the performance or precisely how it supports the mining process. Using 10,000 generated block data, the research investigated these three processes

within the mining process to try ideas and get a first-hand picture of the challenges and provide insight into the nature of the dataset that will pave way for new ideas:

- The time taken for miners to collect a new block sent for mining was analysed and the result shows that it takes the miners an average of 3.65us to take possession of a new block to mine. Figure 3.6 shows a sample of the dataset collected for the analyses.

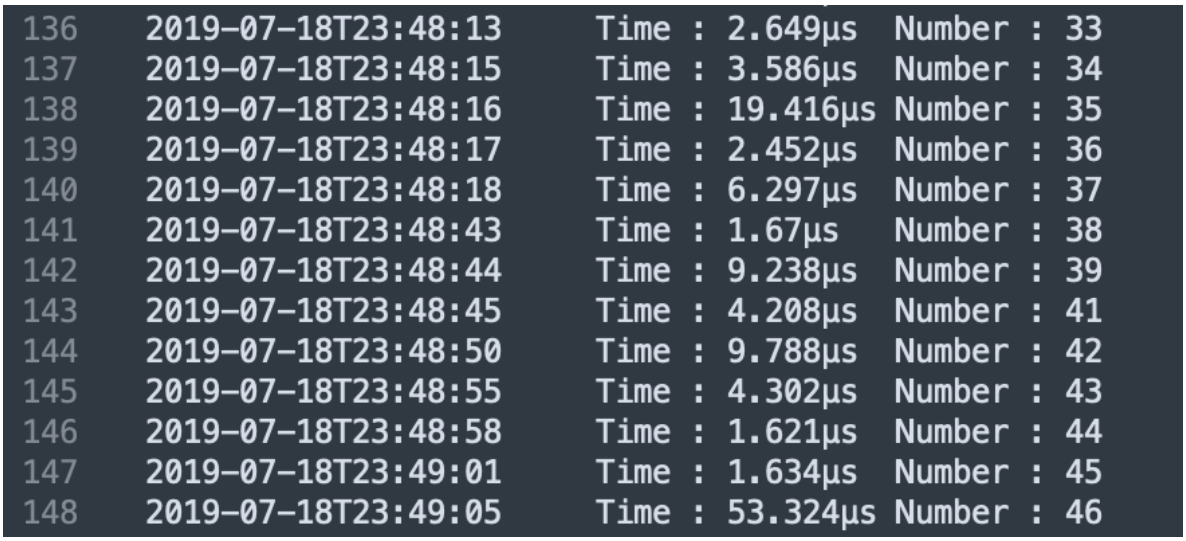

Figure 3.6: Sample dataset of block collection time

Figure 3.6 shows a sample of the dataset collected to analyse the time taken for a block to be collected for mining. The figure shows the date and time of the collection, the time that represents the duration of the collection process and the number representing the block number.

- The block generation time was analysed and the result shows that it takes an average of 2.4 seconds to generate a new block with 131072 (0x20000 in hex) set as the difficulty value. The result of the first miner is what is submitted as a new block, the work of other miners is discarded as mentioned in the review section in chapter 3. The process was further looked into and the nonce searching process was identified as the particular process that takes that time. Figure 3.7 shows a sample of the dataset collected for the analyses.

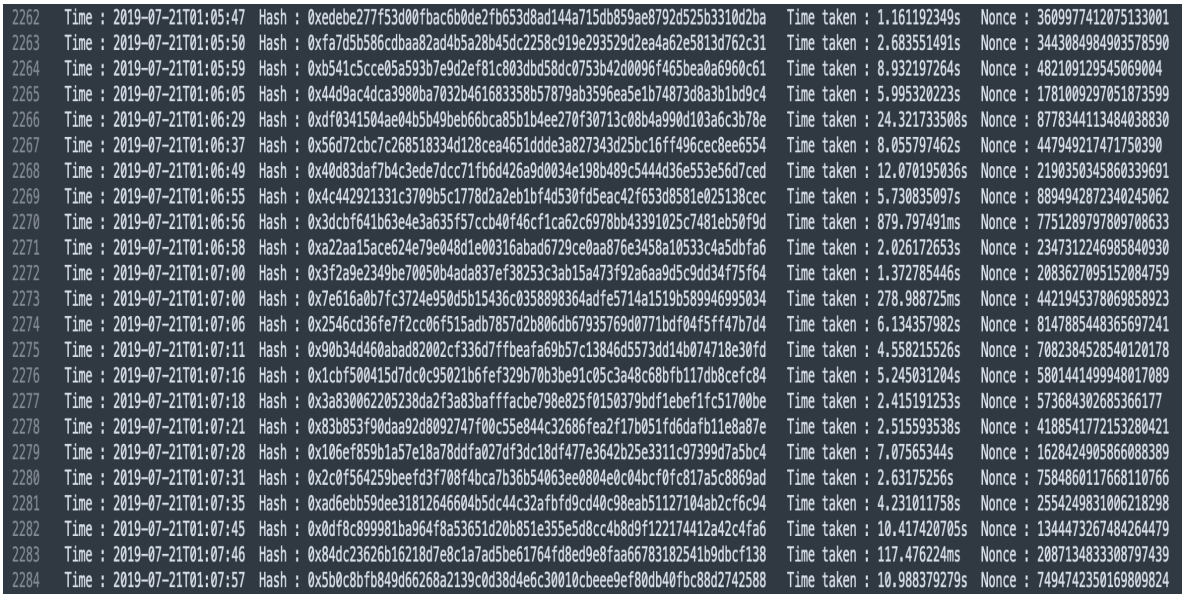

Figure 3.7: Sample dataset of nonce search time

Figure 3.7 shows sample of the dataset collected to analyse the time taken for miners to find a nonce value. The figure shows the date and time of the collection, hash representing the hash string found for the block, time taken that represent the duration of the mining process and nonce representing the nonce value.

- The time taken to insert a mined block into the main blockchain was analysed and the result identified the process to be a very simple one that takes an average of 290us. The figure 3.8 shows a sample of the dataset collected for the analysis.

```
500   2019-07-18T23:59:53 Time : 108.726µs    Block Size : 539.00
501   2019-07-18T23:59:53 Time : 97.166µs Block Size : 539.00 B
502   2019-07-18T23:59:53 Time : 108.8µs  Block Size : 539.00 B
503   2019-07-18T23:59:56 Time : 74.186µs Block Size : 539.00 B
504   2019-07-18T23:59:56 Time : 126.517µs    Block Size : 539.00
505   2019-07-18T23:59:58 Time : 91.516µs Block Size : 539.00 B
506   2019-07-18T23:59:58 Time : 201.796µs    Block Size : 539.00
507   2019-07-18T23:59:58 Time : 104.804µs    Block Size : 539.00
508   2019-07-18T23:59:58 Time : 197.192µs    Block Size : 539.00
509   2019-07-18T23:59:59 Time : 95.174µs Block Size : 539.00 B
510   2019-07-18T23:59:59 Time : 207.44µs Block Size : 539.00 B
511   2019-07-19T00:00:05 Time : 171.873µs    Block Size : 540.00
512   2019-07-19T00:00:05 Time : 136.467µs    Block Size : 540.00
513   2019-07-19T00:00:10 Time : 158.047µs    Block Size : 540.00
514   2019-07-19T00:00:10 Time : 161.813µs    Block Size : 540.00
515   2019-07-19T00:00:14 Time : 99.59µs  Block Size : 540.00 B
516   2019-07-19T00:00:14 Time : 182.333µs    Block Size : 540.00
517   2019-07-19T00:00:16 Time : 115.218µs    Block Size : 540.00
518   2019-07-19T00:00:16 Time : 157.293µs    Block Size : 540.00
519   2019-07-19T00:00:17 Time : 151.236µs    Block Size : 540.00
520   2019-07-19T00:00:17 Time : 235.576µs    Block Size : 540.00
```

Figure 3.8: Sample dataset of block insertion time

Figure 3.8 shows a sample of the dataset collected to analyse the time taken to insert a block onto the chain after mining. The figure shows the date and time of the collection, time representing the duration of the mining process and block size representing the size of the mined block.

The findings have practically addressed the main research question that asks for the main factors affecting the performance protocol by identifying the nonce search process as the most time taken process within the consensus protocol. The nonce needs to be found before transactions are confirmed by adding them into a block and adding the block to the blockchain. Also, they serve as the base for the research question that asks if speeding up the nonce searching process could improve the performance of the protocol. The behaviour/nature of the nonce value was analysed and the result shows that there is randomness in the nonce searching range as shown in figure 5.9 and the randomness increases with an increase in block records.

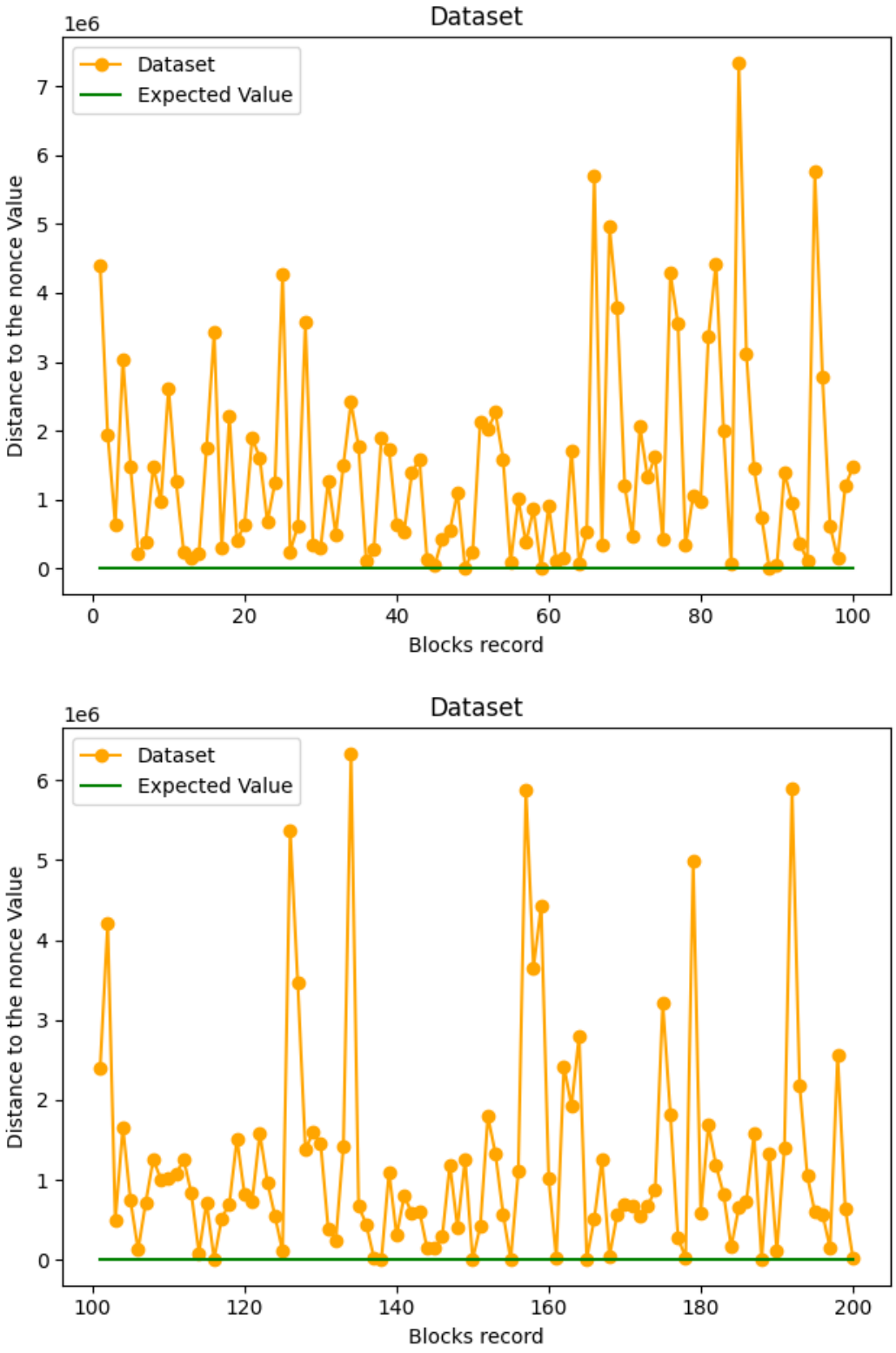

Figure 3.9: Mining distance to the nonce value

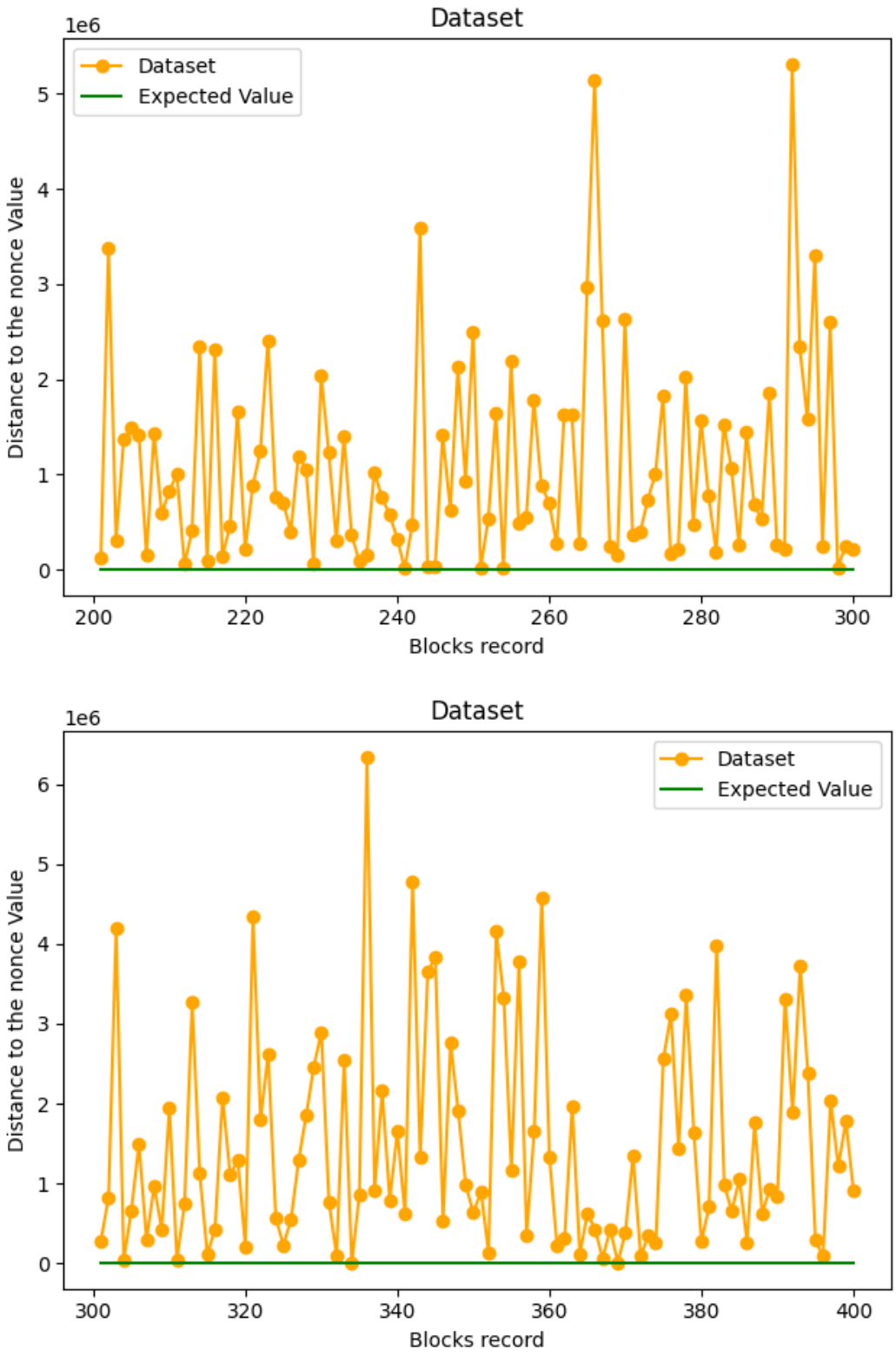

Figure 3.10: Mining distance to the nonce value

Figure 3.9 and 3.10demonstrates the behaviour of 400 nonce values part of the collected dataset. The 400 limit was to allow demonstrating the full result in a few graphs. The

distance to the nonce value represents the range between when searching begins to when the nonce is found. The block record represents the number of blocks used from the dataset. Four graphs were used to demonstrate the result of the 400 records instead of one to ensure readability.

Considering how it takes a long or short time to find the nonce value, if the short values were the only ones to be used in measuring the performance, it is logical the result will be faster compared to the current average. Therefore, the idea of predicting the value to reduce the nonce searching range was developed. Machine learning was seen as the appropriate technique for its ability to learn and predict.

3.3.1 Performance Growth Analyses.

To prove the problem mentioned in the problem statement regarding the performance decreasing with an increase in size. The research analysed the dataset to understand the behaviour of growth and have a clear pathway in choosing the right machine learning model and set a determinant of a significant improvement in the implementation. The average nonce search time was analysed based on a set difficulty value and the result shows there is a linear relationship between the increase in size and the average time taken to generate block as shown in the figure 3.11.

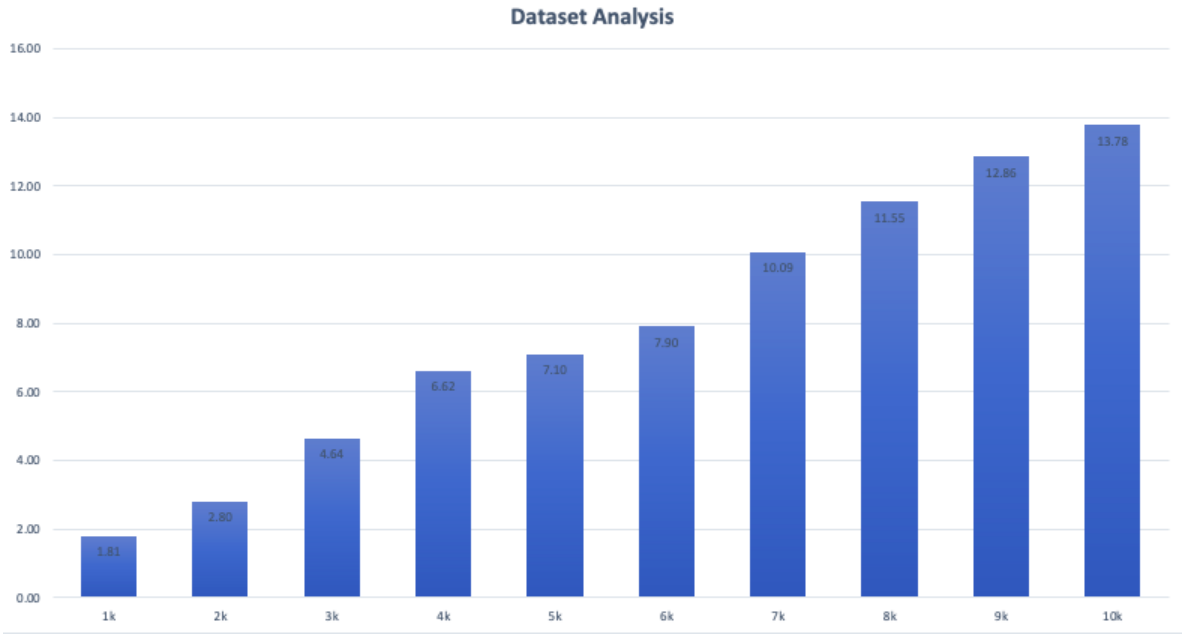

Figure 3.11: The behaviour of the dataset

The graph in figure 3.11 demonstrates the behaviour of the dataset after analysing the performance growth of the dataset. The x-axis represents the amount of block generated while the y-axis represents the average time it takes miner to find nonce value within the network size on the x-axis. As discussed in the literature review, the increase on the x-axis affects the y-axis because of the difficulty target value that increases and results in a longer mining time. The growth stops at some point because the difficulty target value is reduced when the performance goes over a certain average time to maintain a certain efficient level of performance. E.g. Bitcoin maintains a maximum of 10 minutes on average while Ethereum maintains 14 seconds.

3.3.2 Conclusion

The performance analyses have identified the nonce searching process within the block generation process as the most time-consuming aspect of the process. The nonce searching process does not affect any of the decentralisation or security aspects of the protocol. All it does is find the right nonce value that will result in the right block hash. Therefore, modification of the process will not result in scalability issues. Reducing the nonce search space will answer the research question that asked for the right methods of improving performance without altering or modifying any of the protocols.

3.4 Conclusion

The chapter has covered discussions on the dataset collection and analyse and the performance analyses of the PoW consensus protocol through simulation. It discussed the performance analysis and unpacked how the research gap and contribution were identified and highlighted how the results answered the first research question and serve as the basis to support the subquestions as well as one of the hypotheses. It also discussed how data was collected and the amount of data collected.

The pattern of the performance growth was analysed and the result demonstrates a linear growth in the time taken to find a nonce based on an increasing difficulty value. This informed the research about the right model that will best fit the dataset. But, the concern at this stage was whether the randomness of the nonce search shown in

figure 5.10 will allow a linear model to fit the dataset. The result of the experiments in the next chapter was able to address the concern.

Chapter 4

Experiment and Result

The previous chapter has practically analysed the research problem, investigated and addressed some of the research questions and concluded by identifying the behaviour of the dataset to be linear in growth. It also suggested the use of machine learning prediction ability to reduce the nonce search space and provided a collection of datasets for training and evaluating the machine learning model. This chapter focus on experimenting with different machine learning prediction models that fit a linear problem and identifying the model that shows the best results to address the research concern. A quantitative methodology will be used in choosing the right model based on the accuracy of the model.

4.1 Nonce booster Model

The nonce booster model refers to the model the research proposes to improve the performance of the PoW consensus protocol. The model provides optimised performance by reducing the time in the nonce searching process of the block generation process sometimes referred to as the mining process. To reduce the time, the model uses the machine learning ability to learn, identify patterns and predict a value to guide the nonce searching process by giving the miners a value closer to the nonce as the point from which the miners start searching for the nonce.

Ray (2019) emphasised the importance of choosing the right machine learning algorithm for the right problem. The research is informed and guided in choosing the right model by the performance analyses process where linear distribution was identified as the behaviour of the problem. The linearity is because there is a linear growth in the time taken to generate a block with an increase in network size and difficulty target value. The research uses the quantitative research methodology to compare linear problem-supported algorithms to select the best algorithm for the research problem base on accuracy.

Bzdok et al. (2018) have highlighted the importance of data in supervised machine learning as discussed in chapter 4. That is why the study used multiple sources to collect data that is enough to train, test and evaluate the model. Enough data in machine learning depends on the complexity of the problem (Kathole and Chaudhari, 2019). The research has in total collected 1.6 million records of the Ethereum block dataset and over 700,000 records of the Bitcoin block dataset used in implementing the nonce booster model.

4.1.1 Evaluation metrics

The model evaluation matrics used a comparison of its prediction accuracy with the accuracy of the dataset to determine the accuracy improvement. The accuracy was calculated by using the relative error (RE) formula in equation 4.1 is used to determine the magnitude of the absolute error in terms of the actual size of the measurement. The actual value is the nonce value and the predicted value is the point or value at which the search starts. The starting value for the Ethereum dataset is the seed value (as identified and mentioned in chapter 5), the Bitcoin dataset starts from 0 and the nonce booster model uses its prediction value.

$$RE = \frac{actual - predicted}{max(actual)} \tag{4.1}$$

Where (actual) is the nonce value, and the (predicted) value is the point or value at which the search starts

4.1.2 Performance metrics

The performance metrics focused on how good the prediction value is in improving the nonce search range. The goodness and effectiveness of the nonce booster model in reducing the searching range can be defined by how close its prediction is to the nonce value. The only certain thing regarding the model prediction at this stage is the predicted value will have to be a positive integer number with the possibility of sometimes going above the nonce value and in the case of Ethereum, going below the seed value. Therefore, the possible ranges of prediction are defined in figure 4.1 for the Ethereum dataset and 4.1 for the bitcoin dataset.

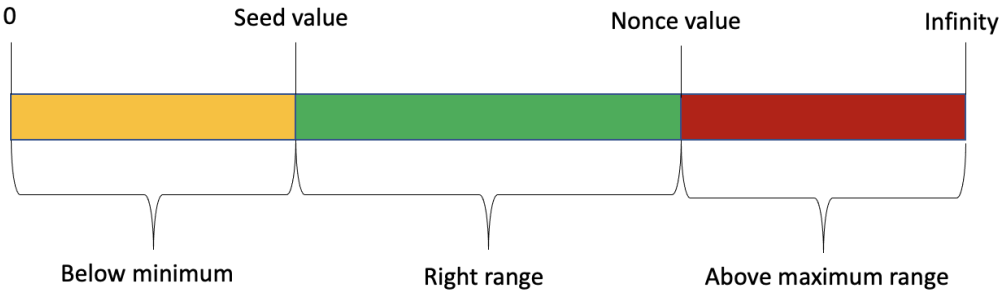

Figure 4.1: Ethereum prediction range description

- Below minimum: is a range within which if a value is predicted, the search will begin from a value lower than the normal starting value.

- Right range: is a range within which if a value is predicted, the search will begin from a value closer to the nonce than the normal starting value.

- Above maximum range: is a range within which if a value is predicted, the search will begin after a value higher than the nonce value leading to an infinite search.

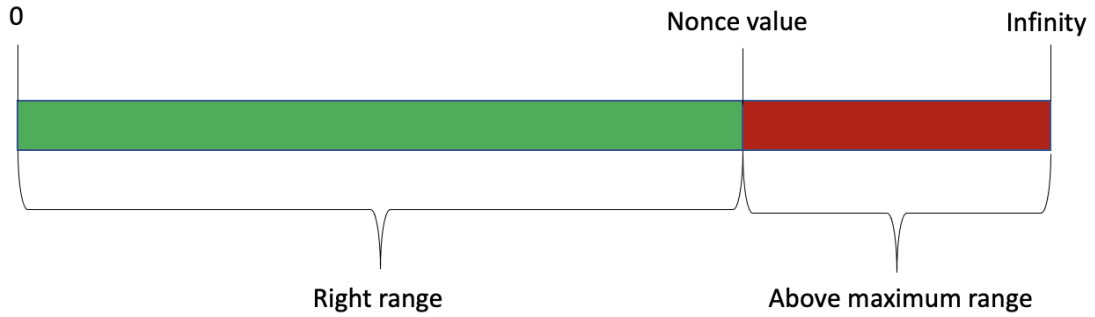

Figure 4.2: Bitcoin prediction ranges description

The model needs to predict values within the right range otherwise it will be a slower performance if the value predicted is below the minimum. The performance will be a disaster if the value predicted is within the above maximum range. The ranges for the Ethereum dataset differ from the Bitcoin datasets used in the evaluation because Bitcoin starts searching for the nonce value from zero, it does not have any value called seed value. Therefore, the Bitcoin dataset has only the right and above the right ranges as seen in figure 4.2.

4.2 Machine learning models

The section discussed the experiments of the machine learning models and their results. It uses a quantitative approach to determine the implementation that best addressed the research question using the performance and evaluation metrics of the nonce booster model. The section also serves as a proof of concept and the process of identifying the best fit model for the research problem. Therefore, the models were implemented with the 400,000 blocks of the dataset collected at the time. The ratio of 80, 20 percent of the dataset was used for training and testing respectfully in line with the recommendation of Gholamy et al. (2018) that suggested the use of 20-30 percent of the dataset for testing gives the best results as empirically studied. A different 400 records were used to test the accuracy and performance of the model. The 400 limit was to allow demonstrating the full result in a few graphs.

Machine learning algorithms selection is based on the algorithm's ability to solve the

research problem - a linear problem. The options are now a Linear Algorithm that can perform classification and regression on a linear problem or a support vector machine that can perform classification and regression on many problems specified through its kernel function including linear problems. Although the problem is linear, it is also important to understand the research problem requires prediction not classification therefore, all other linear algorithms such as Naïve Bayes and logistic regressions that are classifiers will not be useful to the research. Furthermore, to prove Bzdok et al. (2018) emphasis on choosing the right model, Random Forest Regression was selected in the hope of using its overfitting prevention to stop the prediction value from going above the maximum range as shown in figure 4.2 and 4.1. There is parity in all three selected algorithms in terms of their dependant and independent variable but differs in parameters because they use different equations.

4.2.1 Linear Regression Model

The model's technique is referred to as the most widely used statistic technique, it allows for a linear relationship between the dependant variable y and a single predictor variable x. The result of our behaviour analysis in section 4.2.1 suggests a linear model is the best fit for our dataset, therefore, we are going to implement and examine the result of the model with our dataset. The first step in the process is to define our dependent and independent variables, and in our case, the nonce value is the dependent variable that is depending on other headers attributes (independent variables) included in our dataset.

Let Y denote the "dependent" variable and $X_1, X_2, ..., X_k$ denote the "independent" variables and β denote the coefficient estimates that signifies the amount by which change in the dependent variable must be multiplied to give the corresponding average change in the independent variable (the nonce). Then the equation for computing the predicted value of y is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} \tag{4.2}$$

Linear equation

is to be fit to data, which we denote as

$$y_i = x_{i1}, x_{i2}, \cdots, x_{i,p-1}\,, \quad i =_1, \cdots, n \tag{4.3}$$

The observations y_i , where $i = 1, \cdots, n$, will be represented by a vector Y. The unknowns, $\beta_1, \cdots, \beta_{p-1}$, will be represented by a vector β. Let X_{n*p} be the matrix

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix} \tag{4.4}$$

For a given β, the vector of fitted or predicted values, $\bar{Y}$, can be written as

$$\underset{n*1}{\bar{Y}} = \underset{n*p}{X} \quad \underset{p*1}{\beta} \tag{4.5}$$

The least squares problem can then be phrased as follows: Find β to minimize

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{i,p-1}\,)^2 \\ &= ||Y - X\beta||^2 \\ &= ||Y - \bar{Y}||^2 \end{aligned} \tag{4.6}$$

if u is a vector, $||u||^2 = \sum_{i=1}^{n} u_i^2$

If we differentiate S with respect to each β_k and set the derivatives equal to zero, we see that the minimizers $\bar{\beta}_0, \cdots, \bar{\beta}_{p1}$ satisfy the p linear equations.

$$n\bar{\beta}_0 + \bar{\beta}_1 \sum_{i=1}^{n} x_{i1} + \cdots + \bar{\beta}_{p-1} \sum_{i=1}^{n} x_{i,p-1} = \sum_{i=1}^{n} y_i$$
$$\bar{\beta}_0 \sum_{i=1}^{n} x_{ik} + \bar{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{ik}, + \cdots + \bar{\beta}_{p-1} \sum_{i=1}^{n} x_{ik}x_i - p - 1 = \sum_{i=1}^{n} y_i x_{ik}, k = 1, \cdots, p-1 \tag{4.7}$$

These p equations can be written in matrix form

$$X^T X\bar{\beta} = X^T Y \tag{4.8}$$

and are called normal equations. But $X^T X$ is nonsingular in this case, therefore, the formal solution for each β is:

$$\bar{\beta} = (X^T X)^{-1} X^T Y \tag{4.9}$$

We will use the slope intercept formula to form the equation of the line and find our y-intercept and slope:

$$y = mx + b \tag{4.10}$$

Where (m) is the slope, (x) is the intercept and (b) is the y-intercept

To find our slope we use the formula:

$$m = \frac{y_2 - y_1}{x_2 - x_1} \tag{4.11}$$

Where (y) is the dependant variable and (x) is the independent variable

Using our experimentation environment and some open source libraries such as Scikit-learn (sklearn) which is used as a tool for predictive data analysis, the described model was successfully implemented.

Result

The linear regression prediction model was implemented and tested. The model resulted in 94.3 percent accuracy beating the dataset accuracy (92.8) by 1.5 percent. The range of prediction was examined to determine the model performance and 39.3 percent of the predicted values were identified to be within the right range, only one of the values was below the minimum and 60.4 percent were above the maximum. Even though the accuracy is promising, the percentage above the maximum was a big concern for the performance of the model, the concern was addressed section 4.3.

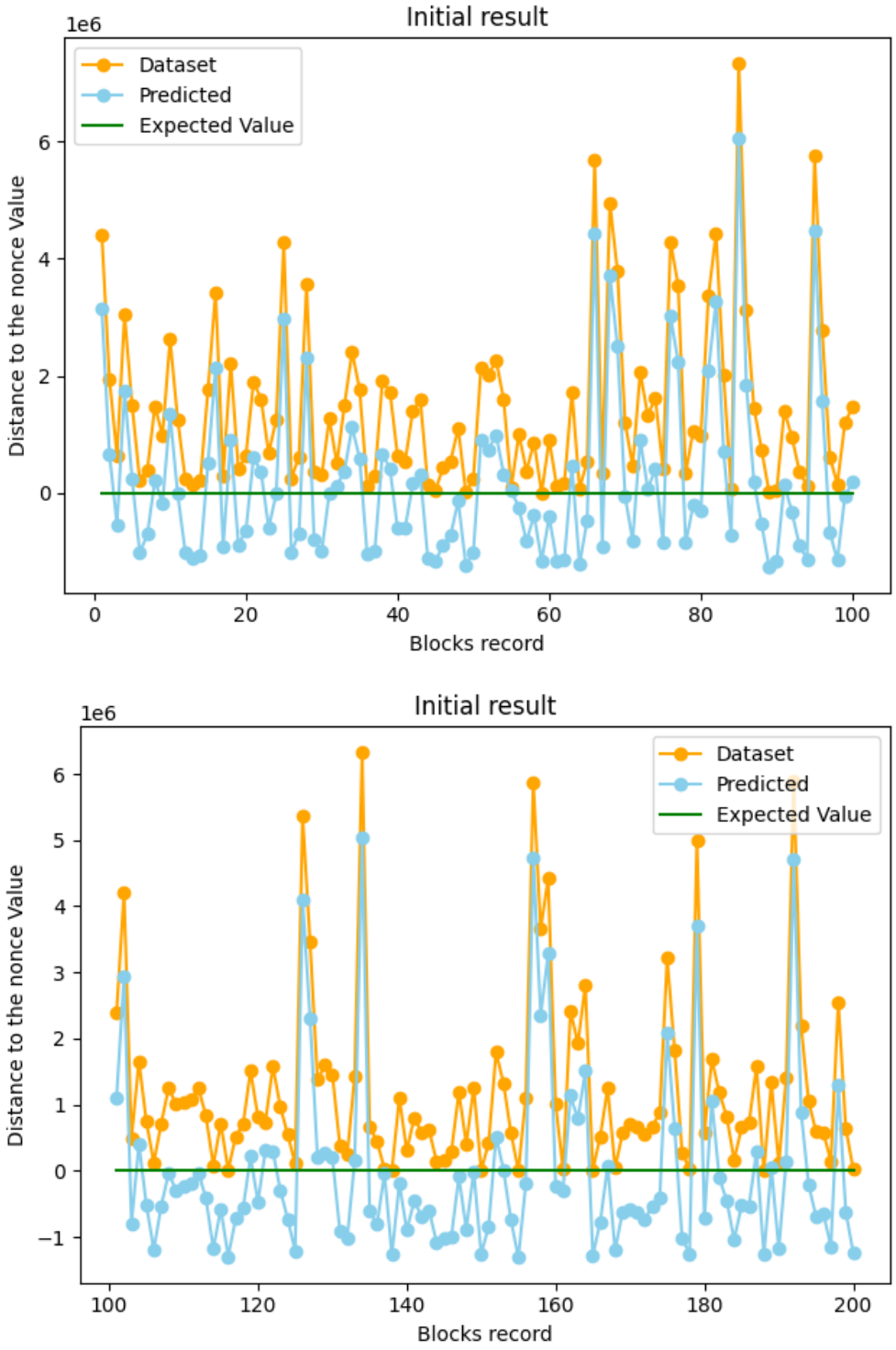

Figure 4.3: The distance of the prediction compared to the mining distance to the nonce value

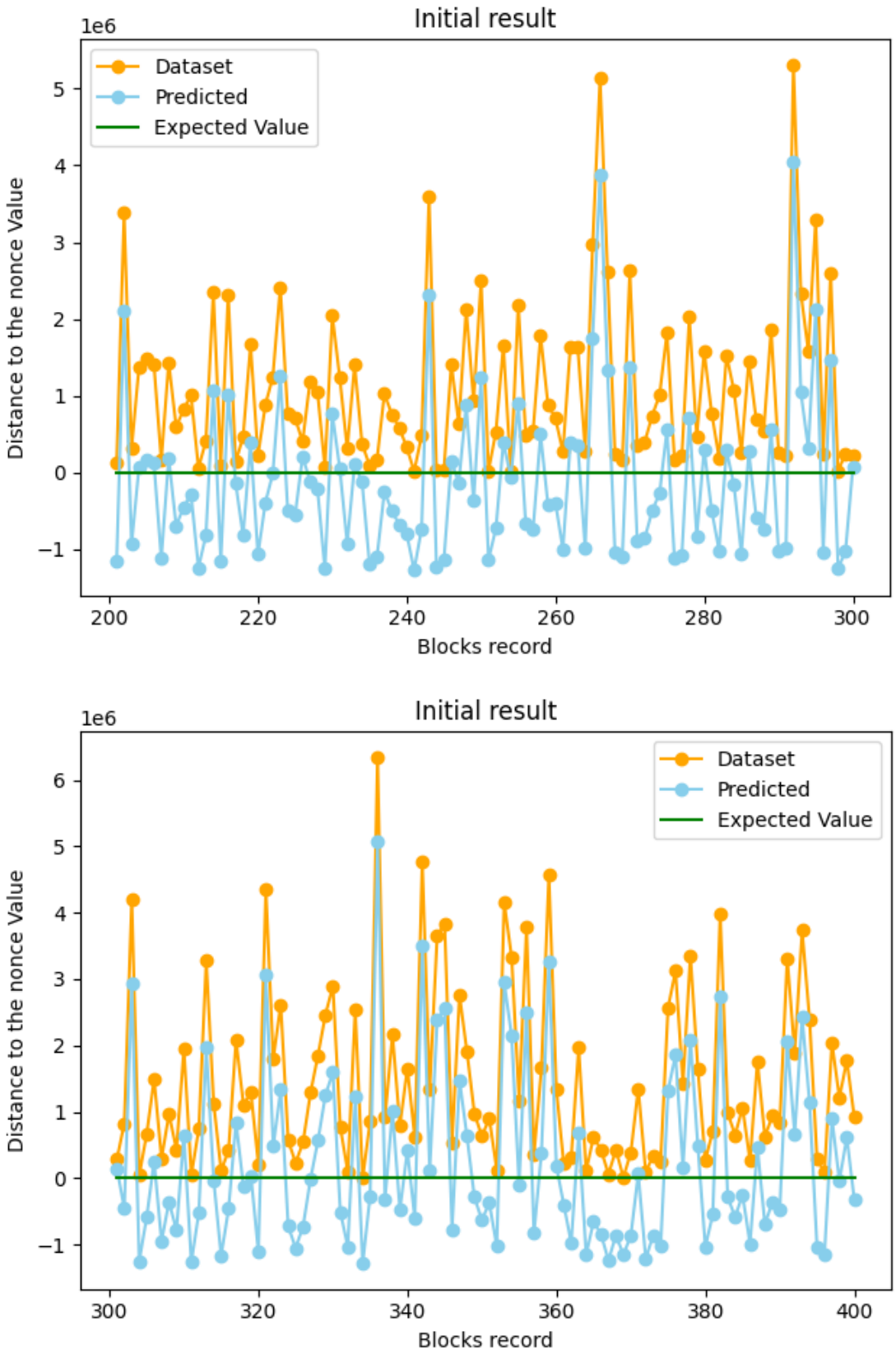

Figure 4.4: The distance of the prediction compared to the mining distance to the nonce value

Figure 4.3 and Figure 4.4 demonstrates the comparison between the prediction result and the dataset (the 400 records used). The 400 limit was to allow demonstrating the full result in a few graphs. The values that have gone below zero are those within the above maximum range. The distance to the nonce value represents the range between when searching begins to when the nonce is found. The block record represents the number of blocks used from the dataset. Four graphs were used to demonstrate the result of the 400 records instead of one to ensure readability.

4.2.2 Support Vector Regression (SVR)

SVR is the part of the Support Vector Machine SVM that supports regression and uses almost the same principle to perform classification and prediction. One of its main benefits is that it minimizes error and gives the flexibility to define how much error the model should accept and it uses a hyperplane in a higher dimension to fit the data. The experiment aims to explore the possibility of achieving maximum accuracy with a minimal error by utilising its principle of maximum margin that shows emphasis on minimizing error in the model prediction. The model has a function that works for different types of problems including linear and can be appropriately specified through the kernel function. Therefore, the model will be implemented with a linear kernel approach.

We need to find a function $f(x)$ with at most deviation from the target y, where x represents the header attributes and the y represents the nonce value. Now, our new objective function and constraints are as follows:

The function is:

$$MIN\frac{1}{2}||w||^2 \tag{4.12}$$

There is a possibility some of the points will fall out of the margin therefore, we need to account for the error that will be larger than $\in$. To do so, we need to add deviation to the function:

$$MIN\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}|\xi_i| \tag{4.13}$$

Result

The SVM model was implemented and tested. The model resulted in 43 percent accuracy which is 49.8 less than the dataset accuracy (92.8). The model gets stuck (run endlessly) when run with dataset values that are not scaled down regardless of the size of the training data. After investigating the reason for the poor performance, it was found that the values lose their accuracy by a big margin after scaling. The Ethereum dataset for example has a value of sixteen digits, scaling it down to one digit with decimals will require a lot of decimals to retain the accuracy. To keep consistency within the decimals means the value decimals get rounded at some points and that results in the values losing their accuracy when converted to the original scale. Therefore, it is not possible to have an accurate result if the model is trained on inaccurate data. To solve the scaling issue, the standard scaler used was changed to normalisation scaling methods provided as a module in the Sklearn library but to no positive effect. Regardless of the scaling approach, the model increases instead of reducing the nonce search space in the predictions.

A theoretical study of the poor performance was studied and Khoong (2021) suggested that SVM is not suitable for a large dataset because the training dataset grows with the dataset to a point where it becomes infeasible to train and use due to computing constraints. But the implementation experimented with a smaller dataset and the performance remains poor regardless. Khoong (2021) suggested of SVM performs poor in classification with an imbalanced dataset and Brus (2021) suggested regression problems are also affected by an imbalanced dataset. The different split ratios were used to turn the balance but there was no improvement in the accuracy.

4.2.3 Random Forest Regression

Random forest is a machine learning model that can be used for both classification and regression problems, which uses a decision tree approach in solving problems. Although the model is not both theoretically and practically the best for a linear problem, the

rationale behind experimenting with the model is to practically prove the point that a model that supports linear problems is the best for our dataset. Random forest prevents overfitting of data, it trains fast and has good accuracy because it takes the average of the result from each decision tree.

Result

Random forest regression was implemented and tested. The model resulted in 6.4 percent accuracy which is 86.4 less than the dataset accuracy (92.8). As expected from a model that is asked to solve a problem it is not fit for, the result shows the model accuracy is worst than the SVM. This further proves the importance of using the right algorithm for the right machine learning problem as discussed in the review in chapter 4. Therefore, the result provides a solid ground for the research to stick to using only linear models.

4.2.4 Conclusion

To ensure the best result and good performance in our implementation, it is imperative to make sure the research is working with the best model. Therefore, it is important to compare the experiment result of our linear regression, support vector regression and random forest model. The comparison is based on the evaluation and performance metrics provided for the nonce booster model. The dataset has an accuracy of 92.8 and a nonce searching range of 17,422,631.

Model	**Accuracy**	**Max Nonce Search Range**	**Behaviour fit**
Linear Regression	94.3	13,785,148	Linear
Support Vector Machine	43	15,011,455,676	Linear
Random Forest Regression	6.4	306,579,150,222,389	No

Table 4.1: Model comparison

From the table 4.1, the linear regression model results to be the most promising of the models both in terms of accuracy and the searching range of the nonce value. This is not surprising because our data behaviour analyses proved our dataset to be linear, thus, it is expected to work well with a linear model. The interesting part is SVR also uses the linear function yet it has a big difference from the linear model. This is because

SVR requires going through a dataset scaling process to ensure the dataset maintains a uniform range of values between all attributes. The analysis of the difference between the scaled value and unscaled value shows that the scaled values are in billions less than the unscaled value thus, the big difference in the searching range. Comparing other results to the random forest and other experiments with nonlinear models that are not documented here justifies why using a model that is inclined with the behaviour of the dataset is important.

4.3 Model Optimisation

The model optimisation started with the parameter turning process where the model experimented with different combinations of the dataset attributes. The combination with the highest consist of attributes that include height, difficulty, and time as general attributes for both Bitcoin and Ethereum and other attributes such as seed value peculiar to Ethereum.

The first question that comes to mind at this stage is: how to ensure the value does not go above the maximum when in reality the maximum value is not known? It is a tricky question that requires a thoughtful solution. The prediction was analysed to determine the difference between the predicted value and the maximum value. It was identified that subtracting the maximum difference between the predicted and maximum values where the predicted is higher than the maximum will bring back the value to the maximum or below.

Since there is no way to check each time the model predicts, if the value is above the maximum, the result of the initial experiment having 60 percent above the maximum value gives the basis for the assumption that most of the predictions will be above the maximum value. Therefore, the maximum difference between the predicted and maximum values from the result of the test data will be subtracted from each predicted value. Also, the result of the initial experiment having 39.3 percent of the predicted value in a good range indicates subtracting the maximum difference in each prediction might take some values outside the good range or below the minimum. Therefore, to ensure none goes below the minimum, the model was modified to always reset to the

value the miners normally start from when the predicted value goes below.

The experiment was carried out again with optimisation and both the model's prediction accuracy and range were improved. The result shows there is a value below the minimum, 203 values at the minimum, 208 within good range, no value above the minimum and there is exactly one value that is the same as the nonce. The new graph is demonstrated in figure 4.5 and 4.6

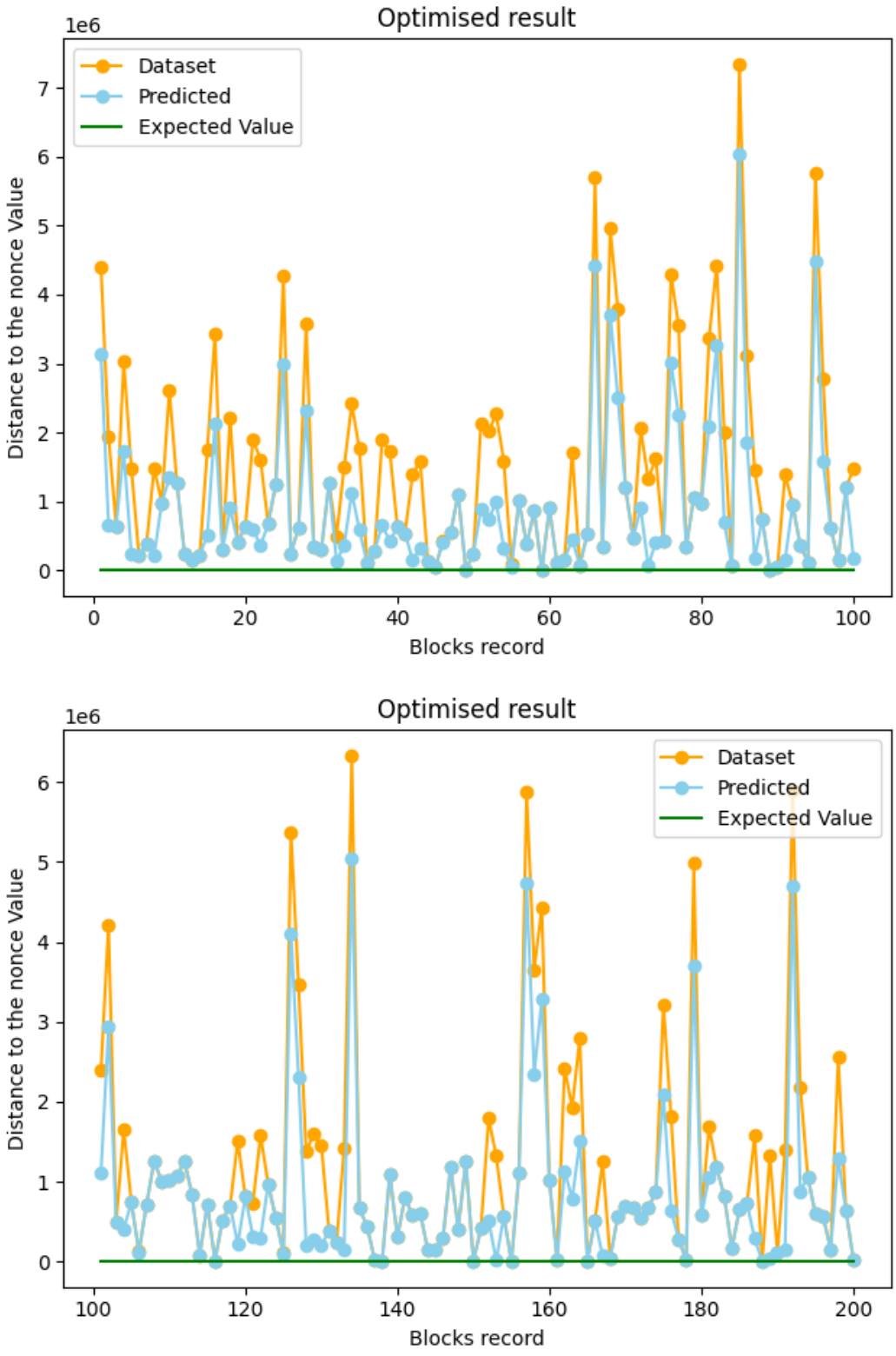

Figure 4.5: The distance of the prediction compared to the mining distance to the nonce value after optimisation

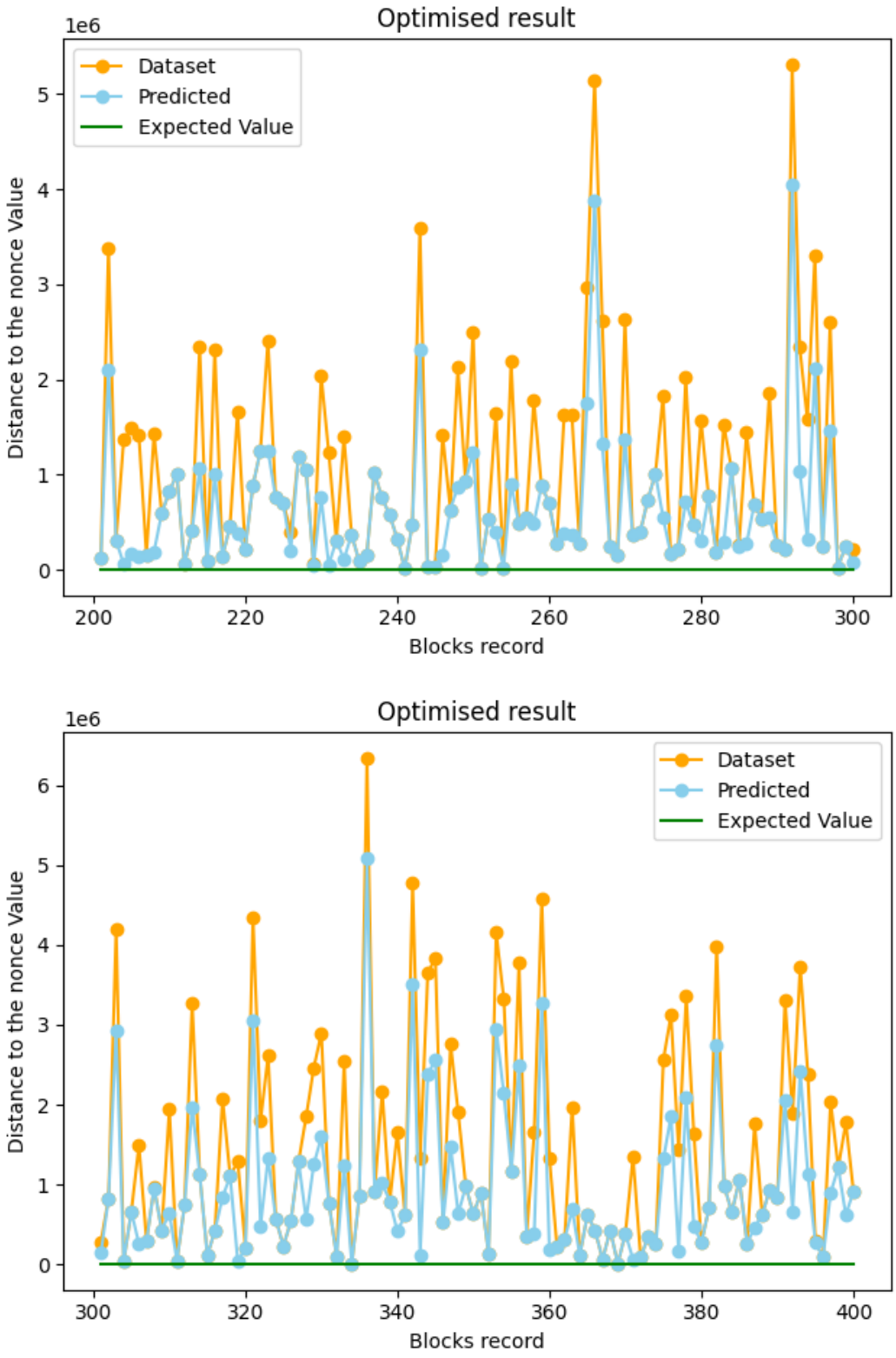

Figure 4.6: The distance of the prediction compared to the mining distance to the nonce value after optimisation

Figure 4.5 and 4.6 demonstrates the comparison between the prediction result and the

dataset (the 400 records used). The 400 limit was to allow demonstrating the full result in a few graphs. The distance to the nonce value represents the range between when searching begins to when the nonce is found. The block record represents the number of blocks used from the dataset. Four graphs were used to demonstrate the result of the 400 records instead of one to ensure readability.

4.4 Performance Evaluation

The discussion of the comparison of the experimented models in section 6.3 has identified the Linear Regression model as the best fit model for the research problem. Therefore, the model implementation in this section expects to result in improved performance of the protocol when tested. There are two implementations one for Ethereum and Bitcoin. The Ethereum implementation also has two implementations. 1) The initial implementation serves as a proof of concept. 2) The final implementation further demonstrates performance improvement.

Before the final implementation of the Ethereum implementation, to ensure efficient performance of the model, the model has been trained again using a larger dataset collected through download and simulation for both Bitcoin and Ethereum respectively as discussed in chapter 5. The implementation was done on another simulation similar to the one used in collecting the dataset. This time, it does not only generate data but also calls the nonce booster model through a restful service to get a number that will replace the seed value as the value from where the nonce search starts. The code used to call the RestFul web service is shown in figure 4.7.

```
167     requestBody, requestErr := json.Marshal(map[string]string{
168         "seed":       strconv.FormatUint(seed, 10),
169         "difficulty": header.Difficulty.String(),
170         "gasLimit":   strconv.FormatUint(header.GasLimit, 10),
171         "gasUsed":    strconv.FormatUint(header.GasUsed, 10),
172         "time":       strconv.FormatUint(header.Time, 10)})
173
174     if requestErr == nil {
175
176         resp, respErr := http.Post("https://localhost:8080/predict", "application/json", bytes.NewBuffer(requestBody))
177
178         if respErr != nil {
179             fmt.Printf(respErr.Error())
180         }
181
182         defer resp.Body.Close()
183
184         bodyBytes, err := ioutil.ReadAll(resp.Body)
185
186         // fmt.Println(string(string(bodyBytes)))
187         predicted, err := strconv.ParseUint(string(bodyBytes), 10, 64)
188         if err == nil {
189             // fmt.Println(string(predicted))
190             nonce = predicted
191         }
192
193     } else {
194         fmt.Println(requestErr.Error())
195     }
```

Figure 4.7: How the nonce booster model was implemented in the simulation

4.4.1 Model Integrated into Ethereum

To study and understand how our model can be integrated into the blockchain system and to determine the impact of our model on the performance of the PoW protocol, an application interface API was developed to allow the integration of the ML model in the simulation network. Before the start of the mining process, the mining protocol communicates with the nonce booster model through the API by sending the new block header attributes and asking for a point at which the miners should start looking for the nonce value. The point is the value predicted by the nonce booster model.

4.4.1.1 Initial Implementation

The initial implementation aimed to test and prove the research hypothesis that says, machine learning can be used to narrow down the block generation time and improve the performance of the proof of work consensus protocol. Previous discussions on experiments have demonstrated how the nonce booster model has been improved with good accuracy, the implementation aimed to determine or translate the improvement in accuracy to performance as in the research context. Different experiments were carried out to ensure the performance of the model after the implementation is captured within different difficulty target values. This was done by restarting the simulation and

manually changing the value of the difficulty target. The result of the evaluation is shown in figure 6.8.

Unlike the previous simulation that was started and allowed to keep running while generating more data and automatically increasing the difficulty, the simulation was restarted after a day with a new manually defined difficulty. This is to allow fast testing of our model instead of having to wait for the network to increase the difficulty automatically when it increases in size. Thus, the x-axis in figure 4.8 represents the difficulty value instead of the size of the blockchain as it was done in figure 3.11. The result shows and compares the average block generation time with and without our model implemented based on a set of different difficulty target values. It also shows improvement in the performance especially when the difficulty increases.

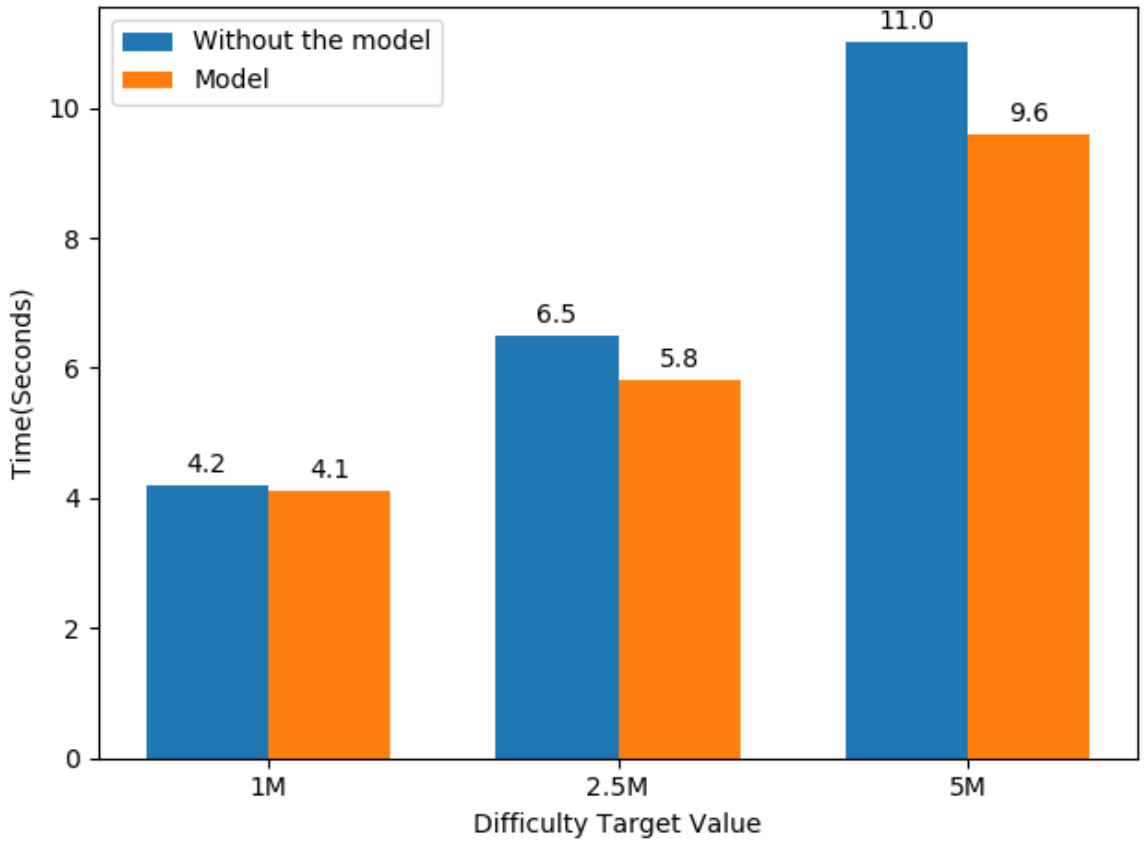

Figure 4.8: Performance improvement as a result of model implementation

Figure 4.8 demonstrate the comparison of the average block generation with and without the nonce booster model. The y-axis represents time in a section and the x-axis

represents the difficulty target value used. The x-axis started at 1M because there was no improvement in the lesser values, and it stopped at 5M because that is the highest difficulty value in the generated dataset used to train and test our model.

Result

The difficulty target value will look very small compared to the current difficulty value of Bitcoin and Etheruem but that justifies why the average time is also much lower. The lack of pattern in the improvement can be attributed to the random behaviour of the mining process as stated earlier. While can still benefit from more optimisation to ensure better performance improvement, the result has tested and proven the research hypothesis true.

4.4.1.2 Final Implementation

The final implementation of the nonce booster model only differs from the initial implementation in the use of a much higher difficulty target value and a larger dataset. A high target value was used to make the simulation network more complex. the nonce booster model was able to outperform the traditional mining process on the Ethereum simulation. The result showed a 30 percent accuracy improvement with a 96.3 percent new accuracy as seen in figure 4.9. It also demonstrated a 4 secs improvement on the block generation time as demonstrated in figure 4.10.

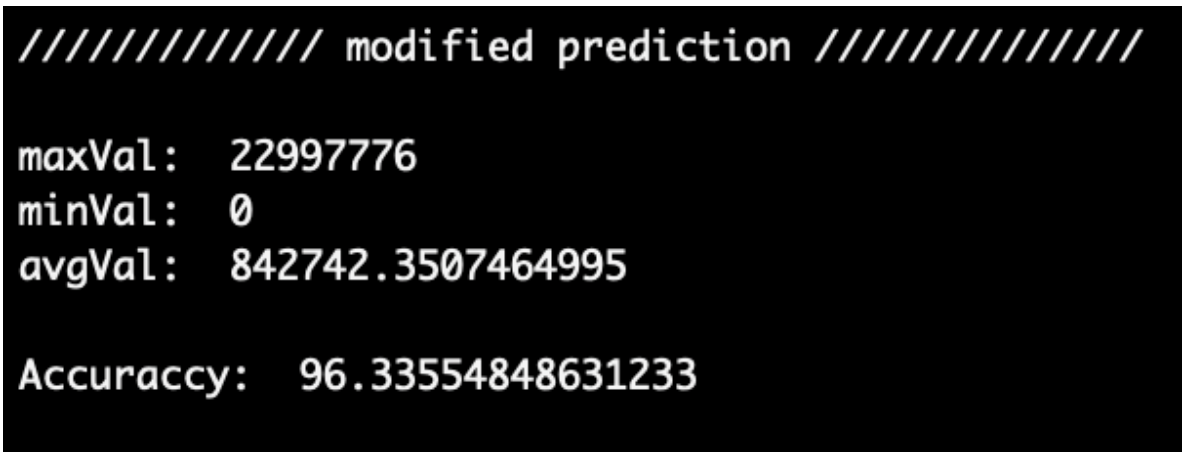

Figure 4.9: The accuracy of the final implementation

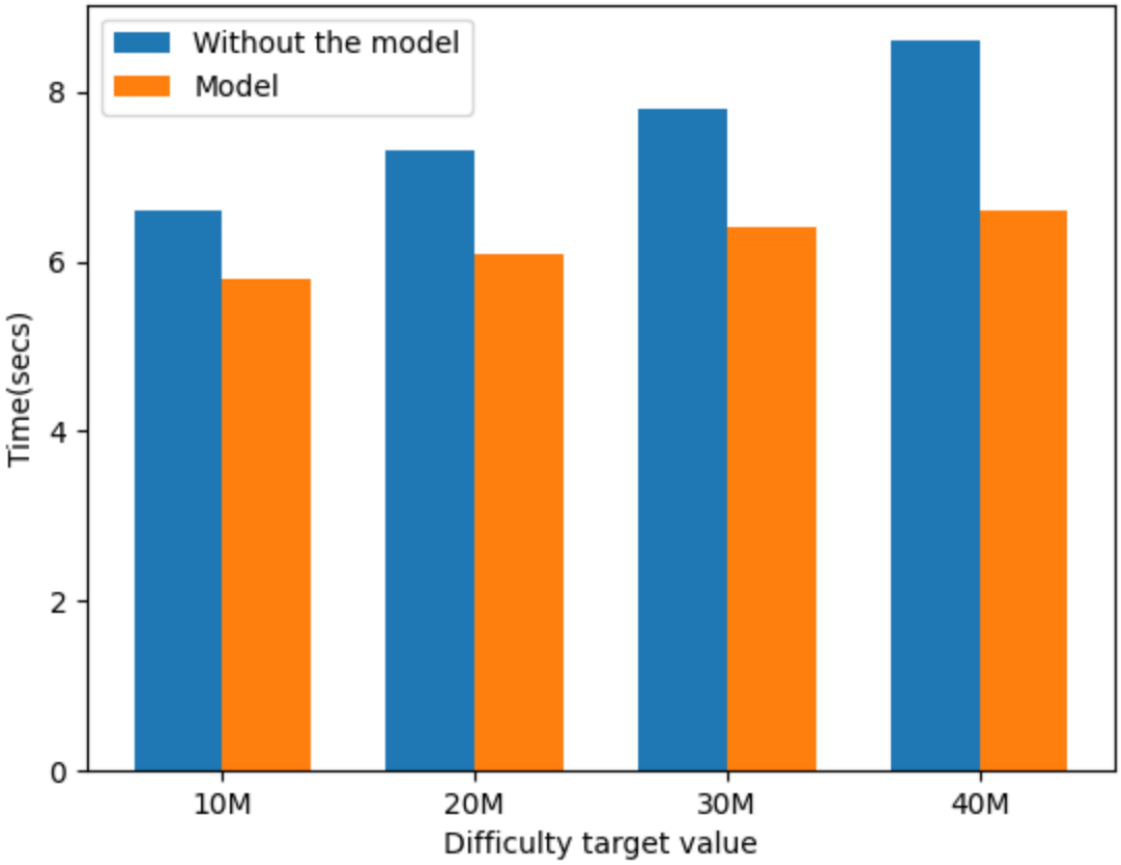

Figure 4.10: Performance improvement as a result of model implementation

Figure 4.10 demonstrate the comparison of the average block generation with and without the improved proposed model. The y-axis represents time in a section and the x-axis represents the difficulty target value used. The x-axis started at 10M to ensure balance in the scale, and it stopped at 40M because that is the highest difficulty value in the generated dataset used to train and test our model.

4.4.2 Model Integrated into Ethereum Bitcoin

This section discusses both the retraining of the research proposed model with the collected Bitcoin dataset and the implementation of the model in the mining process. Unlike Ethereum, both the Bitcoin data collection and implementation did not require simulation of the Bitcoin network. This is because the Bitcoin mining process has been separated from the official network implementation. The official Bitcoin client software has a built-in API called RPC (Remote Procedure Call) that allows interaction with

the blockchain system.

The RPC was used to access the blockchain data using a python script. The dataset was discussed and collected in the data collection section of chapter 5. The data structure was studied carefully before collecting the data to avoid data processing. Unlike the Ethereum implementation, the implementation didn't require a Restful service because both the nonce booster model and Bitcoin mining were implemented in python. Therefore, the model was only imported and a function call was used to get the predicted value as shown in figure 6.11.

```
498  def miner(coinbase_message, address):
499      while True:
500
501          # Get the new block record for mining
502          block_template = rpc_getblocktemplate()
503          weight = 0
504
505          # Extract the required information
506          for t in block_template['transactions']:
507              weight += int(t['weight'])
508          nTx = len(block_template['transactions'])
509          difficulty = rep_getdifficulty()
510          height = block_template['height']
511          version = block_template['version']
512          myTime = block_template['curtime']
513
514          # arrange the requied data for prediction.
515          data = [weight, int(height), int(version), int(myTime), int(difficulty), int(nTx)]
516
517          # Get the predicted value and use it as the new nonce value
518          predicted_value = model.getNonce(data)
```

Figure 4.11: How the nonce booster model was implemented in Bitcoin mining

Results

The model was trained with the seven hundred thousand plus record in the bitcoin blockchain as of the 26th of September 2021 using 80 and 20 percent of the data for training and testing respectively. The result shows 81.1 percent accuracy as seen in figure 6.12, which is a 58.6 percent improvement in accuracy when compared with the calculated accuracy of the dataset. The maximum nonce calculated in the dataset was 4.2 billion while the maximum for our model was 1.4 million. It means using the nonce booster model, bitcoin miners can take 2.8 fewer calculation steps to find the nonce

value.

```
////////////// Bitcoin Implementation //////////////

Data accuraccy: 54.53630563257747
Model accuraccy: 81.18990396524251
Improvement: 58.62611629679603
```

Figure 4.12: The accuracy of the Bitcoin implementation

The result indicates the Bitcoin implementation to be better than the Ethereum implementation even though the model was initially designed and implemented using the Ethereum dataset. Table 6.2 shows mining without the model starts from 0 while mining with the model starts from a number very close to the nonce. Unfortunately, it is not possible to simulate the Bitcoin network as we did with the Ethereum network. But using its approach, the model was successfully implemented with very impressive performance.

Block Number	Miner Start at	Model Prediction	Nonce
702930	0	1,086,201,030	1,118,746,415
702931	0	937,901,919	903,372,318
702932	0	952,056,427	1,142,720,264
702934	0	1,054,639,063	1,610,865,561

Table 4.2: Sample data from the Bitcoin implementation

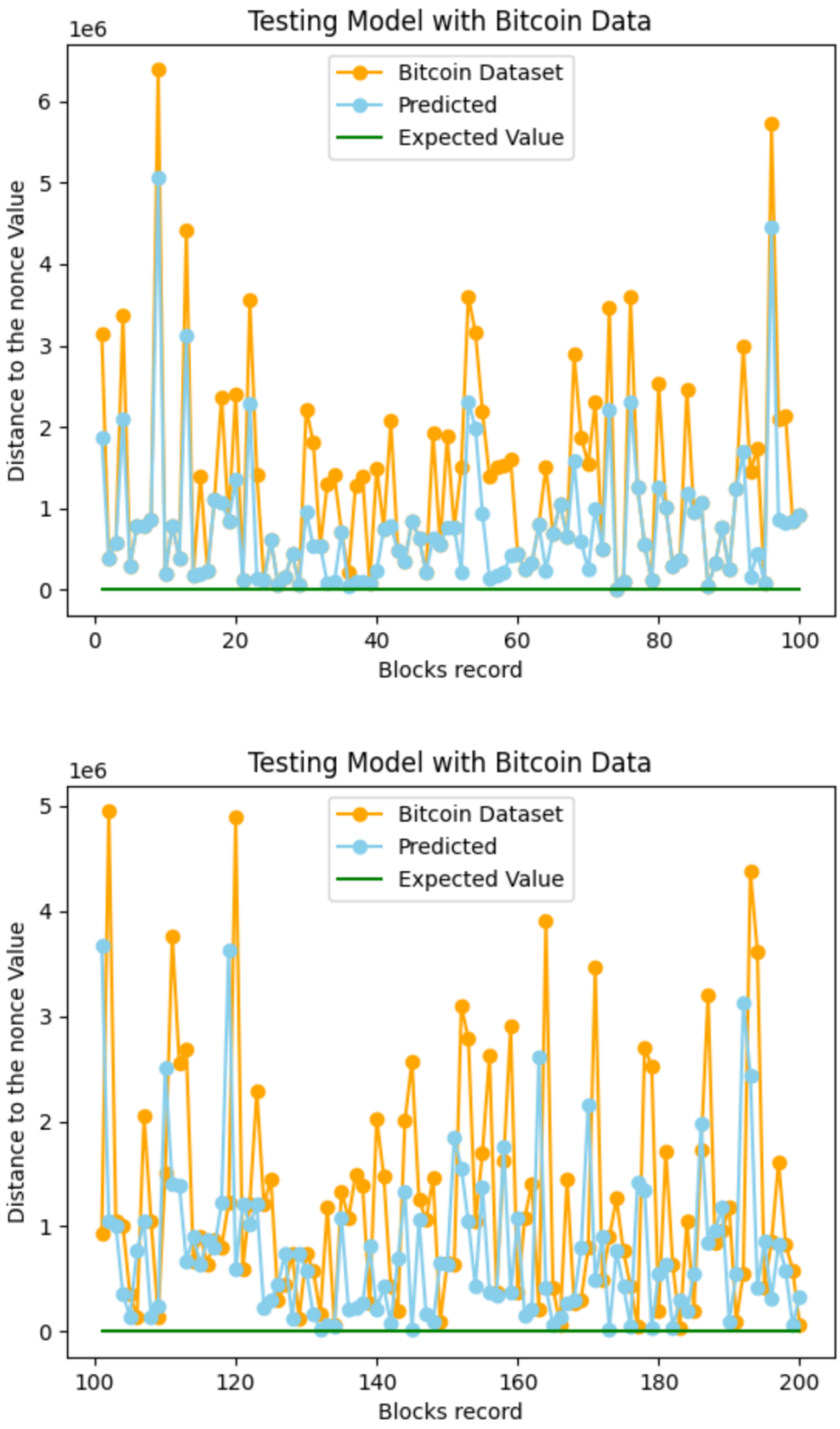

Figure 4.13: Results of the comparison between the bitcoin dataset and model prediction

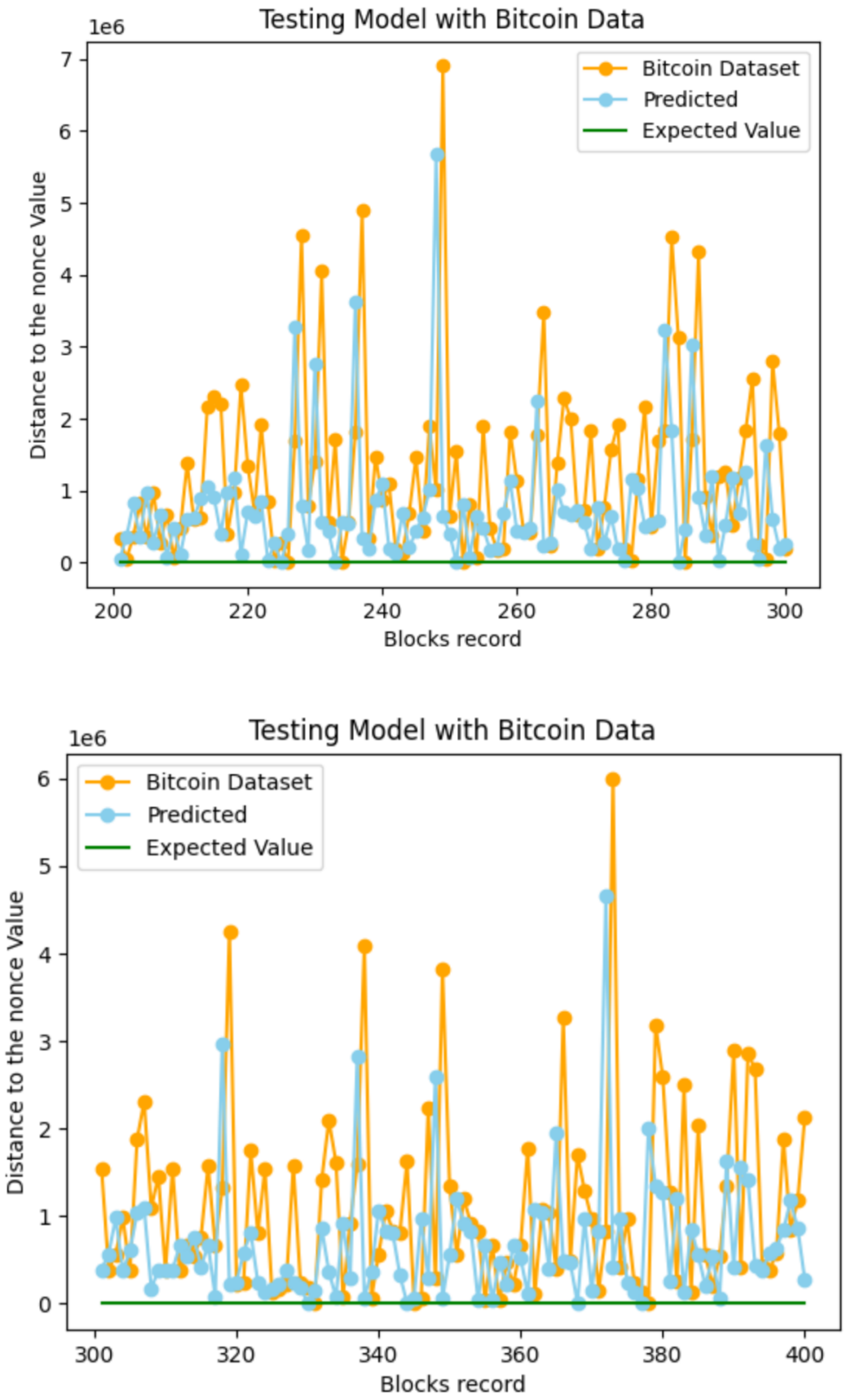

Figure 4.14: Results of the comparison between the bitcoin dataset and model prediction

Figure 6.13 and 6.14 demonstrates the comparison between the prediction result and the dataset (the 400 records used). The 400 limit was to allow demonstrating the full result in a few graphs. The distance to the nonce value represents the range between when searching begins to when the nonce is found. The block record represents the number of blocks used from the dataset. Four graphs were used to demonstrate the result of the 400 records instead of one to ensure readability.

4.4.3 Performance Evaluation Conclusion

The evaluation resulted in reducing the average time of producing blocks by 4 secs using a difficult target value of 40 million. It is a relatively very small difficulty value compared to the current Ethereum difficulty that reads in the trillions. But the research needed to use a difficulty that is within the dataset used to train the model to ensure good model performance and the highest difficulty in the dataset is 40 million. Nevertheless, the Bitcoin implementation has proved the efficiency of the model in a complex setting. The Bitcoin implementation resulted in a 70 percent accuracy which is a 65 percent improvement accuracy improvement against the dataset and a very reduced nonce searching range.

4.5 Conclusion

The chapter discussed the experiment, implementation and result of the nonce booster model. The comparison of the models in the experimentation part of the chapter has identified the linear regression model as the best fit model for the research problem. It has addressed the subquestion in the second part of the research question that asked for the most appropriate machine learning model for the research use case. It also tested true, the hypothesis that says machine learning techniques can be used to narrow down the block generation time and improve the performance of the proof of work consensus protocol. Improving the performance of the protocol by reducing the nonce search space has answered the second research question that asked for the right methods of improving performance without altering or modifying any of the protocols. The implementation part of the chapter has discussed the implementation of the nonce booster model in the research simulation of the Ethereum network and the performant implementation

of the Bitcoin mining process.

The implementation part of the chapter has addressed the third question in the research question asking that asked if applying machine learning techniques will improve the performance of the proof-of-work consensus protocol without sacrificing security or decentralisation. The research has successfully improved the performance of the protocol without trading-off security or decentralisation. The hypothesis that says improving the block generation time without altering any other part of the protocol will allow improved performance without affecting the current level of security or decentralisation has also been tested true.

Let us use this case study to understand and translate how much the nonce booster model has improved Bitcoin block generation time. If you are using a machine that can generate 1 million hashes a sec and you want to find the nonce value for block number 702930 shown in table 6.2. Using equation number 6.8 to calculate the expected time. The expected time to find the nonce value using the normal process is 18.6 minutes. It will take you approximately 1 minute to calculate if you are using the nonce booster model. Note, these values differ on average.

$$ET = \frac{N - SP}{HR}/60 \tag{4.14}$$

Expected time (ET), Nonce (N), Starting point (SP), Hashrate (HR).

Let us now reflect on the improvement in the speed of transactions. Assuming the average nonce finding time is 10 minutes and 5 for the normal process and using the nonce booster model respectively (it strongly depends on the machine's hash rate). If Bitcoin takes an average of 500 transactions per block (it does) and 10,000 transactions are waiting in the transaction pool. All transactions will be added into the blockchain in 3 hours 20 minutes using the normal process and 1 hour 40 minutes using the nonce booster model.

There is a fee called transaction fees in Bitcoin and a gas limit in Ethereum which users pay for their transactions to be given priority and be confirmed faster. The transaction with the highest fees is given priority when fetching transactions from the memory

pool. The option of this fee is used especially when there are a lot of transactions per day which will lead to a longer transaction confirmation time such as in January 2021 when the average transaction per second was 400,000. People's concerns about the transaction fees have made them pay an average high of 60 USD for transaction fees in 2021 (Vigna, 2021). Therefore, the nonce booster model not only improves the confirmation time by reducing the nonce search space but also reduces the need to pay extra to get a transaction confirmed faster.

The performance improvement by the nonce booster model is even more significant when you look at the application of the technology outside the financial sector. For example application of blockchain in the healthcare system, will require a more prompt process and access to data, especially in cases of emergency where there is life at stake. It requires a faster confirmation of records into the blockchain and there is no one to pay a transaction fee. the nonce booster model is positioned as the best option for implementation in healthcare because it provides decentralisation, security and improved speed.

Another example is the immigration system where there are millions of people travelling every day and records are expected to be prompt and accurate. A traveller cannot be asked to pay a fee if he wants his records to be recorded faster and that record needs to reflect fast so that he does have to wait or be denied at another. the nonce booster model is again positioned as the best option for implementation. Although in the case of extreme traffic, the nonce booster model might also not suffice in its current form. But it has shown and proved the ability to improve with an increase with a larger training dataset therefore, continuous training for the nonce booster model will be prudent for the nonce booster model. The result of the optimisation has given the research a reason to confidently believe the model is capable of adapting to the high traffic demand of different applications as long as it will be optimised and trained with a larger dataset.

Chapter 5

Conclusion and discussion

This chapter concludes the research work by summarising how it addresses the research questions and evaluated the hypothesis. It presents a summary of the research approaches, and contributions identified research limitations and suggest future work.

Blockchain has shown the promising potential of dominating the ledger space of the technology industry, although dominance remains far away, it surely has brought decentralised systems and cryptocurrencies to the limelight. The prominent aspect of the technology called the consensus protocol that ensures security and decentralisation have faced the daunting challenge of slowing performance which motivated the emergence of new protocols that ended up with the trilemma of having to trade one of security or decentralisation to get speed. The thesis has investigated the performance issue on the initial protocol (PoW) and was able to improve the performance without trading off security or decentralisation

5.1 Investigations and findings

The research started with the goal of solving the performance and scalability issue of blockchain technology by defining or finding the right protocol the research can improve to enable the system to scale without compromising security or decentralisation. The

literature review was carried out within the research area to gain insight into the whole idea of the technology. State of the art was also studied to gain a wider understanding of the issues, challenges, and how they impact the technology. The research has thoroughly examined the proof-of-Work consensus protocol, comparing it to other proposed solutions to the performance and scalability issue (such as Segwit, sharding and others) to learn from experience and identify limitations in the work done by other researchers, companies and industries to address the performance and scalability issues.

- The research believes that proof of work remains the most reliable protocol in terms of decentralisation and security yet the slowest in terms of transaction per second. Which is one of the reasons the research sticks with the protocol.

- The performance solutions are mostly peculiar solutions designed for their target application or use case and cannot be useful for other use cases. It is not feasible to have a whole new protocol designed for each use case. It is important to have a solution that is applicable to a different sector and use cases otherwise, this can only hinder the adoption of the technology.

- The process of finding the nonce value in the blockchain mining process was identified as the most time taking and power-consuming process within the consensus protocol. The delay in the process is what results in a higher block generation time

- Replacement of the mining and verification process with an approach that requires less number of participants is what led other solutions to have to trade security or decentralisation. The research focused on improving performance within the original design of the protocol.

5.2 Research Novelty and Results

Experiments were carried out to conceptualise the research contribution by analysing the technical process of the technology and testing research ideas gained through the study. The knowledge was used to build the research novel contribution that proposed providing an alternative to the time and power-consuming process of finding the nonce

value in the blockchain mining process without any modification to the protocol's architecture. The chosen alternative technique was the machine learning technique because of its prediction ability - The research used the technique to reduce the nonce search space by predicting a starting value for the miner that is closer to the nonce value. A review of the machine learning technique has been carried out and some models have been implemented and compared.

- The results of the experiments have shown a promising ability to achieve the set goals with a model accuracy of 96.3 and 70 percent for Ethereum and Bitcoin, which is a 30 and 58 percent improvement respectively when calculated by the improved percentage.

- The initial results published have shown promising ability by improving the accuracy by 18 per cent using a small dataset and a promising improvement in accuracy with an increase in training data. The result at the early stage shows our approach is promising in improving the block generation time which is also a big step in validating the research hypothesis

- The final results to be published have shown a significant improvement in performance by achieving a 30 percent accuracy improvement on the Ethereum model and 58 percent on the Bitcoin model

The results of the research model have shown performance improvement and tested the research hypothesis to be true. Thus, the research was able to achieve all aims and objectives to improve the performance of the proof-of-work consensus protocol using machine learning to predict a smaller nonce searching range for the miners.

5.3 Further Discussions

The evaluation result proves that the nonce booster model does not only reduce the waiting time for a transaction to be confirmed by reducing the nonce search space, but it also reduces the need to pay extra fees to get a transaction confirmed faster. It provides a solution that improves the performance of the technology across multiple sectors of its application such as healthcare and the immigration system.

Both research hypotheses tested true by successfully using machine learning to improve the block generation time without altering any other part of the protocol which allowed improved performance without affecting the current level of security or decentralisation. Thus, the research's main contribution can be noted as:

- Improved the overall performance of the blockchain and save cost. Because faster block generation time means transactions can be added to the blockchain faster and that reduces the transaction waiting time. This also means a reduction in the need for users to pay a transaction fee for his/her transaction to be confirmed faster - saving costs.

- Scaled solutions that reduce adoption concerns – the research provides a solution that increases speed without sacrificing security or decentralisation. Even more, the solution can only get better with more training datasets. Transaction confirmation has always been one of the stumbling blocks for the adoption of the technology in some applications and areas that require fast transaction confirmation.

- Reduced the power consumption for Bitcoin by reducing the amount of hash rate used and unused hash created.

The research addressed the research questions as follows:

- The first research question asked for the contributing factor to the performance issue, and the question was answered by identifying the nonce searching process of block generation as the main contributing factor in the performance analyses.

- Question 2 asked for the right technique that can be used to speed up the nonce search process and was answered with the proposal of using the machine learning prediction technique in the performance analyses.

- Question 3.1 asked for the most appropriate machine learning model for the research use case and was answered when the result of the model comparison found linear regression as the best fit model in the research experiment.

- Question 3 asked if applying machine learning techniques improves the performance of the proof-of-work consensus protocol without sacrificing security or decentralisation, and question 2.1 asked if speeding up the nonce search improves the performance without facing the scalability issue were both answered by the amazing results of the nonce booster model.

5.4 Limitations and Future Work

As stated in the previous section, this research presented several contributions to the performance of blockchain technology. Despite the promising results demonstrated in the performance evaluation, it has some limitations and room for improvements and future research directions.

- The difficulty value in the training dataset is very much lower than the current difficulty value in the real Ethereum network. Therefore, the nonce booster model cannot fit in the main Ethereum network or any other with similar complexity until the model is optimised and retrained with accurate data. One way of achieving this is through accessing all blockchain data in the Ethereum network but the problem with this approach is mining in the Ethereum network does not start from zero. It starts from a value called seed that is not found on the blockchain data. As discussed in chapter 5, the seed value was the missing data in the initial downloaded dataset and a critical parameter in the accuracy of the nonce booster model. Alternatively, the data can be generated through simulation while making sure the difficulty value and the number of transactions reflect the Ethereum network.

- The machine learning model will always benefit from more training datasets and it is important to ensure the dataset is complex and accurate enough as the target application of the model. The growing size of the blockchain networks is a concern for the nonce booster model accuracy because of the increase in the difficulty value, the change in timestamp and other parameters used in the nonce booster model training. The more these values chance, the higher it affects the accuracy of the model. As mentioned in the machine learning discussion, the

technique learns from experience and the more the parameters chance the more the model losses familiarity with the data. Therefore, it's important to ensure the model continues learning as more data is generated to ensure the model retains good accuracy at all times.

- Although the data-exploratory process has identified the average behaviour of the dataset to be linear and a linear regression model implemented has proven to be effective in achieving the goal, it is important to explore and find a machine learning algorithm that best fits the randomness of the unaveraged data behaviour as identified in the performance analyses process.

Despite the above-mentioned limitation, the research model is so far the only solution proposed to solve the blockchains' s performance issue that completely avoided trading off security or decentralisation. The research provided a significant contribution to the technology, especially in the industry where the hindering performance of one of the most reliable cryptocurrencies (Bitcoin) is improved, thus, encouraging wider adoption. It also provides a base for reference and learning in the academic sector and a milestone of research other researchers can build on.

References

Abdelsamea, A., El-Moursy, A.A., Hemayed, E.E., Eldeeb, H., 2017. Virtual machine consolidation enhancement using hybrid regression algorithms Virtual machine consolidation enhancement. Egyptian Informatics Journal 18, 161–170. URL: https://doi.org/10.1016/j.eij.2016.12.002, doi:10.1016/j.eij.2016.12.002.

Akcora, C.G., Li, Y., Gel, Y.R., Kantarcioglu, M., 2019. Bitcoinheist: Topological data analysis for ransomware detection on the bitcoin blockchain. arXiv:1906.07852.

Al-Jaroodi, J., Mohamed, N., 2019. Blockchain in Industries: A Survey. IEEE Access 7, 36500–36515. doi:10.1109/ACCESS.2019.2903554.

Al-Saqaf, W., Seidler, N., 2017. Blockchain technology for social impact: opportunities and challenges ahead. Journal of Cyber Policy 2, 338–354. doi:10.1080/23738871.2017.1400084.

Alketbi, A., Nasir, Q., Talib, M.A., 2018. Blockchain for government services — use cases, security benefits and challenges, in: 2018 15th Learning and Technology Conference (L T), pp. 112–119. doi:10.1109/LT.2018.8368494.

Aste, T., Tasca, P., Matteo, T.D., 2017. Blockchain Technologies: The Foreseeable Impact on Society and Industry. Computer 50, 18–28. URL: https://ieeexplore.ieee.org/document/8048633, doi:10.1109/MC.2017.3571064.

Atlam, H.F., Alenezi, A., Alassafi, M.O., Wills, G., 2018. Blockchain with internet of things: Benefits, challenges, and future directions. International Journal of Intelligent Systems and Applications 10, 40–48.

Atzei, N., Bartoletti, M., Cimoli, T., 2017. A survey of attacks on ethereum smart contracts sok, in: Proceedings of the 6th International Conference on Principles

of Security and Trust - Volume 10204, Springer-Verlag, Berlin, Heidelberg. p. 164–186. URL: https://doi.org/10.1007/978-3-662-54455-6_8, doi:10.1007/978-3-662-54455-6_8.

Bach, L.M., Mihaljevic, B., Zagar, M., 2018. Comparative analysis of blockchain consensus algorithms, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1545–1550. doi:10.23919/MIPRO.2018.8400278.

Bayer, D., Street, S., Stornetta, W.S., 1992. Improving the Efficiency and Reliability of Digital , 1–6.

Beck, R., Czepluch, J.S., Lollike, N., Malone, S., 2016. Blockchain - the gateway to trust-free cryptographic transactions, in: ECIS.

Bernardi, L., Mavridis, T., Estevez, P., 2019. 150 successful machine learning models: 6 lessons learned at Booking.com. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , 1743–1751doi:10.1145/3292500.3330744.

Bez, M., Fornari, G., Vardanega, T., 2019. The scalability challenge of ethereum: An initial quantitative analysis. Proceedings - 13th IEEE International Conference on Service-Oriented System Engineering, SOSE 2019, 10th International Workshop on Joint Cloud Computing, JCC 2019 and 2019 IEEE International Workshop on Cloud Computing in Robotic Systems, CCRS 2019 , 167–176doi:10.1109/SOSE.2019.00031.

Binkhonain, M., Zhao, L., 2019. A review of machine learning algorithms for identification and classification of non-functional requirements. Expert Systems with Applications: X 1. doi:10.1016/j.eswax.2019.100001.

Boulos, M.N.K., Wilson, J.T., Clauson, K., 2018. Geospatial blockchain: promises, challenges, and scenarios in health and healthcare. International Journal of Health Geographics 17.

Bravo-Marquez, F., Reeves, S., Ugarte, M., 2019a. Proof-of-Learning: a Blockchain Consensus Mechanism based on Machine Learning Competitions. Proceedings of

the 2019 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPCON) URL: https://www.researchgate.net/publication/330753314.

Bravo-Marquez, F., Reeves, S., Ugarte, M., 2019b. Proof-of-learning: A blockchain consensus mechanism based on machine learning competitions. 2019 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPCON) , 119–124.

Brus, P., 2021. Data imbalance in regression. URL: https://towardsdatascience.com/data-imbalance-in-regression-e5c98e20a807.

Bzdok, D., Krzywinski, M., Altman, N., 2018. Points of significance: Machine learning: Supervised methods. Nature Methods 15, 5–6. doi:10.1038/nmeth.4551.

Cachin, C., Vukolić, M., 2017. Blockchain consensus protocols in the wild. Leibniz International Proceedings in Informatics, LIPIcs 91, 1–16. doi:10.4230/LIPIcs.DISC.2017.1.

Castro, M., Liskov, B., 1999. Practical Byzantine Fault Tolerance. Technical Report.

Celebi, M.E., Aydin, K., 2016. Unsupervised learning algorithms. Unsupervised Learning Algorithms , 1–558doi:10.1007/978-3-319-24211-8.

Chase, B., MacBrough, E., 2018. Analysis of the XRP Ledger Consensus Protocol URL: http://arxiv.org/abs/1802.07242.

Chaudhry, N., Yousaf, M.M., 2018. Consensus Algorithms in Blockchain: Comparative Analysis, Challenges and Opportunities, in: 2018 12th International Conference on Open Source Systems and Technologies (ICOSST), pp. 54–63. doi:10.1109/ICOSST.2018.8632190.

Chauhan, A., Malviya, O.P., Verma, M., Mor, T.S., 2018. Blockchain and Scalability, in: 2018 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C), pp. 122–128. doi:10.1109/QRS-C.2018.00034.

Chaumont, G., Bugnot, P., Hildreth, Z., Giraux, B., 2019. DPoPS : Delegated Proof-of-Private-Stake , a DPoS implementation under X-Cash , a Monero based hybrid-privacy coin , 1–46.

Chen, H., Pendleton, M., Njilla, L., Xu, S., 2019a. A survey on ethereum systems security: Vulnerabilities, attacks and defenses. arXiv:1908.04507.

Chen, X., Ji, J., Luo, C., Liao, W., Li, P., 2018. When machine learning meets blockchain: A decentralized, privacy-preserving and secure design, in: 2018 IEEE International Conference on Big Data (Big Data), pp. 1178–1187. doi:10.1109/BigData.2018.8622598.

Chen, X., Ji, J., Luo, C., Liao, W., Li, P., 2019b. When Machine Learning Meets Blockchain: A Decentralized, Privacy-preserving and Secure Design. Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018 , 1178–1187doi:10.1109/BigData.2018.8622598.

Clarke, S., Craig, I., Wyszynski, M., 2018. Litecoin Cash : The best of all worlds SHA256 Cryptocurrency , 1–9URL: www.litecoinca.sh/downloads/launch{_}whitepaper.pdf.

Conti, M., Sandeep, K.E., Lal, C., Ruj, S., 2018. A survey on security and privacy issues of bitcoin. IEEE Communications Surveys and Tutorials 20, 3416–3452. doi:10.1109/COMST.2018.2842460.

Cui, Z., Gong, G., 2018. The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. NeuroImage 178, 622–637. URL: https://doi.org/10.1016/j.neuroimage.2018.06.001, doi:10.1016/j.neuroimage.2018.06.001.

Dang, H., Dinh, T.T.A., Loghin, D., Chang, E.C., Lin, Q., Ooi, B.C., 2019. Towards scaling blockchain systems via sharding, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, Association for Computing Machinery. pp. 123–140. doi:10.1145/3299869.3319889.

David Yiling Li, guanfeng zeng, Y.L.L.X., 2018. Trinity White Paper An Off--chain Scaling Solution for Neo. Technical Report.

Decker, C., Efe Gencer, A., Juels, A., Croman, K., Eyal, I., Kosba, A., Miller, A., Saxena, P., Shi, E., Gün Sirer, E., Song, D., Wattenhofer, R., Tech, C., 2016. On Scaling Decentralized Blockchains (A Position Paper) Cryptography View project Lightning Network-A Scalability Layer for Bitcoin View project On Scaling Decentralized Blockchains (A Position Paper) Initiative for CryptoCurrencies and Contracts (IC3) 1. Technical Report. URL: https://www.researchgate.net/publication/292782219.

Demir, A., Akilotu, B.N., Kadiroglu, Z., Sengur, A., 2019. Bitcoin Price Prediction Using Machine Learning Methods. 1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC 2019 - Proceedings , 2018–2021doi:10.1109/UBMYK48245.2019.8965445.

Desai, H.B., Ozdayi, M.S., Kantarcioglu, M., 2021. Blockfla: Accountable federated learning via hybrid blockchain architecture, in: Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy, Association for Computing Machinery, New York, NY, USA. p. 101–112. URL: https://doi.org/10.1145/3422337.3447837, doi:10.1145/3422337.3447837.

Dey, S., 2019. Securing Majority-Attack in Blockchain Using Machine Learning and Algorithmic Game Theory: A Proof of Work. 2018 10th Computer Science and Electronic Engineering Conference, CEEC 2018 - Proceedings , 7–10doi:10.1109/CEEC.2018.8674185, arXiv:1806.05477.

Dong, X.L., Rekatsinas, T., 2019. Data integration and machine learning: A natural synergy. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , 3193–3194doi:10.1145/3292500.3332296.

Dorri, A., Kanhere, S.S., Jurdak, R., 2016. Blockchain in internet of things: Challenges and Solutions. 2019 International Conference on Electronics, Information, and Communication (ICEIC) , 1–2URL: http://arxiv.org/abs/1608.05187.

Dorri, A., Kanhere, S.S., Jurdak, R., 2017. Towards an optimized blockchain for iot, in: Proceedings of the Second International Conference on Internet-of-Things Design and Implementation, Association for Computing Machinery, New

York, NY, USA. p. 173–178. URL: https://doi.org/10.1145/3054977.3055003, doi:10.1145/3054977.3055003.

Driscoll, K., Hall, B., Sivencrona, H., Zumsteg, P., 2003. Byzantine Fault Tolerance, from Theory to Reality, pp. 235–248. doi:10.1007/978-3-540-39878-3_19.

van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Machine Learning 109, 373–440. URL: https://doi.org/10.1007/s10994-019-05855-6, doi:10.1007/s10994-019-05855-6.

ETH Zurich, A., Karame, G.O., Wüst ETH Zurich, K., Zurich, E., Ritzdorf ETH Zurich, H., 2016. On the Security and Performance of Proof of Work Blockchains Vasileios Glykantzis Srdjaň Capkun URL: https://bitcoin.org/en/developer-reference{#}data-messages.

Eyal, I., Sirer, E., 2013. Majority is not enough: Bitcoin mining is vulnerable. doi:10.1007/978-3-662-45472-5_28.

Fanning, K., Centers, D.P., 2016. Blockchain and Its Coming Impact on Financial Services. Journal of Corporate Accounting & Finance 27, 53–57. URL: https://onlinelibrary.wiley.com/doi/epdf/10.1002/jcaf.22179, doi:10.1002/jcaf.22179.

F.Y, O., J.E.T, A., O, A., J. O, H., O, O., J, A., 2017. Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology 48, 128–138. doi:10.14445/22312803/ijctt-v48p126.

Gao, W., Hatcher, W.G., Yu, W., 2018. A Survey of Blockchain: Techniques, Applications, and Challenges, in: 2018 27th International Conference on Computer Communication and Networks (ICCCN), pp. 1–11. doi:10.1109/ICCCN.2018.8487348.

Ghasemi, F., Mehridehnavi, A., Pérez-Garrido, A., Pérez-Sánchez, H., 2018. Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. Drug Discovery Today 23, 1784–1790. URL: https://doi.org/10.1016/j.drudis.2018.06.016, doi:10.1016/j.drudis.2018.06.016.

Ghimire, S., 2018. Fig. 7. proof of work flowchart. URL: https://www.researchgate.net/figure/Proof-of-Work-Flowchart_fig6_331040157.

Ghimire, S., Selvaraj, H., 2019. A survey on bitcoin cryptocurrency and its mining. 26th International Conference on Systems Engineering, ICSEng 2018 - Proceedings doi:10.1109/ICSENG.2018.8638208.

Gholamy, A., Kreinovich, V., Kosheleva, O., 2018. A pedagogical explanation a pedagogical explanation part of the computer sciences commons. URL: https://scholarworks.utep.edu/cs_techrephttps://scholarworks.utep.edu/cs_techrep/1209.

Harlev, M.A., Yin, H., Langenheldt, K.C., Mukkamala, R., Vatrapu, R., 2018. Breaking bad: De-anonymising entity types on the bitcoin blockchain using supervised machine learning, in: HICSS.

Hasan, M., Islam, M.M., Zarif, M.I.I., Hashem, M., 2019. Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. Internet of Things 7, 100059. URL: https://doi.org/10.1016/j.iot.2019.100059, doi:10.1016/j.iot.2019.100059.

Hua, G., Zhu, L., Wu, J., Shen, C., Zhou, L., Lin, Q., 2020. Blockchain-based federated learning for intelligent control in heavy haul railway. IEEE Access 8, 176830–176839. doi:10.1109/ACCESS.2020.3021253.

Jalalzai, M.M., Busch, C., Richard, G., 2019. Proteus: A Scalable BFT Consesus Protocol for Blockchains URL: http://arxiv.org/abs/1903.04134.

Jang, H., Lee, J., 2018. An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. IEEE Access 6, 5427–5437. doi:10.1109/ACCESS.2017.2779181.

Janowski, T., 2015. Digital government evolution: From transformation to contextualization. Government Information Quarterly 32, 221–236. URL: https://www.sciencedirect.com/science/article/pii/S0740624X15000775, doi:https://doi.org/10.1016/j.giq.2015.07.001.

J.D. Bruce, 2017. The Mini-Blockchain Scheme. cryptonite Rev 3. URL: http://cryptonite.info/files/mbc-scheme-rev3.pdf.

Jourdan, M., Blandin, S., Wynter, L., Deshpande, P., 2019. A probabilistic model of the bitcoin blockchain, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2784–2792. doi:10.1109/CVPRW.2019.00337.

Kamble, S.S., Gunasekaran, A., Kumar, V., Belhadi, A., Foropon, C., 2021. A machine learning based approach for predicting blockchain adoption in supply chain. Technological Forecasting and Social Change 163, 120465. URL: https://www.sciencedirect.com/science/article/pii/S0040162520312919, doi:https://doi.org/10.1016/j.techfore.2020.120465.

Karame, G.O., 2016. On the security and scalability of Bitcoin's blockchain, in: Proceedings of the ACM Conference on Computer and Communications Security, Association for Computing Machinery. pp. 1861–1862. doi:10.1145/2976749.2976756.

Kasireddy, P., 2017. How does ethereum work, anyway? URL: https://www.preethikasireddy.com/post/how-does-ethereum-work-anyway.

Kathole, A.B., Chaudhari, D.N., 2019. Pros & Cons of Machine learning and Security Methods 21, 6–14.

Khalil, R., Gervais, A., . Revive: Rebalancing Off-Blockchain Payment Networks. Technical Report.

Khan, N., Mi˘, J., 2018. Blockchain Tradeoffs and Challenges for Current and Emerging Applications: Generalization, Fragmentation, Sidechains, and Scalability. 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) , 1582–1587doi:10.1109/Cybermatics.

Khoong, W.H., 2021. When do support vector machines fail? URL: https://towardsdatascience.com/when-do-support-vector-machines-fail-3f23295ebef2.

Kim, H., Kim, S.H., Hwang, J.Y., Seo, C., 2019. Efficient privacy-preserving machine learning for blockchain network. IEEE Access 7, 136481–136495. doi:10.1109/ACCESS.2019.2940052.

Kim, S., Kwon, Y., Cho, S., 2018. A Survey of Scalability Solutions on Blockchain, in: 2018 International Conference on Information and Communication Technology Convergence (ICTC), pp. 1204–1207. doi:10.1109/ICTC.2018.8539529.

Kim, Y.B., Kim, J., Kim, W., Im, J., Kim, T., Kang, S., Kim, C.H., 2016. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. PLOS ONE 11, e0161197. doi:10.1371/journal.pone.0161197.

Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E., 2006. Machine learning: A review of classification and combining techniques. Artificial Intelligence Review 26, 159–190. doi:10.1007/s10462-007-9052-3.

Krawiec, R.J., Housman, D., White, M., Filipova, M., Quarre, F., Barr, D., Nesbitt, A., Fedosova, K., Killmeyer, J., Israel, A., Tsai, L., 2016. Blockchain: opportunities for health care. ComputerWeekly.com , 14URL: https://www2.deloitte.com/us/en/pages/public-sector/articles/blockchain-opportunities-for-health-care.html.

Kumar, R., Khan, A.A., Zhang, S., Kumar, J., Yang, T., Golalirz, N.A., Zakria, Ali, I., Shafiq, S., Wang, W., 2020. Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging.

Kwon, Y., Kim, H., Shin, J., Kim, Y., 2019. Bitcoin vs. Bitcoin cash: Coexistence or downfall of bitcoin cash? Proceedings - IEEE Symposium on Security and Privacy 2019-May, 935–951. doi:10.1109/SP.2019.00075.

Lacity, M., 2018. Addressing key challenges to making enterprise blockchain applications a reality. MIS Q. Executive 17.

Lamport, L., Shostak, R., Pease, M., 1982. The Byzantine Generals Problem - Lamport, Shostak, Pease. Technical Report 3.

Li, K., Li, H., Hou, H., Li, K., Chen, Y., 2017. Proof of Vote: A High-Performance Consensus Protocol Based on Vote Mechanism & Consortium Blockchain, in: 2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International

Conference on Data Science and Systems (HPCC/SmartCity/DSS), pp. 466–473. doi:10.1109/HPCC-SmartCity-DSS.2017.61.

Lin, F., Qiang, M., 2019. The Challenges of Existence, Status, and Value for Improving Blockchain. IEEE Access 7, 7747–7758. doi:10.1109/ACCESS.2018.2888697.

Lin, I., Liao, T.C., 2017. A survey of blockchain security issues and challenges. Int. J. Netw. Secur. 19, 653–659.

Lu, Y., 2019. The blockchain: State-of-the-art and research challenges. Journal of Industrial Information Integration 15, 80–90. doi:10.1016/j.jii.2019.04.002.

Luu, L., Narayanan, V., Zheng, C., Baweja, K., Gilbert, S., Saxena, P., 2016. A secure sharding protocol for open blockchains. Proceedings of the ACM Conference on Computer and Communications Security 24-28-October-2016, 17–30. doi:10.1145/2976749.2978389.

Madan, I., 2014. Automated bitcoin trading via machine learning algorithms.

Mao, D., Wang, F., Hao, Z., Li, H., 2018. Credit evaluation system based on blockchain for multiple stakeholders in the food supply chain. International Journal of Environmental Research and Public Health 15, 1627. doi:10.3390/ijerph15081627.

Mazieres, D., 2015. The stellar consensus protocol: A federated model for internet-level consensus. Stellar Development Foundation , 1–45URL: https://www.stellar.org/papers/stellar-consensus-protocol.pdf, doi:10.1021/ja982417z.

McNally, S., Roche, J., Caton, S., 2018. Predicting the price of bitcoin using machine learning, in: 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), pp. 339–343. doi:10.1109/PDP2018.2018.00060.

Mechkaroska, D., Dimitrova, V., Popovska-Mitrovikj, A., 2018a. Analysis of the Possibilities for Improvement of BlockChain Technology, in: 2018 26th Telecommunications Forum (TELFOR), pp. 1–4. doi:10.1109/TELFOR.2018.8612034.

Mechkaroska, D., Dimitrova, V., Popovska-Mitrovikj, A., 2018b. Analysis of the Possibilities for Improvement of BlockChain Technology, in: 2018 26th Telecommunications Forum (TELFOR), pp. 1–4. doi:10.1109/TELFOR.2018.8612034.

Mendling, J., Weber, I., Aalst, W.M.P., vom Brocke, J., Cabanillas, C., Daniel, F., Debois, S., Di Ciccio, C., Dumas, M., Dustdar, S., Gal, A., García-Bañuelos, L., Governatori, G., Hull, R., La Rosa, M., Leopold, H., Leymann, F., Recker, J., Reichert, M., Zhu, L., 2018. Blockchains for Business Process Management - Challenges and Opportunities. ACM Transactions on Management Information Systems . In press.

Mingxiao, D., Xiaofeng, M., Zhe, Z., Xiangwei, W., Qijun, C., 2017. A review on consensus algorithm of blockchain. 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) , 2567–2572.

Mitchell, T.M., 1997. Machine Learning. 1 ed., McGraw-Hill, Inc., USA.

Mitra, R., 2019. What is facebook libra cryptocurrency? [the most comprehensive guide]- part 2. URL: https://blockgeeks.com/guides/what-is-facebook-libra-cryptocurrency-the-most-comprehensive-guide-part-2/.

Monrat, A.A., Schelén, O., Andersson, K., 2019. A survey of blockchain from the perspectives of applications, challenges, and opportunities. IEEE Access 7, 117134–117151.

Montague, P., 1999. Reinforcement Learning: An Introduction, by Sutton, R.S. and Barto, A.G. Trends in Cognitive Sciences 3, 360. doi:10.1016/s1364-6613(99)01331-5.

Mugunthan, V., Rahman, R., Kagal, L., 2020. Blockflow: An accountable and privacy-preserving solution for federated learning. ArXiv abs/2007.03856.

Nakamoto, S., 2008. Bitcoin: A Peer-to-Peer Electronic Cash System , 112URL: https://bitcoin.org/bitcoin.pdf.

O'Dwyert, K.J., Malone, D., 2014. Bitcoin mining and its energy footprint. IET Conference Publications 2014, 280–285. doi:10.1049/cp.2014.0699.

Ouyang, L., Yuan, Y., Wang, F.Y., 2020. Learning markets: An ai collaboration framework based on blockchain and smart contracts. IEEE Internet of Things Journal , 1–1doi:10.1109/JIOT.2020.3032706.

Palai, A., Vora, M., Shah, A., 2018. Empowering Light Nodes in Blockchains with Block Summarization, in: 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS), pp. 1–5. doi:10.1109/NTMS.2018.8328735.

Poon, J., Buterin, V., 2017. Plasma: Scalable Autonomous Smart Contracts. Technical Report. URL: https://plasma.io/.

Poon, J., Dryja, T., 2015. The Bitcoin Lightning Network. Lightning.Network i, 1–22. URL: http://lightning.network/.

Puthal, D., Malik, N., Mohanty, S.P., Kougianos, E., Das, G., 2018. Everything You Wanted to Know about the Blockchain: Its Promise, Components, Processes, and Problems. IEEE Consumer Electronics Magazine 7, 6–14. doi:10.1109/MCE.2018.2816299.

Rahman, M.A., Hossain, M.S., Islam, M.S., Alrajeh, N.A., Muhammad, G., 2020. Secure and provenance enhanced internet of health things framework: A blockchain managed federated learning approach. IEEE Access 8, 205071–205087. doi:10.1109/ACCESS.2020.3037474.

Raju, S., Rajesh, V., Deogun, J.S., 2017. The case for a data bank: An institution to govern healthcare and education, in: Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance, Association for Computing Machinery, New York, NY, USA. p. 538–539. URL: https://doi.org/10.1145/3047273.3047275, doi:10.1145/3047273.3047275.

Ramachandran, G.S., Wright, K.L., Krishnamachari, B., 2018. Trinity: A Distributed Publish/Subscribe Broker with Blockchain-based Immutability URL: http://arxiv.org/abs/1807.03110.

Ray, S., 2019. A Quick Review of Machine Learning Algorithms. Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Par-

allel Computing: Trends, Prespectives and Prospects, COMITCon 2019 , 35–39doi:10.1109/COMITCon.2019.8862451.

ur Rehman, M.H., Salah, K., Damiani, E., Svetinovic, D., 2020. Towards blockchain-based reputation-aware federated learning, in: IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 183–188. doi:10.1109/INFOCOMWKSHPS50562.2020.9163027.

Rosic, A., 2020. Decentralized scalability -a quick comparison of smart contract platforms: Lightning network, raiden, plasma and rif lumino payments. URL: https://blockgeeks.com/guides/decentralized-scalability/.

Salman, T., Zolanvari, M., Erbad, A., Jain, R., Samaka, M., 2019. Security services using blockchains: A state of the art survey. IEEE Communications Surveys and Tutorials 21, 858–880. doi:10.1109/COMST.2018.2863956.

Sankar, L.S., Sindhu, M., Sethumadhavan, M., 2017. Survey of consensus protocols on blockchain applications, in: 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 1–5. doi:10.1109/ICACCS.2017.8014672.

Saravanan, R., Sujatha, P., 2018. A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification, in: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 945–949. doi:10.1109/ICCONS.2018.8663155.

Sathya, R., Abraham, A., 2013. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. International Journal of Advanced Research in Artificial Intelligence 2. doi:10.14569/ijarai.2013.020206.

Scherer, M., 2017. Performance and Scalability of Blockchain Networks and Smart Contracts. Technical Report.

Scicchitano, F., Liguori, A., Guarascio, M., Ritacco, E., Manco, G., 2020. A deep learning approach for detecting security attacks on blockchain, in: ITASEC.

Sharma, A., 2018. The beginners guide for ethereum mining and casper update. URL: https://medium.com/hackernoon/the-beginners-guide-for-ethereum-mining-and-casper-update-45b9ca938698.

Sharma, S., Agrawal, J., Agarwal, S., Sharma, S., 2013. Machine learning techniques for data mining: A survey. 2013 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2013 doi:10.1109/ICCIC.2013.6724149.

sheinix, 2020. The bitcoin blockchain. URL: https://medium.com/coinmonks/the-bitcoin-blockchain-a3eb996f7140.

Shrivastava, V., Kumar, S., 2019. Utilizing Block Chain Technology in Various Application Areas of Machine Learning. Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019 , 167–171doi:10.1109/COMITCon.2019.8862203.

Singh, A., Thakur, N., Sharma, A., 2016. A review of supervised machine learning algorithms. Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016 , 1310–1315.

statista, 2018. Global blockchain market size 2017-2027. URL: https://www.statista.com/statistics/1015362/worldwide-blockchain-technology-market-size/.

Stornetta, W.S., Haber, S., 1991. How to Time-Stamp a Digital Document. Journal of Cryptology 3, 99–111. URL: https://www.anf.es/pdf/Haber{_}Stornetta.pdf.

Sullivan, C., Burger, E., 2017. E-residency and blockchain. Comput. Law Secur. Rev. 33, 470–481.

Sun Yin, H., Vatrapu, R., 2017. A first estimation of the proportion of cybercriminal entities in the bitcoin ecosystem using supervised machine learning, in: 2017 IEEE International Conference on Big Data (Big Data), pp. 3690–3699. doi:10.1109/BigData.2017.8258365.

Swan, M., 2015. Blockchain: blueprint for a new economy. OReilly.

Tasatanattakool, P., Techapanupreeda, C., 2018. Blockchain: Challenges and applications. International Conference on Information Networking 2018-Janua, 473–475. doi:10.1109/ICOIN.2018.8343163.

Thin, W.Y.M.M., Dong, N., Bai, G., Dong, J.S., 2018. Formal Analysis of a Proof-of-Stake Blockchain, in: 2018 23rd International Conference on Engineering of Complex Computer Systems (ICECCS), pp. 197–200. doi:10.1109/ICECCS2018.2018.00031.

Tonelli, R., Lunesu, M.I., Pinna, A., Taibi, D., Marchesi, M., 2019. Implementing a Microservices System with Blockchain Smart Contracts, in: 2019 IEEE International Workshop on BloNakamotockchain Oriented Software Engineering (IWBOSE), pp. 22–31. doi:10.1109/IWBOSE.2019.8666520.

Toyoda, K., Zhang, A.N., 2019. Mechanism design for an incentive-aware blockchain-enabled federated learning platform, in: 2019 IEEE International Conference on Big Data (Big Data), pp. 395–403. doi:10.1109/BigData47090.2019.9006344.

Vasin, P., 2014. BlackCoin's Proof-of-Stake Protocol v2 Pavel. Self-published , 2URL: https://blackcoin.co/blackcoin-pos-protocol-v2-whitepaper.pdf.

Vats, V., Zhang, L., Chatterjee, S., Ahmed, S., Enziama, E., Tepe, K., 2018. A comparative analysis of unsupervised machine techniques for liver disease prediction, in: 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 486–489. doi:10.1109/ISSPIT.2018.8642735.

Vigna, P., 2021. Crypto and its many fees: What to know about the hidden costs of digital currency. Wall Street Journal URL: https://www.wsj.com/articles/crypto-and-its-many-fees-what-to-know-about-the-hidden-costs-of-digital-currer~:text=On%20the%20Bitcoin%20network%2C%20the.

Wang, T., Liew, S.C., Zhang, S., 2021. When blockchain meets ai: Optimal mining strategy achieved by machine learning. arXiv:1911.12942.

Winter, G., 2019. Machine learning in healthcare. British Journal of Health Care Management 25, 100–101. doi:10.12968/bjhc.2019.25.2.100.

Wood, G., 2017. Ethereum: A Secure Decentralised Generalised Transaction Ledger. EIP-150 REVISION. 2017 , 33URL: https://ethereum.github.io/yellowpaper/paper.pdf.

Xin, W., Zhang, T., Hu, C., Tang, C., Liu, C., Chen, Z., 2017. On Scaling and Accelerating Decentralized Private Blockchains, in: 2017 ieee 3rd international conference on big data security on cloud (bigdatasecurity), ieee international conference on high performance and smart computing (hpsc), and ieee international conference on intelligent data and security (ids), pp. 267–271. doi:10.1109/BigDataSecurity.2017.25.

Xu, X., Pautasso, C., Zhu, L., Gramoli, V., Ponomarev, A., Tran, A.B., Chen, S., 2016. The blockchain as a software connector, in: 2016 13th Working IEEE/IFIP Conference on Software Architecture (WICSA), pp. 182–191. doi:10.1109/WICSA.2016.21.

Yaga, D., Mell, P., Roby, N., Scarfone, K., 2018. Blockchain Technology Overview - National Institute of Standards and Technology Internal Report 8202. NISTIR 8202 , 1–57URL: https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8202.pdf, doi:10.6028/NIST.IR.8202.

Yang, F., Zhou, W., Wu, Q., Long, R., Xiong, N.N., Zhou, M., 2019. Delegated Proof of Stake With Downgrade: A Secure and Efficient Blockchain Consensus Algorithm With Downgrade Mechanism. IEEE Access 7, 118541–118555. doi:10.1109/access.2019.2935149.

Yli-Huumo, J., Ko, D., Choi, S., Park, S., Smolander, K., 2016. Where Is Current Research on Blockchain Technology?A Systematic Review. PLOS ONE 11, e0163477. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0163477, doi:10.1371/journal.pone.0163477.

Yong, B., Shen, J., Liu, X., Li, F., Chen, H., Zhou, Q., 2020. An intelligent blockchain-based system for safe vaccine supply and supervision. Int. J. Inf. Manag. 52. URL: https://doi.org/10.1016/j.ijinfomgt.2019.10.009, doi:10.1016/j.ijinfomgt.2019.10.009.

Yu, Y., Liang, R., Xu, J., 2018. A Scalable and Extensible Blockchain Architecture,

in: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 161–163. doi:10.1109/ICDMW.2018.00030.

Zheng, Z., Xie, S., Dai, H., Chen, X., Wang, H., 2017a. An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends. Proceedings - 2017 IEEE 6th International Congress on Big Data, BigData Congress 2017 , 557–564doi:10.1109/BigDataCongress.2017.85.

Zheng, Z., Xie, S., Dai, H., Chen, X., Wang, H., 2017b. An overview of blockchain technology: Architecture, consensus, and future trends, in: 2017 IEEE International Congress on Big Data (BigData Congress), pp. 557–564. doi:10.1109/BigDataCongress.2017.85.

Zheng, Z., Xie, S., Dai, H.N., Chen, X., Wang, H., 2017c. Blockchain Challenges and Opportunities: A Survey. International Journal of Web and Grid Services .

Zhou, S., Huang, H., Chen, W., Zhou, P., Zheng, Z., Guo, S., 2020. Pirate: A blockchain-based secure framework of distributed machine learning in 5g networks. IEEE Network 34, 84–91. doi:10.1109/MNET.001.1900658.

Appendix A

Appendix

A detailed understanding of the problems around blockchain technology requires a solid understanding of the concept and mechanism that gave life to the technology. Therefore, this section is going to unpack the technical complexity of the technology.

A.1 Blockchain Architecture

The structure of blockchain technology is represented by sequel blocks that enclosed a list of transactions which is stored as a flat-file. It uses two important data structures: Pointers that keeps the information about the location of each block. Linked lists used in keeping a sequential order of blocks with help of the pointer. The distributed structure of blockchain technology enforces the need for each participant within the network maintains, approves, and updates all record in the blockchain. This network consists of many computers that only agree to alter the record through a consensus of the whole network ensuring all records and procedures are in order. Thus, it provides an excellent data validity, security and trust in the data integrity.

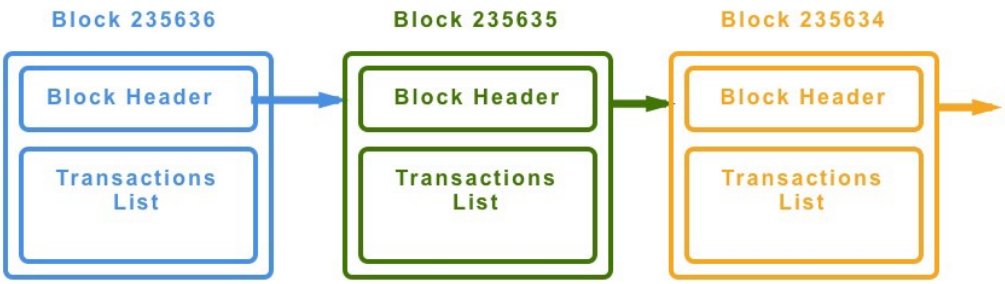

Figure A.1: An example of blockchain with a continues growth
Source (sheinix, 2020)

Figure 1 above illustrates an example of a blockchain, the link between the blocks is done by an attribute in the block header called parent hash, it holds the hash values of the previous block. Each block must have the parent hash except for the genesis block that doesn't have any parent block. Each block consisting of two sections: the block header and the block body. The block header contains information used in identifying a particular block and every blockchain has a different set of attributes included in the block header base on their operational requirement. The block body is composed of a transaction counter and some set of transactions (the maximum number of transactions depends on the size of the block).

A.1.1 Block structure

As mentioned earlier, the attribute in the header are base on the blockchain and its requirements. Figure 1 shows the block header in bitcoin context while figure 2 does the same in Ethereum context:

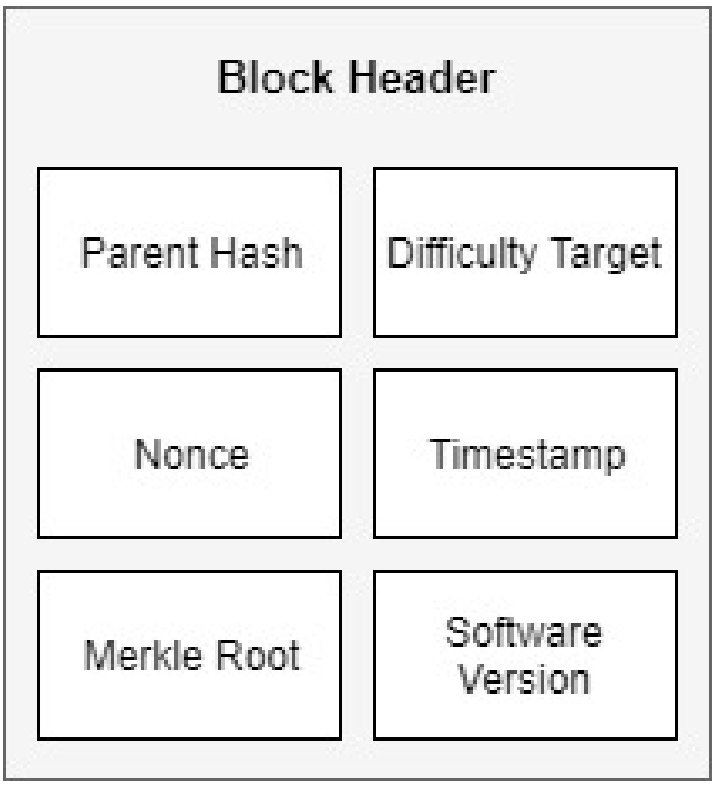

Figure A.2: Bitcoin block header

Bitcoin block header attributes

- *Block version:* indicates which set of block validation rules to follow or the bitcoin version number.
- *Parent hash:* the hash value of the previous block.
- *Merkle tree root hash:* the hash value of all the transactions in the block.
- *Timestamp:* current timestamp in Unix's time() at this block's inception (seconds since 1st January 1970).
- *Difficulty target:* The difficulty target of the block.
- *Nonce:* a number find by miners used in generating the correct hash

Ethereum block header attributes

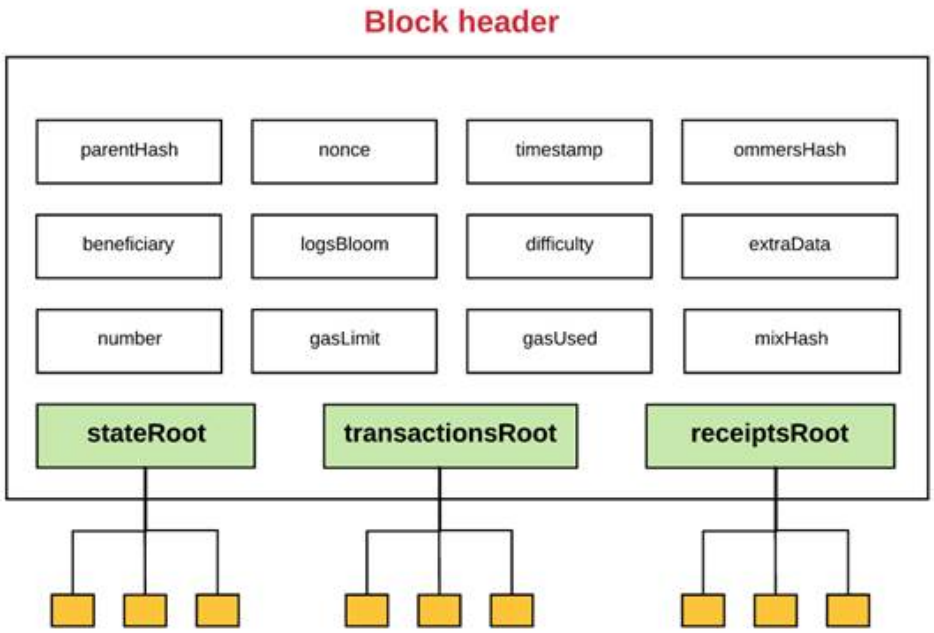

Figure A.3: Ethereum block header

Source (Kasireddy, 2017)

Extra attributes found in Ethereum block header:

- *ommersHash:* The Keccak 256-bit hash of the current block list of ommers.
- *beneficiary:* The 160-bit address of the miner to collected the successful fees of mining this block.
- *logsBloom:* The log information in Bloom filter (data structure).
- *extraData:* An arbitrary byte array containing data relevant to this block. This must be 32 bytes or less than.
- *number:* A count of the number of the ancestor blocks. (The genesis block has a number of zero).
- *gasLimit:* The current limit of gas expenditure per block.
- *gasUsed:* The total gas used by transactions in this block.

- *mixHash:* A 256-bit hash which, combined with the nonce, proves that a sufficient amount of computation has been carried out on this block.

- *stateRoot:* The Keccak 256-bit hash of the root node of the state trie, makes it easy for a light client to verify anything about the state.

- *transactionsRoot:* The Keccak 256-bit hash of the root node of the trie that contains each transaction in the transactions list portion of the block.

- *receiptsRoot:* The Keccak 256-bit hash of the root node of the trie structure that contains the receipts of each transaction in the transactions list portion of the block.

A.1.2 The core component and concept of blockchain

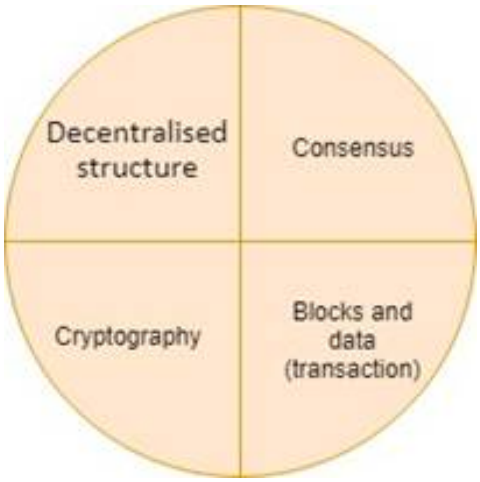

Figure A.4: The main component of blockchain technology

A.1.2.1 Decentralisation

The decentralised structure of blockchain technology made data more secure by removing the reliance on central point storage. It brought about accuracy and trust in all data stored in blockchain technology because it keeps multiple copies of the data across all participating nodes located in many different parts of the world and a complex consensus between all nodes is required before any form of modification on the data. Any successful manipulation will have to occur across all or the majority of the nodes

making it almost impossible. Any attack on one of the nodes cannot lead to losing data or its validity since all information is recorded on multiple nodes that can easily synchronise data based on consensus (Yli-Huumo et al., 2016).

A.1.2.2 Consensus

A consensus can be defined as a set of rules that are used in governing a distributed system. It plays a significant role in blockchain technology's success in eliminating the need for a central monitory by providing a secured channel for all nodes to communicate and reach a collective agreement before any action is performed on the network. The efficiency of the blockchain or decentralised system in both performance and security aspect is highly affected by the efficiency in the consensus algorithm. The protocol has to process and validate transactions before a record can be added into the block (Bach et al., 2018). Therefore it plays a pillar role in the blockchain system as well as any other distributed system. Proof-of-Work (PoW) consensus protocol was the first to be developed and is used by both Bitcoin and Ethereum (Yli-Huumo et al., 2016). PoW will be discussed in the next section while other consensus protocol will be discussed in depth in the state of the art section.

A.1.2.3 Cryptography

Cryptography is one of blockchain's most important features, it plays a key role in achieving immutability in the blockchain system. The techniques bring privacy and confidentiality by making sure only the intended receiver can read a message sent which is necessary when communicating over any untrusted medium. It also helped in proving identity and securely sharing crypto keys. Blockchain technology has used cryptography in so many ways that include generating hash for the transaction and blocks data when storing the history of transactions in a Merkle tree. It is not only used in making data immutable, it is also used in the authentication (Salman et al., 2019).

A.1.2.4 Blocks and transactions

After a transaction is performed, the details are published to the blockchain network through a place called transaction pool where all unconfirmed transactions wait to be verified and validated. A transaction is only confirmed if added in a block that has

been added to the blockchain. They both are very important parts of the technology because they represent the data and its storage that the whole complexity of the system is trying to secure (Yli-Huumo et al., 2016).

A.2 Conclusion

The chapter discussed the important technical aspects of the blockchain technology and the PoW consensus protocol for better understanding of discussions in the chapters that follow. The core components of the technology were discussed. The discussion highlighted the importance of decentralisation and security in the blockchain system which provides a better insight into the concerns around the scalability issue forcing trading off one of the important attributes. The research also discussed the PoW consensus protocol and the idea behind the mining process and mentioned how the difficulty level influences the mining speed. To conclude, The amount of trust that can be given to blockchain depends on its level of decentralisation and the security of its protocol, and the difficulty level influences the mining speed.