International Journal of Advanced Technology and Engineering Exploration, Vol 9(95)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2021.876322

Review Article

A comprehensive review of significant learning for anomalous transaction detection using a machine learning method in a decentralized blockchain network

Sabri Hisham*, Mokhairi Makhtar and Azwa Abdul Aziz
Faculty of Informatic and Computing, Universiti Sultan Zainal Abidin,22000 Besut,Terengganu, Malaysia

Received: 04-July-2022; Revised: 20-October-2022; Accepted: 21-October-2022

Abstract
Blockchain is a distributed ledger technology (DLT) that enables decentralized applications (DApps) such as Hyperledger, Bitcoin, Ethereum, decentralized finance (DeFi), and non-fungible token (NFT) to operate on it. Bitcoin is a popular application that leverages blockchain technology (BT) to provide secure transactions in a decentralized environment. Among the main reasons for BT being utilized is to eliminate the dependence of the process on third parties such as lawyers, insurance firms, and banks to execute approvals, verifications, and signatures. Due to the immutability and transparency of this technology, it has been deployed in several industries, including agri-food, supply chain, automotive, pharma, healthcare, insurance, land registration, and higher education, to increase verification, security, and traceability. Modern research indicates that blockchain distributed ledger may still be susceptible to various privacy, security, and dependability challenges, despite their impressive characteristics. To resolve these challenges, it is essential to notice abnormal behavior in a timely manner. Consequently, machine learning (ML) approaches with anomaly detection play a comprehensive role in preventing fraud. This paper explores the technological integration of anomaly detection models into BT and compares supervised and unsupervised ML techniques for detecting rogue and legitimate transactions. The studies for the review come from three well-known databases: Web of Science (WoS), Google Scholar and Scopus. Using sophisticated search tactics, specialized keywords such as Bitcoin, Ethereum, blockchain, fraud detection, anomaly detection, data mining, and ML are employed. These findings are based on three main points (1) the background of BT (2) Blockchain integration with ML, and (3) the ML approach for supervised learning and unsupervised learning. According to the findings of this study, supervised learning is the most prevalent method for studying anomalous detection in blockchain networks. This study also equips researchers with the knowledge necessary to perform research on the detection of blockchain network anomalies using ML techniques.

Keywords
Blockchain, Ethereum, Bitcoin, Machine learning, Anomaly detection, Artificial intelligence, Fraud detection.

1.Introduction
Blockchain technology has changed the global trading of assets. A blockchain can be viewed as a connected ledger managed by a distributed peer-to-peer (P2P) network. Blockchain offers distinctive characteristics such as transactional privacy, the immutability of data, transparency and cryptographic, among others. These features paved the door for blockchain to develop numerous technology solution, including voting applications [1,2], internet of things (IoT) [3,4], and supply chain management (SCM) [5,6], among others. The increasing desire for technological advancements stimulated the development of BT.

A blockchain logs the transfer of assets during transactions and arranges them into blocks. Cryptographic algorithms connect blocks to their predecessors, establishing a blockchain. Satoshi Nakamoto popularized the blockchain in 2008 as the public ledger of Bitcoin transactions [7].

Bitcoin is the first popular blockchain technology (BT) and the first real implementation of a cryptocurrency ecosystem. Ethereum is the principal distributed blockchain network that now supports digital smart contracts and the second-largest cryptocurrency, known as Ether. Despite these advantages, BT is susceptible to certain assaults and problems. Security concerns and weaknesses, such as majority assaults [8], weak smart contracts [9], eclipse attacks [10], and phishing schemes [11], have

*Author for correspondence

been significant obstacles to BT. The Ethereum blockchain has become an important platform despite security issues such as phishing scams, which account for about half of all Ethereum cybercrime [12]. In addition, the distributed ledger made publicly accessible by the Ethereum blockchain network would undoubtedly be classified as big data ecosystem. Furthermore, manually searching through all of these transactions to discover any transactions suspected of exhibiting abnormality characteristics would be impractical or time-consuming. In theory, machine learning approach would help distinguish between transactions that exhibit normal and abnormal behavior among user address by learning the attributes that correspond to either normal or abnormal conduct. Blockchains are susceptible to numerous types of harmful assaults and fraudulent behavior, which may be detectable by analyzing transaction patterns. Consequently, BT can profit from using machine learning (ML) algorithms via their capacity to evaluate, learn, and improve with large amounts of data. Detecting abnormalities has been studied extensively for ages. Numerous independent ML approaches have been created and utilized for anomaly detection in various applications. The process of identifying abnormal patterns in a transaction is relatively difficult to detect [13]. Therefore, anomaly identification is used in various applications. For example, anomalous detection is accomplished by integrating an ML approach with blockchain in a study on intruder detection in drone technology of unmanned aerial vehicle (UAV) [14]. Other extensively used applications for identifying cybercrime anomalies include crypto-jacking [15], phishing [16], frauds [17], money laundering [18], crypto-ransomware [19], and Ponzi schemes [20]. Another example is the detection of abnormalities in blockchain-IoT-based application for the construction of a secure framework [21], decentralized IoT data in smart cities [21], and monitoring wastewater reuse and electricity usage in smart grids [22].

Overall, it is crucial to note that anomaly detection is one of the critical topics for securing future blockchain networks and that a large amount of work is being conducted on this subject from multiple viewpoints, which will be discussed in this study. A detailed evaluation of these studies includes (1) a basic understanding of blockchain architecture, (2) a review of blockchain and ML integration, and (3) identifying the principal anomaly detection methods using ML approaches (such as supervised, unsupervised learning), datasets, metric measurement, tools, and the types of data they exploit.

This review is structured as follows: The second part explains how the article selection process is carried out. The third section provides an overview of blockchains such as smart contracts, versions, consensus, properties, and classifications. Section 4 and 5 gives an overview of the combination of blockchain-based anomaly detection techniques. In section 6, the study's findings are discussed and section 7 shows the conclusions and directions of future studies.

2.Method

2.1Articles selection strategy

The selection of articles related to anomaly detection in the blockchain network involves a systematic article selection process (see *Figure 1*). Thus, three online databases, namely Scopus, Google Scholar, and web of science, are used for searching the articles. This process is carried out based on the steps outlined in the Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) for the selection of documents (identification, screening, and eligibility).

In general, the article search technique begins with an extraction of relevant articles from major journal databases (WoS, Scopus, and Google Scholar). This search encompasses a variety of primary subject categories, including computer science, security, information systems, artificial intelligence (AI), engineering, decision science, and mathematics. However, this database has millions of articles published in journals from over the world. Since the primary study focuses on the detection of abnormalities in the block chain network, the search for papers is conducted using a search string related to the study's title. "Anomaly detection," "blockchain," "Bitcoin," "Ethereum," "ML," "abnormal," "suspicious," "malicious," and "fraud" are among the most popular search terms. The search string used to locate items in the database yielded 919 results. This complete study is comprised of excerpts from papers published between 2017 and 2022.

The extracted articles will then be recognised and analysed, while the irrelevant articles removed. Initially, the article search was limited to journal articles and did not include the categories of books, trade journals, or conference proceedings. As far as the language of the article is concerned, the selection includes only English-language articles. Therefore,

the article version written in a language other than English is excluded. After undergoing this filtering procedure, 476 articles were retained while 443 were excluded.

Finally, the paper is finalised through the eligibility procedure. This procedure involves a final screening that eliminates duplicate items. Out of a total of 476 items, only 133 remained after this process. Therefore, based on the study, 343 articles were excluded since they did not emphasise blockchain, machine learning, or detecting whether something is incorrect.

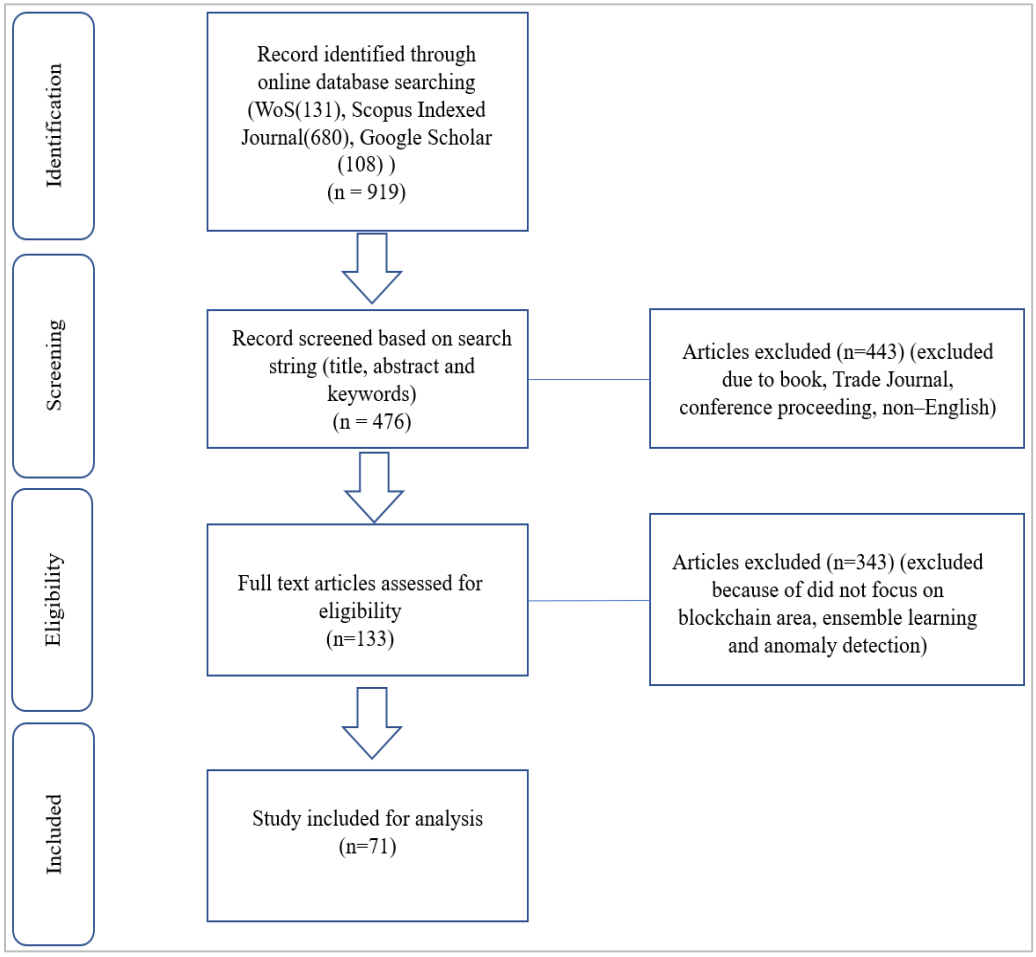

Figure 1 The study's data flow diagram

2.2 Workflow review analysis

This study is described in depth in accordance with the analysis of the work flow that refers to the articles chosen according to the subtheme (see *Figure 2*). First, a comprehension of blockchain technology will be offered. It examines blockchain architecture, the contrast between blockchain and conventional databases, blockchain versions, consensus processes, blockchain classification, blockchain characteristics, blockchain platforms, and smart contracts. The integration of BT and ML was subsequently analysed. This section describes how both of these technologies can be advantageous to both parties. In the section on anomaly detection analysis, it describes the use of ML models to detect anomalous blockchain transactions. This analysis includes a review of earlier studies as well as the use of supervised learning and unsupervised learning techniques for anomaly identification. In spite of this, ML model construction (training, testing) necessitates the usage of suitable data sets, as these determine the performance of the final model created. Therefore, the datasets and tools section describe the analysis of prior studies, covering the sorts of datasets (live datasets, social media, open datasets, and tool datasets) and the application of tools

(application programming interface (API), development tools, software, and smart devices). Lastly, it involves analysing the sorts of performance indicators utilised to quantify experimental outcomes in past investigations.

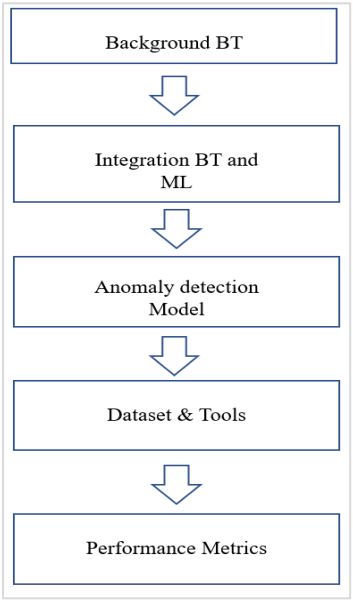

Figure 2 Workflow review analysis diagram

3.Decentralized blockchain network

3.1Blockchain architecture

Blockchain was first discussed after Satoshi Nakamoto introduced a decentralized cryptocurrency. The main function of blockchain is to keep updated news entries and authorization processes for each network participant. All the transactions are maintained in a particular order with the collection of blocks. The blockchain architecture consists of private, public, and consortium [23]. In addition, the architecture of blockchain consists mostly of the six layers depicted in *Figure 3*, comprising hardware, data, the network, the consensus, and the application. The majority of blockchain applications are hosted on a server in an on-premise or cloud data center and operate on a P2P basis. Conceptually, P2P networks operate on a large scale of connected nodes (computers) for data sharing, validation, verification, and storing all the transactions in the ledger [23]. On the P2P basis, all the programs run as a client-server concept, in which the request from client browsers is returned with the result.

The data structure in the blockchain is stored and structured in a list of blocks at the data layers. It comprises of two fundamental components, namely pointers and linked lists. Clearly, linked blocks refer to a series of linked lists containing data pointing to the prior block. Because each blockchain block might include thousands of transaction records, the Merkle (binary tree of hashes) algorithm and Merkle tree root [24] were employed to produce the final hash value. In general, hash values are utilized as input and output identifiers. In the complete block, each transaction output is only valid once used as an input for the entire blockchain [25]. The blockchain operates on consensus features, cryptography, and Merkle trees. As seen in *Figure 4*, each block contains the Merkle tree's root hash, as well as critical information such as difficulty, nonce, data, block number, timestamp and the preceding block's hash. Thus, a Merkle tree provides security, reliability, and incontestability. In addition, the security and integrity of the data stored in the distributed blockchain are assured. To reach this point, each transaction must be digitally signed with a secure key(private key) and validated by a signer.

A P2P network is a data transmission mechanism and the primary network-level communication architecture. It preserves network integrity by allowing nodes to interact and synchronize with one another. In a blockchain setting, these nodes consist of numerous computers that are interconnected via a blockchain network in order to conduct transactions. The consensus layer is one of the crucial layers in blockchain architecture, including Ethereum, Hyperledger, and Bitcoin. Apart from that, this layer is comprised of consensus algorithms mechanism such as proof of authority (PoA), proof of work (PoW), proof of stake (PoS) and practical byzantine fault tolerance (PBFT). These algorithms run the ranking process, block validation, and confirm that all nodes reach consensus. The application layer of a blockchain architecture consists of layers that interact with end-users via the programming logic embedded in a contract (smart contract in Ethereum, chain code in Hyperledger) based on business cases, rules, and algorithms. This application type is known as a DApps. Smart contracts include procedures, models, assets, rules, and business logic that function without needing a third party with blockchain-based trust. Typically, the application layer will communicate with end-users through the execution layer's application interface. These two layers interact through the API, application framework, and smart contracts. Concurrently, the validation process and consensus among nodes are conducted on the blockchain.

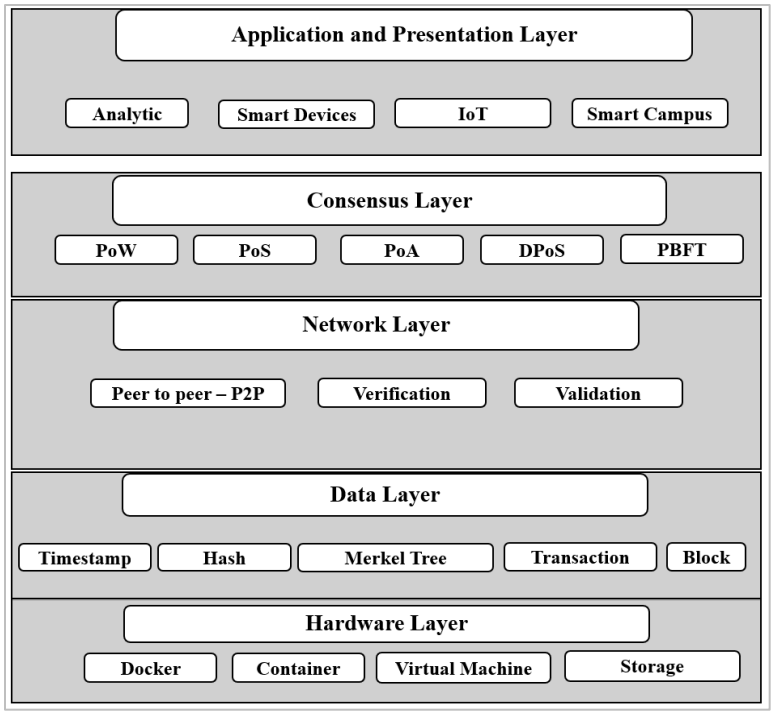

Figure 3 Blockchain architecture

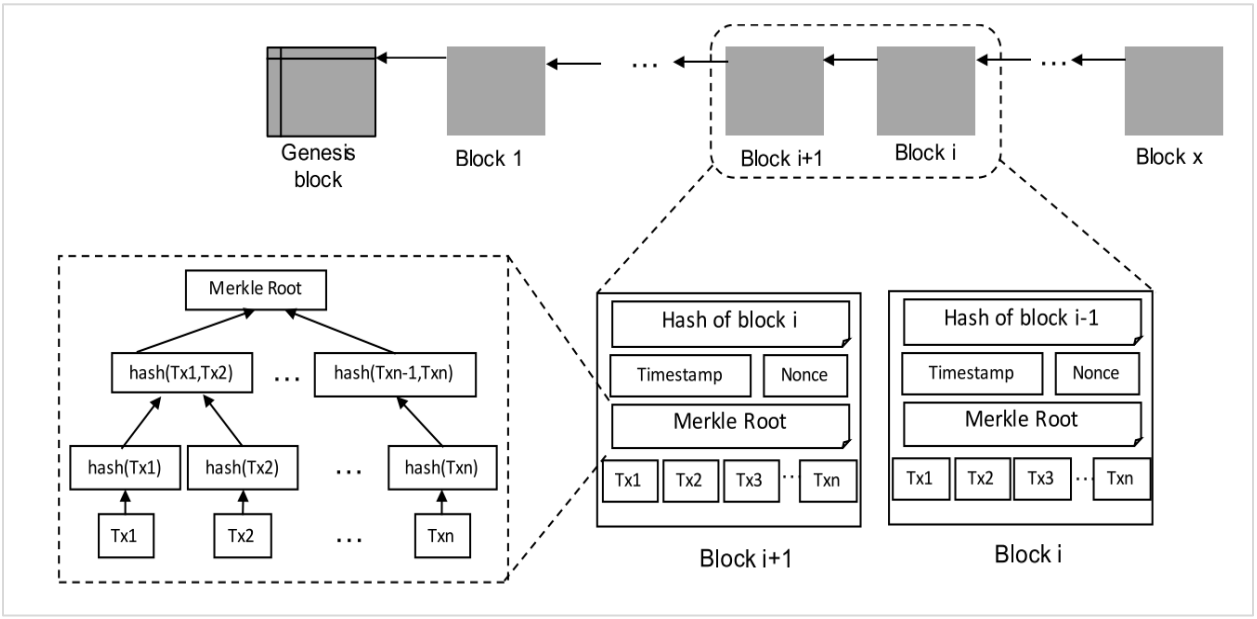

Figure 4 Block structure [14]

3.2 Blockchain and traditional database

Conceptually, blockchain and distributed ledger technology are distributed and decentralized systems that store structured data in public ledgers and blocks. Pertaining to trust, blockchain eliminates using trusted third parties on whom databases rely, improving the content's veracity and reliability [26]. There is a variation in data structure between blockchain and conventional databases. In contrast to databases, blockchains store data in blocks. Over time, databases embraced and utilized a relational model that enabled increasingly complicated data collection methods by linking information from numerous databases. An administrator can modify,

manage, update, and control a database. In addition, administrators can undertake database administration tasks such as speed improvement, tuning, monitoring and database size reduction. Generally, a huge database slows down the performance index. Thus, administrators employ optimization techniques to increase the database's performance. A recursive database allows you to alter or delete a record if you have permission to do so. The contrast between blockchains and databases is summarized in *Table 1*. A blockchain, unlike a database, maintains data in the chain of blocks, including hash code information from previous blocks. This blockchain is extremely safe and hack-proof due to the cryptographic techniques that protect it. The SHA-256 algorithm is the most widely used secure hashing algorithm and function. Nevertheless, SHA-256 is predominantly used for one-way encryption [27].

Table 1 Compares blockchain and databases

Criteria	Database	Blockchain
Architecture	Centralized	Decentralized/Distributed
Performance	Connected Centralized Server: slow	Many blockchain Nodes: faster
Security	Append data to the database without tracing	Traceable Block Transaction
Downtime for system update	When updating the server, the system goes down.	No system downtime. Other blockchain nodes can be updated at any moment.
Access Control	Admin has complete access to the database.	Smart Contract: Permission or access is protected and encrypted. No Admin
System Backup	Must do backup database: Weekly/Monthly	Original Copied: Each blockchain node
DDOS	The entire system hangs and shuts down.	One node down. Other nodes are running
Application Development and Deployment	It at least requires web server hosting services. There is no direct programming in the database. Only SQL programming for data analysis.	All programming and data are in one place via smart contract programming. No need for hosting services. Everything is executed by blockchain nodes.

3.3 Blockchain version

Blockchain Version 1.0 was introduced by Hall Finley in 2005, who implemented DLT, marking its first cryptocurrency-based application. In 2008, Satoshi Nakamoto launched Bitcoin, which leverages BT 1.0 [28].

After the problems of digital currency being spent twice and digital transactions being performed without a trusted third party were resolved, Bitcoin became a popular method of conducting financial transactions on the blockchain network [29]. Thus, any participant can conduct valid Bitcoin transactions with this version, and this type is predominantly used for currencies or payments.

In Version 1.0, Bitcoin mining was inefficient, and the network lacked scalability. Hence, the latest version of blockchain Version 2.0 addresses these issues. In this iteration, smart contracts will be added to the blockchain in improvement to cash management. This smart contract is an executable software that automatically verifies the previously established requirements, such as facilitation, verification, or enforcement, and minimizes transaction cost efficiency. The adoption of Ethereum after Bitcoin has resulted in an increase in transactions on public networks in a short time. These scenarios have changed most domains from focusing on cryptocurrency to DApps. This led to blockchain Version 3.0, which focused on developing DApps in health, industry, digital government, and other areas. *Figure 5* depicts the chronology of blockchain network versions.

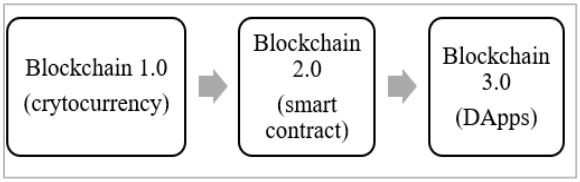

Figure 5 The history of the blockchain

3.4 Blockchain consensus

Authors in [30] explains consensus algorithms that enable a safe updating of the shared ledger. A common method for error detection in the decentralized ledger is distributing the shared data over multiple network replicas. Consequentially, the method utilized in such operations is the consensus

algorithm, and it is often important for processes to agree on the desired data value during calculation. The consensus algorithms consist of various types, such as PoW, PoS, PBFT, PoA and Delegated Proof of Stake (DPoS). *Table 2* compares the PoW and the PoS.

Table 2 PoW and PoS comparison [31]

Property	**PoW**	**PoS**
Efficiency of Energy	No	Yes
Tolerated Power	Less than 25 percent computing power	Less than 51 percent stake
Hardware	Very essential	No Needed
Forking	When two nodes find an appropriate nonce	Very difficult
Attack for double-spending	Yes	Difficult
Speed of Block Creating	Low according to the variant	Fast
Pool Mining	Yes. It is preventable	Extremely difficult to avoid
Example	Bitcoin	Nextcoin

Previous research related to consensus algorithms was conducted by [32]. This work introduced the first variant of the consensus algorithm named PoW. For a related understanding, the PoW consensus can be seen in the mining process. In general, the mining process increases the turnover of transactions in one block compared to another block. Each public node in the network is eligible to be a miner for new block validation before being added to the network. However, the PoW consensus method consumes too much electricity. Nonetheless, the phenomena of the affluent getting richer may manifest in the PoS consensus procedure. The alternate solutions are produced in the network by the initial miner. Each block varies in terms of estimation computation difficulty and quantity. The data will be processed to be stored in the blocks and converted to a unique hash value according to the cryptographic method. Therefore, one-way hashes are hash values that are encrypted on one side and cannot be reversed to the original data. Further, due to its high energy consumption, PoW is environmentally unfavorable. As a result of this consensus adoption of Ethereum, PoW is now commonly employed to incorporate BT in applications.

PoS is an algorithm comparable to PoW. In terms of consensus mechanisms, the PoS has no competition compared to the PoW [33]. If the given validator cannot confirm the transaction, the network will select the next validator and continue doing so until another node can confirm the transaction. Miners in PoS must demonstrate possession of the required amount of cash. It is expected that wealthy individuals would be less inclined to assault the network. It implements the CASPER protocol when in PoS. However, PoS will be implemented on Ethereum in the near future. PoS is more efficient and conserves more energy than PoW. As the mining cost is almost zero, it is possible that attacks will occur. Many blockchains begin with PoW and transition to PoS over time.

PBFT is an algorithm that was developed in the 90s to solve Byzantine tolerance errors. It is widely used in BT and distributed computing. The major BT Hyperledger employs the PBFT consensus [34]. Note that DPoS was introduced based on the evolution of PoS. Basically, this algorithm uses the concept of voting to select a representative (called a witness) for the purpose of validating the next block. These delegates are selected by collecting tokens into betting groups and linking them to specific delegates.

PoA operates through nodes that have been granted authorization only to verify transactions. Therefore, it needs to be approved by a majority of the authorities.

3.5 Blockchain classification

Current BT systems generally fall into three types: Public BT, Private BT, and Consortium BT.In the Public BT, anyone connected to a public network and able to access all records is allowed to participate in the process by consensus. These participants will be rewarded if they have reached an agreement. Consequently, the authors [35] explains that anyone can join a Public BT network and engage without authorization because it is completely open. To encourage additional users to join the network, there is typically an incentive structure in place. The formation and operation of Public BT are secure. Thus, transactions in the blockchain go through an expensive consensus mechanism process even though each member is free to join any node in the network. The hash value is a cryptographic mechanism that makes it impossible and difficult to change the information(data) on the blocks. In addition, each node in a blockchain network can be anonymous,

which is one of the properties that protect the privacy of its members [36]. One of the most popular and widely used Public BT networks is Bitcoin. Since Bitcoin operates publicly, it has become a major drawback because of the high processing power required for the mining process (placing data in a ledger). This process is carried out in PoW, which requires consent at each node by implementing a complex cryptographic mechanism. This concept of openness has an impact on the absence of privacy.

For Private BT or permissioned networks, only nodes from a single company would be permitted to join the consensus mechanism process and offer consumers the ultimate privacy they desire. Thus, each member requires an invitation from the initiator of the network or from someone who has been authorized for access permission to the network [37]. In addition, this is a measure to limit and control access to members who want to join the network node to implement the business process that has been developed. In this case, they need to get an invitation and permission first before accessing the network node.

The Consortium BT is also known as the federated blockchain, which consists of several organizations that are administered on one platform and are managed by one entity. The Consortium BT established by multiple companies is somewhat decentralized, as only a subset of nodes would be chosen to ascertain consensus. This type's consensus process is slower than Private BT but faster than Public BT. There are fewer known players in a Consortium BT. As a voting-based system, it ensures low latency and good performance. Every node can read or write transactions, but none may add a block. To confirm this block, every node (or a supermajority) must do so. If this rule is not met, the block cannot be inserted. As for Consortium BT, it is applicable to numerous business applications. The blockchain framework used to develop business applications using the Consortium BT network is Hyperledger. Furthermore, due to enhanced security measures and the participation of various companies, this version of the blockchain is also more resistant to hacking.

3.6 Blockchain features

BT has existed for a considerable amount of time and continues to be in the public eye. Bitcoin, a prominent cryptocurrency, initially brought the technology to public attention. In addition, additionally, blockchain manages data through distributed ledgers. Any node connected to the network will get a copy of the ledger, and it is very different from the centralized database approach. The failure of a centralized database could result in data loss, and blockchain solves this issue. Transparency of the transactions is another significant benefit. The following properties have been identified for the BT, as indicated in *Figure 6*.

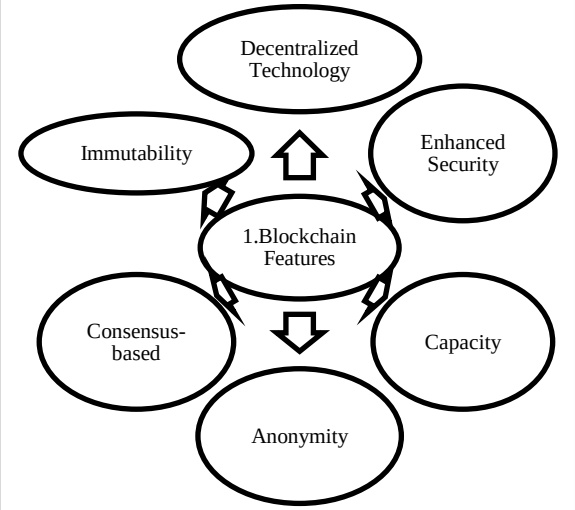

Figure 6 Blockchain feature

Decentralized Technology means the network is decentralized, meaning it is not rights by a central authority or managed by a single individual. A node that connects a network and then renders it decentralized is the crucial feature of the blockchain. As the system does not require any regulatory authority, we can store our assets on it immediately over the web. Thus, it may store anything, including cryptocurrencies, vital papers, contracts, and other digital items of value. With the aid of blockchain, we will have direct control over them using the private key. Thus, the decentralized system restores the common people's authority and property rights. Multiple computers, often known as nodes, house the blockchain ledger. In BT, central authority is not required to verify information on nodes in a P2P manner[38]. Consensus protocols are a set of rules and methods used to validate transactions and maintain information consistency and integrity.

Among the fascinating characteristics of BT is immutability. Also known as un-tamperability [39], immutability refers to something that cannot be changed or altered. The unchanging and permanent nature of the network is an important feature in the BT environment. The agreement needs to be made by

the majority to change the data to make it tamper-proof. Apart from that, the data block stores the timestamp in the hash value to ensure the data is permanent [40]. However, once a transaction is approved into blocks, the data cannot be altered, deleted, or restored to the original[41].

In terms of security enhancement, misappropriation by certain parties can be avoided because the blockchain eliminates the role of the central authority to ensure that there is no modification in the network. In addition, the protection aspect is enhanced through the cryptography mechanism for the encryption process. The method of using private keys (disclosed to the owner) and public keys (disclosed to anyone) is the practice of cryptography to ensure the level of security of transactions is guaranteed. This key is used as access for secure transactions with real ownership and immutability. Each block in the distributed ledger has its own unique hash and includes the hash of the previous block. Changing or attempting to alter the data would necessitate changing all the transaction hash (Tx Hash) which is nearly impossible. Therefore, public keys and private keys are required to access data to perform transactions. Additionally, the decentralization of data structures in BT uses P2P consensus to avoid the occurrence of failures in the data. It differs from a centralized design that is prone to disruption and failure.

The consensus algorithm is what makes BT effective. It is an important component of every blockchain and a defining feature. A simple explanation is that consensus is a mechanism that helps make decisions within network nodes after the agreement process is accepted. The P2P model is practised within the blockchain to generate democratic decision agreements by each node. Technically, the ledger will be copied and distributed to other node blocks by consensus after a new or existing block addition transaction takes place [42].

Anonymity conceals the identity of users and maintains their identities confidential. Therefore, this mechanism allows transaction verification to be carried out without knowing the identity and disclosing personal information. This is the aspect of trust that uses algorithms for data transactions between nodes. The information transfer is done anonymously, and the information on the node does not need to be disclosed.

Lastly, BT is its ability to enhance the capacity of a whole network. Thousands of interconnected computers can be more powerful than a few centralized servers. The smart contract system is another amusing fact. This can expedite the settlement of any type of contract. This is one of the greatest advantages and benefits of BT to date. Digital smart contracts on the blockchain may automatically generate transactions, make decisions, and store data. Using particular consensus protocols [38], all system nodes can automatically exchange and verify data.

3.7 Blockchain platform difference

Emerging blockchain platforms are essentially indistinguishable from core BT at this point. In a blockchain, data structures are stored in blocks that contain timestamp information and previous blocks. In addition, it contains cryptographically signed digital records that are shared through a distributed ledger. Digital assets consist of tangible and intangible objects such as patient records, security, currency, etc. There are three main blockchain platforms: Bitcoin, Ethereum, and Hyperledger. *Table 3* displays a comparison of this blockchain platform.

Table 3 Blockchain platform

Scope	**Bitcoin**[7]	**Ethereum** [43]	**Hyperledger fabric**[44]
Consensus	PoW	PoW, PoS	PBFT
Currency	BTC	Ether	None
Smart contract	Yes	Yes	None
Miner Participation	Public	Public, Private, Hybrid	Private
Operation Mode	Permission less	Permission less	Permissioned
Governance	Bitcoin Developer	Ethereum Developer	Linux Foundation
Application	Cryptocurrency	Yes	Yes
Data Access	Public Network	Public and Private Network	Authorize (Private Network)

Bitcoin is a decentralized digital currency that is transferred between two parties without the need for intermediaries such as regulators,banking,insurance or other financial institutions. With Bitcoin, digital or

virtual currencies known as cryptocurrencies gained prominence. Hyperledger [44] is a multi-project open source collaborative initiative formed by The Linux Foundation in December 2015 to improve various industry BT. Hyperledger Fabric is a modular enterprise architecture platform for creating private and permissioned blockchains. Members of the blockchain network must register with a Membership Service Provider (MSP). In the Hyperledger, several channels can be created, each of which consists of a separate ledger to be accessed by several groups of authorized users. Ethereum is a distributed blockchain network platform that enables a P2P network for securely executing and verifying smart contracts (program code). In Ethereum, a smart contract is a programme code developed to allow parties to transact without middlemen. Thus, all records cannot be modified and securely distributed across decentralized network nodes, allowing participants to gain full ownership of data transactions. In general, Ethereum accounts are run on the basis of transactions between the sender and the recipient that have been signed and spent digital money (Ether) for each transaction. Technically, Ethereum uses a machine state approach to manage transactions. The transfer of transactions from state to final state begins with the genesis state [45] that holds the various data.

3.8 Smart contract

Szabo introduced smart contracts in 1994 [46]. Initially, smart contracts were not fully utilized until BT became popular. Smart contracts have begun to gain attention after the rapid development of BT in various sectors. The adaptation of smart contract technology has led to the norm for the preparation of contracts in business procedures. On the other hand, conventional contracts run transactions through a middleman, as opposed to smart contracts that are enforced automatically without the intervention of a middleman. The implementation of smart contracts has seen improved management quality, operating cost savings, and time as well as risk reduction [47]. A smart contract consists of programme code developed to run and simulate business functions in the real world. The programme will be run once it meets the set criteria or conditions. This allows the agreement document to be executed automatically after activation occurs in the smart contract when all conditions are met. A smart contract is synonymous with contracts developed in the Ethereum environment [48]. Program code is developed in bytecode format and execute in the environment of the Ethereum virtual machine (EVM). The main language(programming) for smart contract development in the Ethereum environment is Solidity [49]. Meanwhile, chain code is a smart contract developed in the Hyperledger environment [21]. Among the main languages for chain code development are Java, Go, and Node.js.

4.Integration of blockchain and ML

Blockchain is a decentralized ledger that stores data in blocks and links them using cryptographic techniques. Based on write-only database capabilities, blockchain activities are implemented in a decentralized distributed P2P environment. Such cryptocurrencies are commonly used applications in the blockchain realm, such as Bitcoin and Ethereum. Nowadays, BT is increasingly widely used in various sectors, including health, industry, education, pharmacy, finance, and so on. As a result, it has opened up space for irresponsible parties to hack and misuse this technology for personal gain. Eventually, consumers fall victim to fraud, data leakage, loss of digital money, and so on. Despite these benefits, BT is not hundred percent safe, and it is still prone to attacks and vulnerabilities [50]. For example, a large number of Ponzi schemes have been devised to steal money from honest users, and a large volume of fraudulent accounts are being created routinely to carry out money laundering. Therefore, an effective method is to use ML technology to make early detection of transaction patterns, whether normal or abnormal. Researchers are concerned about the combination of BT and ML nowadays. Among them is research to create new ecosystems that decentralize infrastructure, data storage management, administration and ML-based applications.

ML using traditional centralized databases is changeable and unreliable. The database administrator has complete control over the database and has full access to it. Therefore, the combination of ML technology in the environment can ensure that the data does not change and remains. The role of smart contracts also makes the process done without third parties and automated using ML. There are many benefits when these two technologies are combined. This situation has inspired researchers to propose and build blockchain-ML-based solutions to enhance the effectiveness of electronic health record management and SCM and empower security aspects within the IoT and networks [51, 52]. It is more interesting when the adaptation of the ML algorithm in the smart contract is impossible to do [53]. *Figure 7* displays a representation of the architecture for ML adaption in a BT-based application.

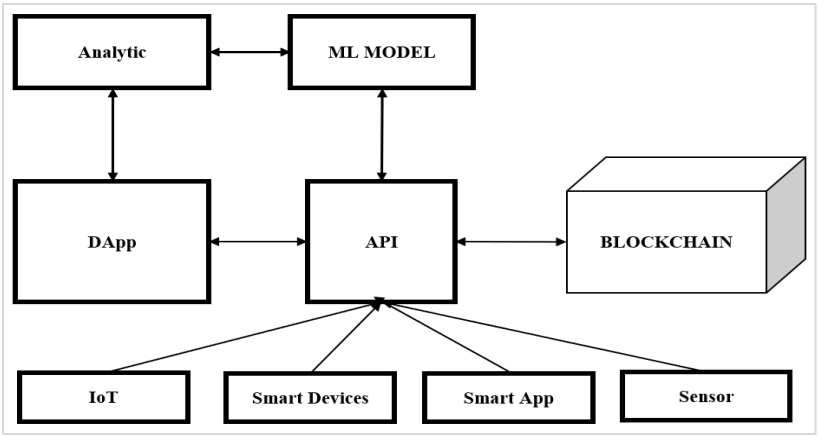

Figure 7 Blockchain ML adoption

Blockchain applications in the construction of ML models provide many advantages compared to traditional database ecosystems that manage data centrally. *Figure 8* demonstrates the benefit of ML in the blockchain network. ML algorithms have incredible possibilities for learning.

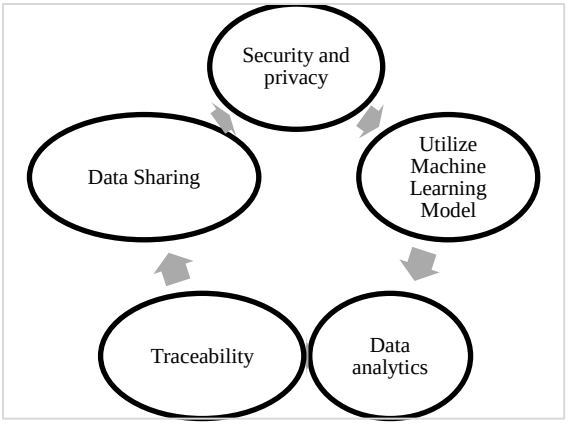

Figure 8 The Advantages of Using ML in blockchain

These skills can be applied to the blockchain to make the chain smarter than previously. Furthermore, we can create much better ML models leveraging the decentralized data architecture feature of BT. This integration can be useful in the improvement of the privacy, security and transparency of the distributed ledger network. For example, in improving the management of sensitive data in the health sector, Medrec features secure accountability, confidentiality, and authentication[54]. Computational features found in ML can reduce processing time and improve aspects of data sharing. This goal has been achieved through a combination of cryptographic mechanisms, distributed storage management, and consensus operations by BT. The advantages found in blockchain have resulted in better data value and avoiding data silos [55].

In the analytical aspect, the ML model is developed using data stored in a blockchain network to make predictions. This can be seen in past research using a combination of data sources from blockchain, smart device tools, sensors, and IoT devices. Integrated integration between ML and blockchain provides real-time data analysis. This can be seen in terms of improving data quality, such as avoiding data duplication and cleaning up the data. In addition, the integration of these two technologies has provided space for researchers to study aspects of data structures. The study focuses on data security issues for personal data protection and data analysis issues to produce predictions for people's habits [56].

Integration of ML models can assist assure the sustainability of terms and conditions that were agreed upon before. In addition, the ML model is updated according to the nodes(chain) environment of the distributed blockchain network. The traceability feature in the blockchain allows data to be tracked in detail, starting from the genesis initiator block. This has increased the value of transparency and traceability of the data recorded on the distributed node network. Pertaining to the traceability of the BT in IoT, we can also analyze the hardware of different machines so that ML models will not stray from the learning path for which they are allocated in the environment.

5.Anomaly detection method

Blockchains can be exposed to several forms of harmful assaults and fraudulent activities [57], which possibly can be discovered by studying the transaction patterns. Thus, unusual behaviour in data

transaction patterns is termed abnormal detection [58]. Authors [59] explains it as a process which is used to detect typical patterns in data which are different from the normal behaviour of the complete dataset. The ML algorithms would help distinguish between transactions behavior among user accounts by learning the associated attributes that pertain to either anomalous or normal behavior. The method of identifying normal and abnormal transaction patterns is also called outlier detection. Therefore, security on blockchain networks can be protected using an abnormal detection approach. This is done by making a prediction about how often abnormal transactions will happen based on data that has been looked at to find signs of fraud or attack in blockchain transactions.

5.1 ML classification model

The ML method is a prominent model used for anomaly detection in the distributed blockchain network. In general, the ML model is capable of generating predictive data (as output) in the future based on the experience of past data (as input). Thus, the following section shows the ML experiments with anomaly detection analysis in the blockchain network. The workflow process implemented in the ML model is illustrated in *Figure 9*.

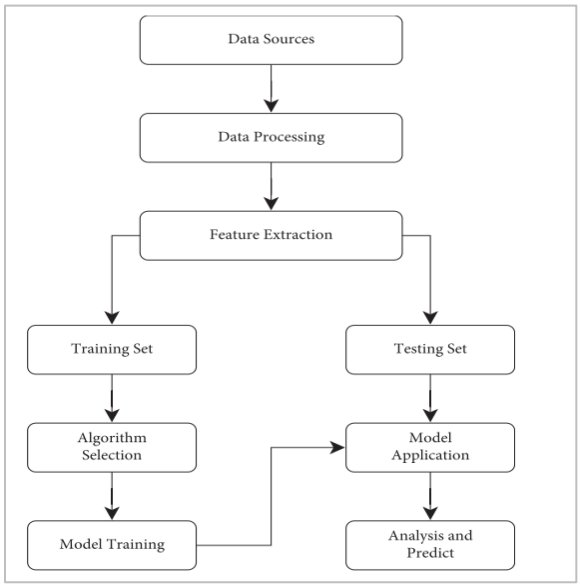

Figure 9 Typical workflow of ML [56]

This process is initiated in the training process through pre-processing of the raw(original) data input. Then, feature extraction is performed, and models are trained using this data. Feature extraction should also be performed in the training phase to produce a final model using test data, and finally, the analysis process is performed. ML provides several types of algorithms on the appropriate model based on the data set provided to produce intelligent results. Basically, the ML model being developed is based on training data, where each part of this data has inputs and outputs. In the training phase, analysis of the estimated distance between inputs and outputs is used to help generate the model. Subsequently, further analysis of normal or abnormal behaviour in transactions is identified and estimated. Four primary methods of ML are reviewed such as supervised, unsupervised, reinforcement learning (RL), neural network (NN) and deep learning (DL). Nevertheless, in this review, we focused on supervised and unsupervised approach. The difference between the two approaches is shown in *Table 4*. Correspondingly, *Figure 10* depicts the ML taxonomy.

5.1.1Supervised learning

In ML and AI, supervised learning is one of the most widely used methods to make predictions. This method trains new algorithms to make more accurate predictions by feeding them sets of data that have been labelled [60]. In the validation process, the adjustment of weights based on data input is made to train the model to reach a good level of suitability. Next, a learning supervision process is implemented to produce the desired model as an output using training data. Clearly, the inputs and outputs in the training data set will help model learning. The accuracy of the model was determined using measurement methods and modified to reach an error-free level. For example, the author [61] used supervised ML approaches to detect fraudulent activity. In general, the classification of models in supervised learning is based on two problem scenarios: classification and regression.

Classification refers to how algorithms are used to classify data based on specific categories. The classification process involves the identification of data entities by recognizing data for labelling purposes. There are several models commonly used in this classification technique, namely Random Forest (RF), Ensemble Methods, K-Nearest Neighbour (KNN), Decision Trees (DT) and Support Vector Machines (SVM). Besides, there are also studies to improve the accuracy and performance of the classification model in making predictions through the ensemble learning approach.

In general, the ensemble method combines predictions generated from several individual models,

also called weak models, into new and strong models. Typically, individual models are biased because they consist of many variants. Thus, the main intention of the ensemble learning method is to reduce the value of variants and bias by mixing them to become a strong and high-performing new model [62]. In addition, an idea-based ensemble method to help make decisions based on several views [63]. Therefore, the final decision is based on a predictive ensemble that goes through a voting and averaging procedure. The ensemble technique is made up of two main steps: building the classification of the ensemble and combining or integrating the ensemble [64]. However, there are researchers who employ a three-phase strategy by incorporating an ensemble pruning step between ensemble construction and ensemble combination [64]. In the real environment, ensemble techniques are also widely used in DL approaches. Therefore, a study conducted using medical datasets has proved that the accuracy of predictions is high using the method of combination of classifiers compared to individual classifiers in the DL approach [65].

An important component of the regression technique is the link between the dependent and independent variables. This method is often used to produce forecasts in this scenario, such as sales forecasts, pricing, markets, stocks, etc. Polynomial regression, logistical regression, and linear regression are three popular regression approaches [66].

Table 4 Comparison of supervised and unsupervised

Item/Learning Method	Supervised	Unsupervised
Type of Data	Labeled data is used to train algorithms.	When dealing with unlabeled data, algorithms are used.
Level of Complexity	Easier method	Difficult to calculate
Level of Accuracy	Method that is extremely exact and reliable	Method that is less accurate and reliable

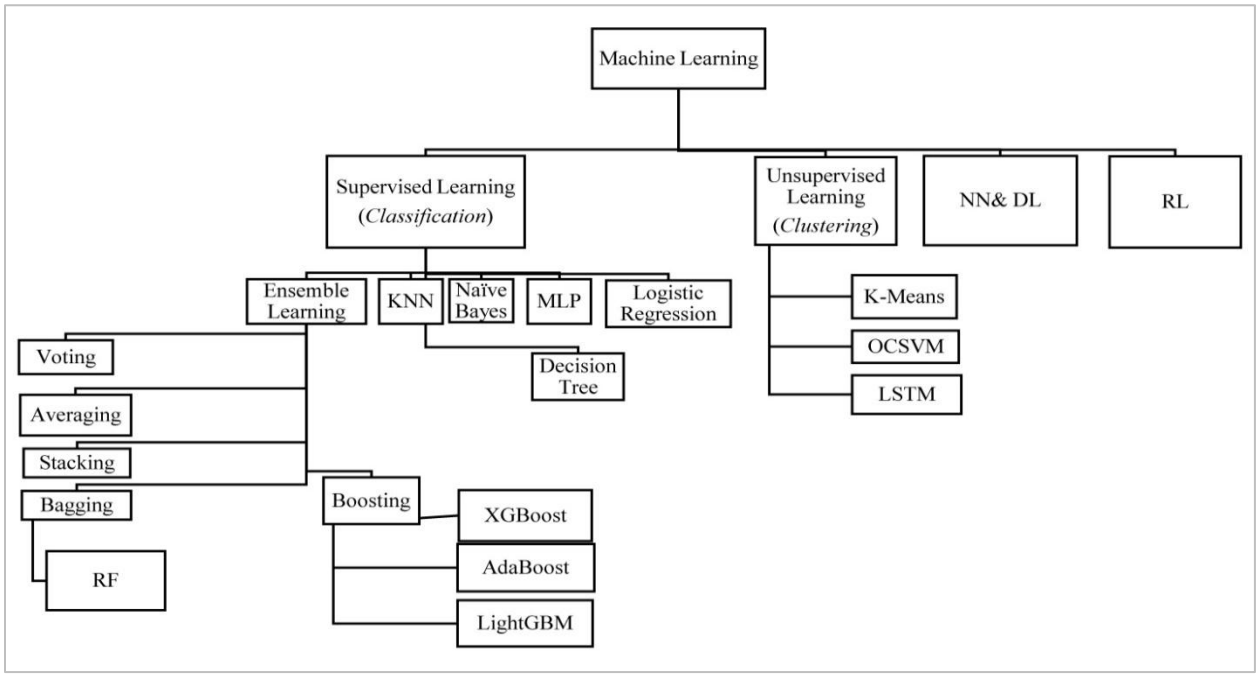

Figure 10 Illustrates the ML taxonomy

5.1.2 Unsupervised learning

Essentially, unlabeled datasets used for the purpose of model training are called unsupervised learning. In unsupervised learning, it is important to look at hidden data and know which data can be split up into different sets. In terms of supervision, users do not need to monitor the model as unsupervised learning is one of the ML techniques. Thus, the process of identifying transaction patterns is determined by a model that functions alone. Among the advantages of unsupervised learning is that it allows users to complete complex processing tasks as opposed to supervised learning. However, it is becoming more

unpredictable in unsupervised learning compared to other models. Examples of unsupervised learning models include NN, anomaly detection, and clustering. Further, within the Bitcoin network, fraud is detected using unsupervised learning techniques [67]. Therefore, among the things that cause unsupervised learning techniques to be chosen are knowing unknown patterns in data transactions, finding features for categorization, analyzing and labelling in real-time, and the fact that unlabeled data is more easily obtained than labelled data, i.e., requiring manual intervention. Generally, the classification of unsupervised learning based on problem challenges is divided into clustering and association. Among the important ones is clustering in unsupervised learning. It entails the detection of patterns as well as data structures with uncategorized data sets. Among the common models for clustering are K-Means, K-NN, and Principal Component Analysis (PCA) [68]. The data elements in the database will be built into associations according to the rules of the association. As a whole, correlations between variables in the database are known through unsupervised learning techniques.

5.2 Analysis of anomaly detection model

The goal of this review is to collect research articles about how supervised and unsupervised learning can be used for anomaly detection. In this review, a total of 71 previous research published papers related to the detection of abnormalities in ML are analyzed from 2017 to 2022. For details, past research related to supervised learning is shown in *Table 5*, while *Table 6* lists unsupervised learning. The data from this obtained study article is arranged on a graph by classification ML type to aid in the review's analysis (see *Figure 11)*. The analysis of this figure has shown that a total of 59 research articles were conducted using the supervised learning approach, which makes this approach the most dominant of the entire research articles. This analysis also shows that unsupervised learning began to be used by researchers in 2017 and continued to be used until 2022. Overall, from 2017 to 2022, supervised learning has been very widely used in abnormal detection research.

Table 5 The selected research article using supervised learning

Year	Reference	Type	Model	Blockchain	Application	Conclusions/findings
2017	[69]	Conference	RF + XGBoost	Bitcoin	HYIP (High Yield Investment Program)	With a true positive count of less than 4.4 percent, a total of 83 HYIP cases were successfully detected.
2017	[70]	Conference	Ensemble	Blockchain-based (Ripple)	Anomaly detection	Concluded that the detection of anomalies in human behavior is very important in traditional or modern payment systems
2017	[71]	Journal	RF	Bitcoin	Fraud detection	RF (achieved 99.99 percent), GLM logistic, and boosted regression models all outperform each other by more than 90 percent.
2018	[17]	Journal	RF	Cryptocurrency	pump and dump scams	The average AUC score across all coins is 0.74.
2018	[60]	Journal	RF	Bitcoin	Ponzi Scheme	Overall, the RF classifier is marginally more effective, according to the studies.
2018	[60]	Journal	RF	Bitcoin	Ponzi Scheme	The results of the study showed a true positive value of 1 percent, and the top classifiers managed to detect 31 ponzi schemes.
2018	[72]	Journal	Gradient Boosting algorithm (ensemble)	Bitcoin	Anomaly Detection	Obtain a 77 percent accuracy rate, and an F1-score of 0.75. The best performance comes via gradient boosting.
2018	[73]	Journal	XGBoost	Ethereum	Ponzi Scheme	The results demonstrate that 45 of the 54 contracts (83 percent) are clever Ponzi schemes.
2019	[2]	Journal	secureSVM	Blockchain-based	IoT	Proved secure SVM's efficiency and security
2019	[19]	Journal	Ensemble DT	Bitcoin	crypto-ransomware	The LA proposed model produced the lowest FPR value (1.56 percent) by analyzing classifying hostage software compared to RF, Naive Bayes (NB), and Ensemble (NB with RF).
2019	[20]	Journal	J48	Ethereum	Ponzi Scheme	Among the three classifier models, J48 has the best recall of 0.872.

Year	Reference	Type	Model	Blockchain	Application	Conclusions/findings
2019	[61]	Journal	RF	Ethereum	Fraudulent Accounts	Three alternative classifiers were investigated, and RF yielded the best results in recall and false-positive rate.
2019	[74]	Journal	RF	Bitcoin	High Yield Investment Programs (HYIP)	A study result of 93.75 indicates the accuracy of fraud detection in Bitcoin
2019	[75]	Journal	Ensemble + Cascading ML (CML)	Bitcoin	Ponzi Scheme	The results of the study showed the precision value was 0.95, the F-score was 0.79, and the recall value was 0.69
2019	[76]	Journal	RF	Bitcoin	Address Identification	Voting-based methods have shown better performance than non-voting-based methods (simulated data of more than 200K Bitcoin addresses) calculated based on F1 score values, recall, and precision.
2019	[76]	Conference	RF + node2vec	Ethereum	Network Traffic	We measured the nodes' features and their connections' properties by thoroughly evaluating the dataset.
2019	[77]	Journal	LightGBM	Bitcoin	Address Identification	Macro-F1 performance was highest with 87 percent and 86 percent performance using LightGBM
2019	[78]	Conference	RF	Ethereum	Vulnerability detection	The model is able to discover vulnerabilities efficiently and fast. The results of our model's evaluation of 49502 real-world smart contracts confirm its usefulness and efficiency.
2019	[79]	Journal	SVN + Decision Tree + RF	Ethereum	Security Analysis	Our model correctly identified a critical software flaw (accuracy of 95 percent).
2019	[80]	Journal	Naïve Bayes	Cryptocurrency	crypto jacking detection	Accuracy 0.973 Average F-Score: 0.973
2019	[81]	Conference	XGBoost	Cryptocurrency	pump and dump scams	After applying the model to the entire time series of 172 coins, the model identified 612 pump-like occurrences.
2019	[82]	Journal	RF	Cryptocurrency	anomalous transactions	Employed a high-precision RF technique to identify suspicious wallets
2020	[18]	Conference	Ensemble	Bitcoin	Anti-Money Laundering (AML)	The result is accuracy (98.13 percent), Precision (99.11 percent), Recall F1 (71.93 percent) and score false 83.36 percent
2020	[22]	Journal	KNN	Blockchain-based	Electricity Network	The rates of incidence of anomalies we used were 1, 2, 3, and 4 percent, respectively. Over the course of 50 runs, we evaluated the rate of effective anomaly identification.
2020	[48]	Journal	LightGBM	Ethereum	Honeypot	The honeypot on a smart contract was successfully detected, with the study results being an F1 value (0.93) and AUC value (0.99).
2020	[82]	Journal	RF	cryptocurrency	Fraudulent Transactions	The result is Precision (0.96 percent), Recall (0.96 percent) and F1 Score (0.96 percent)
2020	[83]	Journal	RF	Cryptocurrency	Pump and Dumps	n/a
2020	[84]	Journal	Ensemble + XGBoost	Ethereum	Honeypots	Even when all contracts relating to one honeypot technique were removed from training, the ML models demonstrated to generalize effectively.
2020	[85]	Journal	KNN	Bitcoin	Abnormal transaction	The results of the analysis show that KNN successfully detects suspicious transactions at the nodes.
2020	[86]	Journal	Isolation forest	Blockchain-based	Battery Health	In comparison to the well-known anomaly detection system, experiment results show

Year	Reference	Type	Model	Blockchain	Application	Conclusions/findings
						that the method enhances the F-score values by up to 25.65 percent.
2020	[87]	Journal	SVM	Ethereum	IoT	The result is Evaluation Metrics (0.99), Accuracy (0.9998), Recall (1) and F1-Score (0.9998)
2020	[88]	Journal	Ordered Boosting	Ethereum	Ponzi Scheme	On a real-world dataset, the new model obtains a 98 percent F-score, greatly outperforming existing techniques.
2020	[89]	Journal	RF	Cryptocurrency	Cryptojacking	On our dataset, the BRENNTDROID tool can detect miners with 95 percent accuracy.
2020	[90]	Journal	SVN + KNN	Blockchain	Malicious Users	KNN and SVM are better choices because they require a third of the resources of CNN algorithm and have accuracy values higher than 0.9, which is 0.9 percent lower than CNN.
2020	[91]	Conference	XGBoost	Ethereum	Malicious Account	That assessment is 96.21 percent accurate, with only 3 percent false positives.
2020	[92]	Conference	GCN (Graph Convolutional Network)	Bitcoin	Money Laundering	The result is F1-Score (0.773), Recall (0.678), Accuracy (0.974) and Precision (0.899).
2020	[93]	Journal	RF	Ethereum	fraudulent behaviour	The findings of this study showed the detection of fraudulent behaviour produces good results using RF
2020	[94]	Journal	RF	Bitcoin	money laundering	The results of the study showed an inability to detect illegal (abnormal) activities using unsupervised learning techniques.
2020	[95]	Journal	Ensemble (RF, Stacking Classifier, and AdaBoost)	Ethereum	Malicious Transaction	The ensemble approaches perform well (F1 score of 0.996).
2020	[96]	Journal	XGBoost	Ethereum	illegal activity	XGBoost classification mode swiftly and successfully detects illicit behaviour on the Ethereum network.
2020	[97]	Journal	Ensemble DT	Bitcoin	illicit entities	According to the study's findings, 66 percent of users were correctly classified using the proposed model
2020	[98]	Journal	Multi-layer Perceptron (MLP)	Bitcoin	Scam	We found 6,395 addresses explicitly offered by scam cases by actively hunting for them and using ML to identify the findings.
2021	[9]	Journal	XGBoost	Ethereum	Vulnerability Detection	The Ethereum Micro-F1 and Macro-F1 give a Turing-complete Ethereum Virtual ContractWard that is over 96 percent accurate.
2021	[14]	Journal	KNN algorithm (Stacking ensemble)	Blockchain-based	intrusion detection	Studies show that ensemble stacking techniques increase the level of effectiveness by 95 percent in forecasting analysis.
2021	[99]	Journal	RF + Adaboost + SVM	Ethereum	Fraudulent Transactions	The result of accuracy is 0.97 percent
2021	[100]	Journal	AdaBoost	Ethereum	Phishing	The result analysis for data label (Phishing) is Precision-0.83, F1 score-0.74 and Recall-0.66.). The result analysis for the data label (No Phishing) is Precision-0.92, F1 score-0.94 and Recall-0.97.
2021	[101]	Journal	XGBoost	Bitcoin	Blockchain simulator	A product called BlockEval formulated the first simulator to use real Bitcoin data for simulation purposes.
2021	[102]	Conference	RF	Ethereum	Credit Card Fraud detection	This study shows that RF produces high true positive values compared to other models.
2021	[103]	Journal	Ensemble	Ethereum	Under-priced	The DT is the best strategy of the other

Year	Reference	Type	Model	Blockchain	Application	Conclusions/findings
			(Decision Tree, RF, KNN, SVM and Naïve Bayes)		DoS attack	models by giving good results with F and AUC-ROC score values.
2021	[104]	Journal	RF	Ethereum	Fraud Detection	This study showed the adjustment of the data set, and the characteristics gave good results (99%) in terms of accuracy of all the classifiers tested.
2021	[105]	Journal	Naïve Bayes	Bitcoin	untrusted users of cryptocurrency transaction services	For both datasets, the accuracy finds that the Nave Bayes algorithm performs better than other classification algorithms.
2021	[106]	Journal	Ensemble (DT)	Ethereum	malicious accounts	Studies using the ensemble technique (ExtraTreesClassifier) have successfully detected suspicious accounts with a balanced accuracy with a range of 87.2 and 88.7.
2021	[107]	Journal	Isolation forest	blockchain-based	IoT	The research results indicate malicious attempts were successfully detected.
2021	[108]	Journal	RF	Bitcoin	Fraudulent Transactions	The results of the analysis show that RF has achieved the highest results compared to other models, that is, the value of F1 (95.9%).
2021	[109]	Journal	Logistic Regression	Ethereum	fraud detection	We were able to forecast fraud transactions with 94 percent confidence using both approaches, which is promising.
2021	[110]	Journal	RF	Ethereum	fraud detection	The results demonstrate that by employing the three algorithms, time measurements improve significantly, and the RF approach improves the F measure.
2022	[111]	Journal	Isolation forest	Blockchain	Social Media	This study uses tree algorithms for abnormal detection and gives good results.
2022	[112]	Journal	Adaboost	Blockchain-based	IoT	The results showed that Adaboost produced good results with precision (97.9%), recall (95%), and F-score (96.3%).
2022	[113]	Journal	Ensemble Boosting	Cryptocurrency	anomaly detection	Studies show the Ensemble Boosting technique produces good performance compared to other models.

Table 6 Selected research paper using unsupervised learning

Year	Reference	Type	Model	Blockchain	Application	Conclusions
2018	[15]	Journal	LSTM + RNN	cryptocurrency	crypto jacking	The results of the research yielded the accuracy of the BMDetector prototype (93.04%)
2019	[114]	Journal	DBSCAN	Hyperledger fabric	Water Network	The result is accuracy (0.94155), PR (0.9433), Recall (0.9969) and F1 score (0.96937)
2019	[115]	Journal	OCVM+ K-Means	Bitcoin	Anomaly Detection	The study successfully detected 6 DDOS attacks and 5 multiple spending attacks.
2019	[116]	Journal	Gaussian Mixture Model	Bitcoin	Anomaly Detection	The results showed that abnormal users were successfully detected among the entire user population in this study.
2020	[117]	Journal	LSTM	Blockchain-based	Electricity Network	The results show the effectiveness of the proposed framework for identifying abnormal patterns in transactions. FPR (0.01 percent), UNSW-NB15: DR (99.80 percent), FPR (2.93), and Power System: DR (96.27 percent) were the readings.
2020	[118]	Journal	LSTM	cryptocurrency	Cryptocurrency	For the analysis of crypto investment

Year	Reference	Type	Model	Blockchain	Application	Conclusions
					Deception	transactions, LSTM achieves 98.99 percent accuracy.
2020	[119]	Journal	K-means	Ethereum	behavioural traits in transactions	This study looked at behaviours on transactions using supervised learning and unsupervised learning.
2021	[120]	Journal	OCSVM	Ethereum	Performance Testing	This study has successfully detected Ponzi schemes (1,621 cases) with an F-score (96 percent).
2021	[121]	Journal	LSTM + RNN	Blockchain-based	Privacy Protection	This study uses Matthews Correlation Coefficient (MCC) measurement to make a prediction, and the result is a range value between (-1 to 1).
2021	[122]	Journal	LSTM	Ethereum	Anomaly Detection	This study focuses on the evaluation of anomaly detection in smart contracts, covering the type of contract and identifying malicious contracts.
2021	[122]	Journal	Deep Autoencoder NN	Ethereum	Anomaly Detection	The result is Precision (0.988), AUC (0.9891), Recall (0.988) and F1-Score (0.988).
2022	[16]	Journal	OCSVM	Ethereum	Phishing	The result is Precision (0.927), Recall (0.893) and F-Score (0.908)

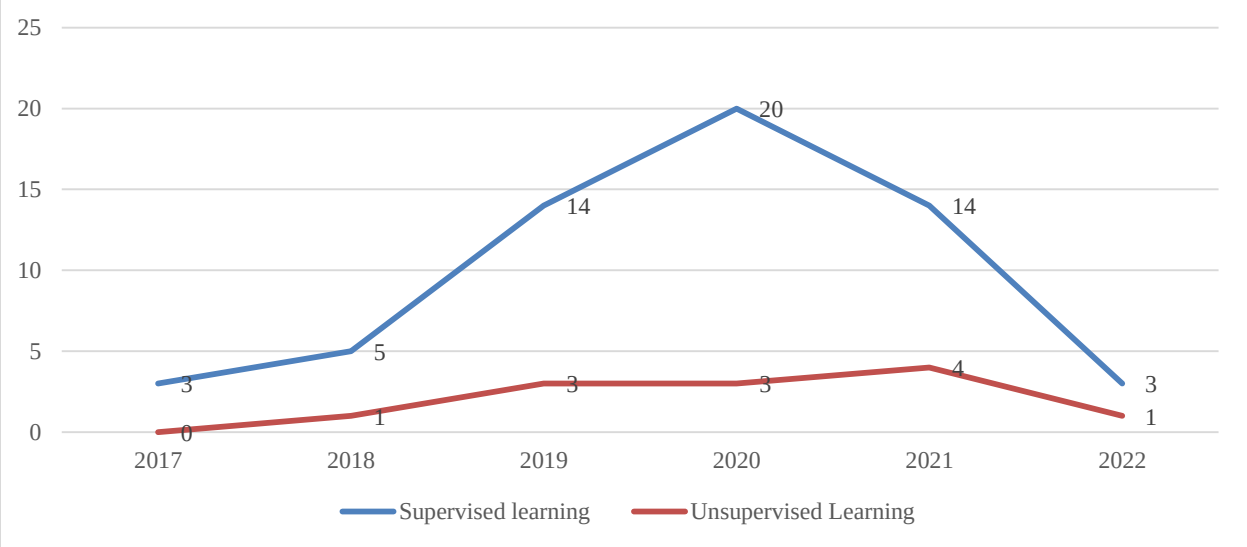

Figure 11 Classification of anomaly detection per year

Now we will go over the analysis of the ML model in additional depth, as given in *Table 7* and displayed in graft as represented in *Figure 12*. According to the data in *Figure 11*, RF is utilized in 22 research publications and is the most popular model utilized from 2017 to 2022. In addition, researchers have applied the ensemble approach in 10 papers, with an increasing tendency from 2017 to 2022.On the other hand, 6 research papers employed the XGBoost model for anomaly detection categorization.

Table 7 ML and its classifiers used in research

ML	Classifier model	Reference
Supervised Learning	Ensemble Method	[14, 18, 19, 70, 72, 75, 84, 97, 95, 103, 106, 113]
	AdaBoost	[99, 100, 112]
	GCN (Graph Convolutional Network)	[92]
	Isolation Forest	[86, 107, 111]
	J48	[20]

	KNN	[22, 85, 90]
	LightGBM	[48, 77]
	Logistic Regression	[109]
	MLP (Multi-layer Perceptron)	[98]
	Naïve Bayes	[80,105]
	RF	[4, 17, 50, 60, 61, 71, 74, 76, 79,78,82,89,83,93,94,99,102,104,108,110]
	secureSVM	[2, 87]
	SVM	[79,86,90,99]
	XGBoost	[9,17,73,81,96,91,101]
	Decision Tree	[79]
Unsupervised Learning	OCVM (one-class support vector machine)	[16,115,114,120,123]
	DBSCAN	[21]
	Deep Autoencoder NN	[122]
	Gaussian Mixture Model	[116]
	K-means	[114,115,119]
	LSTM	[15, 117,118,121,122]

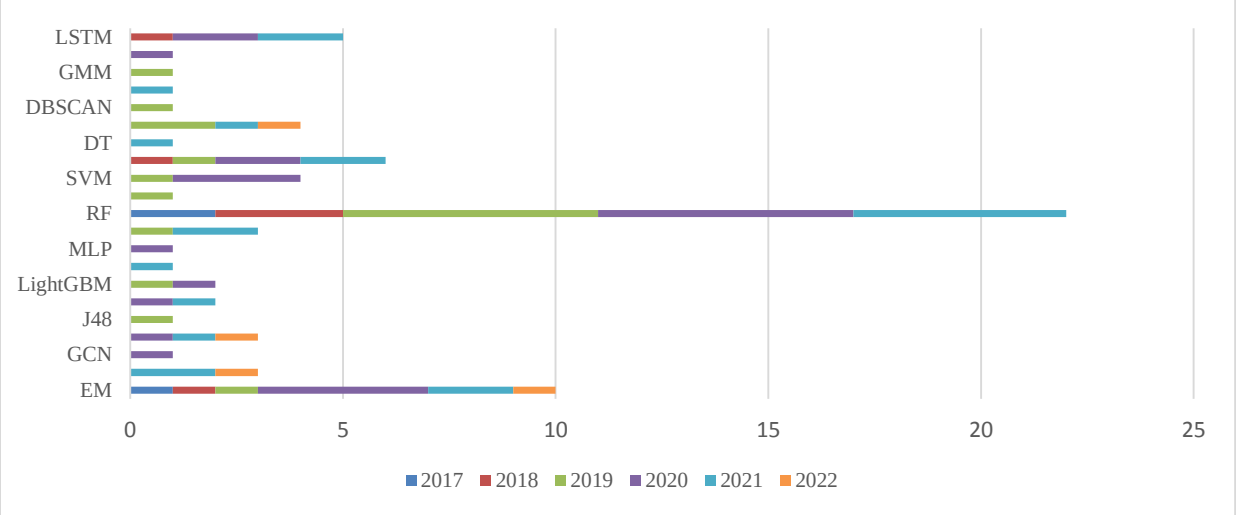

Figure 12 Anomaly detection model per year

5.2.1 Dataset and tools

Referring to the authors [13], anomaly detection is a change in the operating system that will occur if suspicious transactions are identified during data processing. Therefore, the main objective is to adapt anomaly detection for early detection of traction that shows suspicious patterns on the entire data set. The authors previously mentioned that they use datasets acquired from a single source to uncover abnormalities in data transactions using normal and abnormal labelling in a previous research paper. Consequently, [124] offered a supervised technique for detecting fake Ethereum accounts. This research has randomly extracted data taken from Etherscan (349999 wallets) and identified a total of 2200 wallets used for criminal purposes. Researchers, on the other hand, utilized more than one data source to develop this study to classify typical and abnormal transactions. For a better analysis, there are two sets of Ethereum transactions on etherscan.io. The first data set contains about 420 fraudulent wallets found in the etherscamdb.info database, whereas the second data set contains non-fraudulent actions found in the etherscan.io database [93]. Datasets are an important

aspect of the development of ML models. Apart from that, the method or way the data is obtained is dependent on the tool to access the data. This is described in *Table 8*, which lists the datasets and tools used in previous research to produce the respective studies. Referring to *Table 8*, a total of 32 different data sets were used in most experiments.

Besides, from the information listed in *Table 8*, we summarize that the datasets may be grouped as live datasets, media social, open datasets, and tools datasets, as illustrated in *Figure 13*. From the data in *Figure 13*, it is evident that the most commonly used group of datasets in this selected research paper was open datasets.

The frequency analysis of the datasets used in the previous study is shown in *Figure 14*. Datasets accessed live from Etherscan sources are the most popular approach used by researchers. In addition, 5 research publications employed open datasets from Elliptic and 4 research studies adopted open datasets from the Computational Biology Lab (University of Illinois) and Bitcoin.

Tools are ways or means for obtaining data for analysis, integration, development, and study that come in the form of application software, web platforms (API, Web Services), libraries, and smart devices. These technologies are used to access data as datasets during the data collecting phase of the creation of ML models. Following the data processing and ML model construction phase, these tools are categorized into APIs, Development Tools, Software, and smart devices, as shown in *Figure 15*.The utilization of development tools like Python, Weka, and others is prevalent in research articles about anomaly detection applications.

Analysis in *Figure 16* displays the frequency with which different tools are employed in research articles. The most often used method in research for anomaly detection is access to Etherscan via APIs, which is described in 10 research articles. In addition, 7 research articles leveraged the Python library as tools for ML model construction. Previous studies also employed a range of methodologies or instruments to generate more accurate models. For example, the authors [103] utilizes Ganache as a personal Ethereum blockchain, a database to keep live Ethereum transactions collected via service APIs (Etherscan API), and web3.py to request and receive transaction execution results from the Ethereum in the face of an under-priced denial of service (DoS) attack.

Table 8 Previous research articles with the use of different datasets and tools

Tools	Dataset
• Binance API[81]	• anonymity-in-bitcoin.blogspot.com[108]
• Chrome Web Driver[118]	• Binance[82][81]
• Ethereum Client[16][122]	• Telegram Data[17][81][83]
• Etherscan API[9][16][20][50][61][82][96][93][100][122]	• Twitter data[17][83]
• Geth Client[96]	• Bitcointalk[60][88][83][108]
• Github[60][104]	• Etherscan[9][16][20][50][61][73][82][96][93][100][119][122]
• Google Big Query[88][104][119]	• https://goo.gl/k5PCOZ[69]
• Honeybadger[48]	• Computational Biology Lab (University of Illinois)[71]
• Parity Client[122]	• Google BigQuery[88][104]
• Pycharm[82]	• PonziTect[88]
• Python[18](89)	• Elliptic[18][50][94](89)[118]
• R[69][85][102]	• EtherScamDB[16][96][93]
• RS-232 connection[122]	• https://goo.gl/CvdxBp[20]
• Scikit-learn [70][82][108][102]	• gz.blockchair.com/bitcoin/[116]
• Selenium[118]	• Honeybadger[48]
• Telegram API[17][81][83]	• goo.gl/ToCho7[60]
• Text-blob[118]	• github.com/bitcoinponzi[60]
• Twitter API[17][83]	• blockchain.info/tags[60][74]
	• Reddit[60][83]
	• rissgroup.org/[19]
	• blockchain.com[85][115]
	• blockchair.com/dumps[85]
	• chainalysis.com[72]
	• Kaggle[110][109][113]
	• Ripple Bank[70]
	• WalletExplorer.com[74][98]
	• srg.site.uottawa.ca/bgsieeesb2020/[98]
	• bitcoincharts.com[98]

Tools	Dataset
	• VJTI blockchain Lab[97] • Xapo.com[74] • ULB[102] • Cryptocurrency Market Data[17] • Water dataset[122]
• Ethereum Client[95] • Etherscan API[79][78][84][106][103] • Honeybadger[84] • Scikit-learn[103] • BMDetector[15] • Solidity[79] • Ganache[103] • Web3[103] • Stratum[80] • WEKA[80] • VirusTotal[89] • BRENNTDROID[89] • ZombieCoin[123] • Smart Meter[121]	• Etherscan[79][78][95][84][106][103] • BMDetector[15] • Block Explorer[79] • KDD99[14] • Stratum[80] • Smart Meter[121] • VirusTotal[89] • BRENNTDROID[89] • ZombieCoin[123]
• Etherscan API[114] • Google Big Query[114][120] • Scikit-learn[77][76] • Management System[114] • NetFlow[76]	• Etherscan[114] • bitcointalk.org[77] • Google BigQuery[114][120] • blockchain.info/tags[77] • Media social login data[111] • NetFlow[76] • WalletExplorer.com[77][76][75]
• IOT Meterr[21][22] • Python[87] • Scikit-learn[107] • Web3[87] • MATLAB[87] • VirusShare[86] • IOTPOT[86] • BlockEval[101] • SimPy[101] • Smart Power Network[117]	• bitcoind[101] • blockchain.info/tags[101] • IoTID20[112] • KDD99[107] • IOT Meter[21][22] • NSL-KDD[87] • UNSW-NB15[87] • UCI ML [2] • Smart Power Network[117] • VirusShare[86] • IOTPOT[86] • NASA Battery[86]

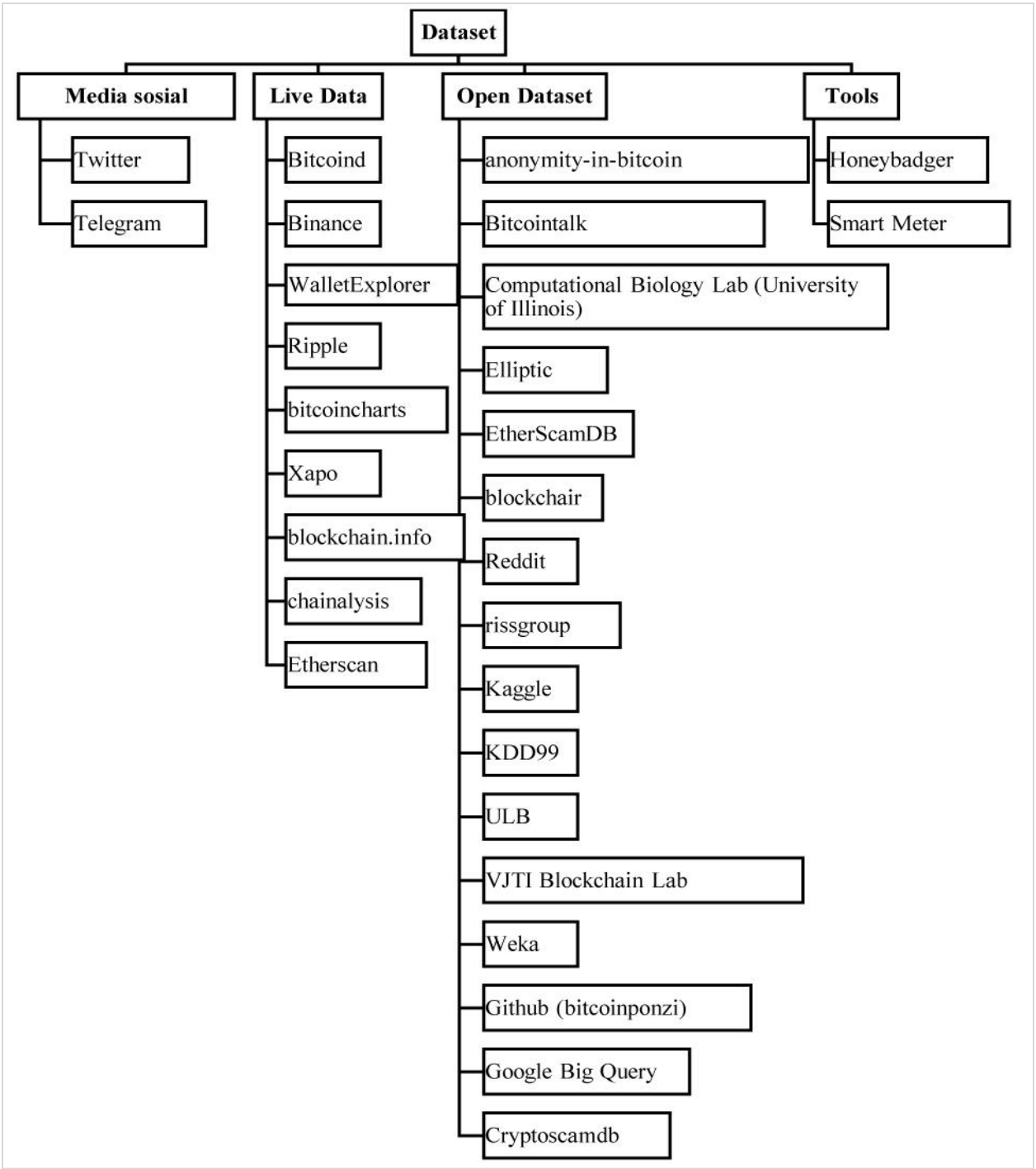

Figure 13 Illustrates the blockchain dataset from a different source

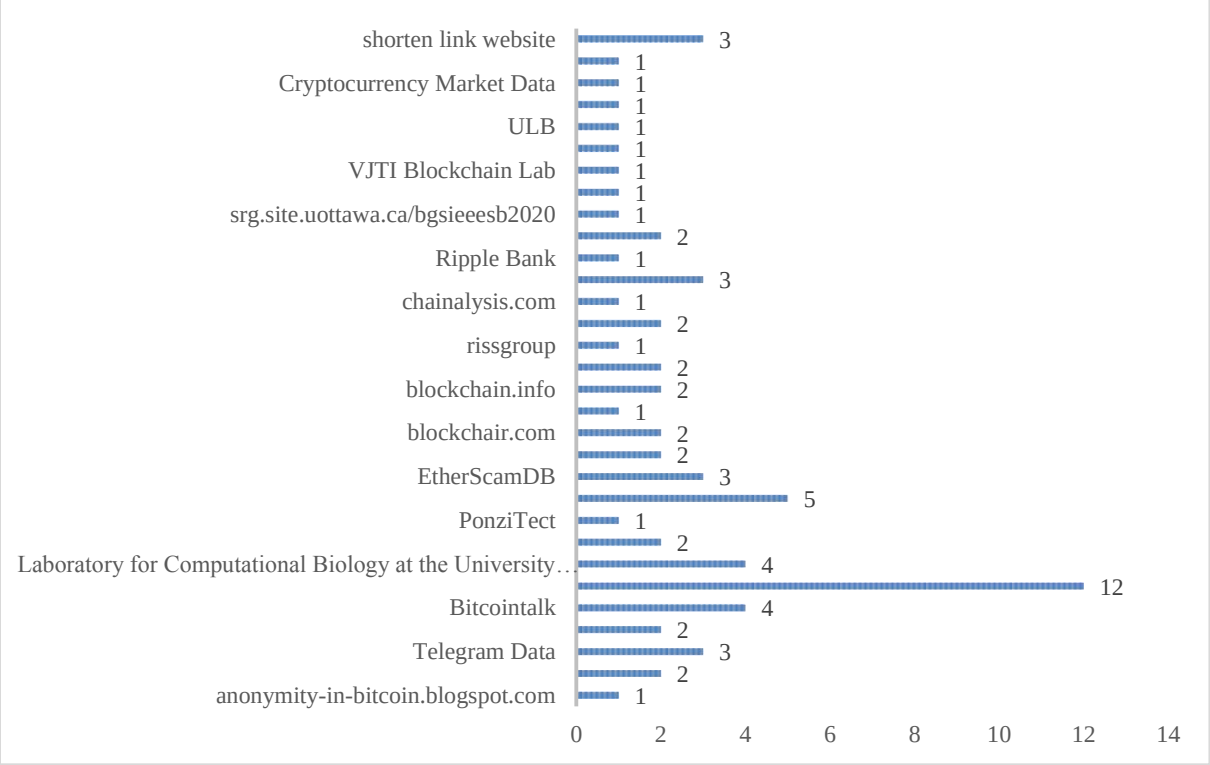

Figure 14 Utilized datasets in collected research articles

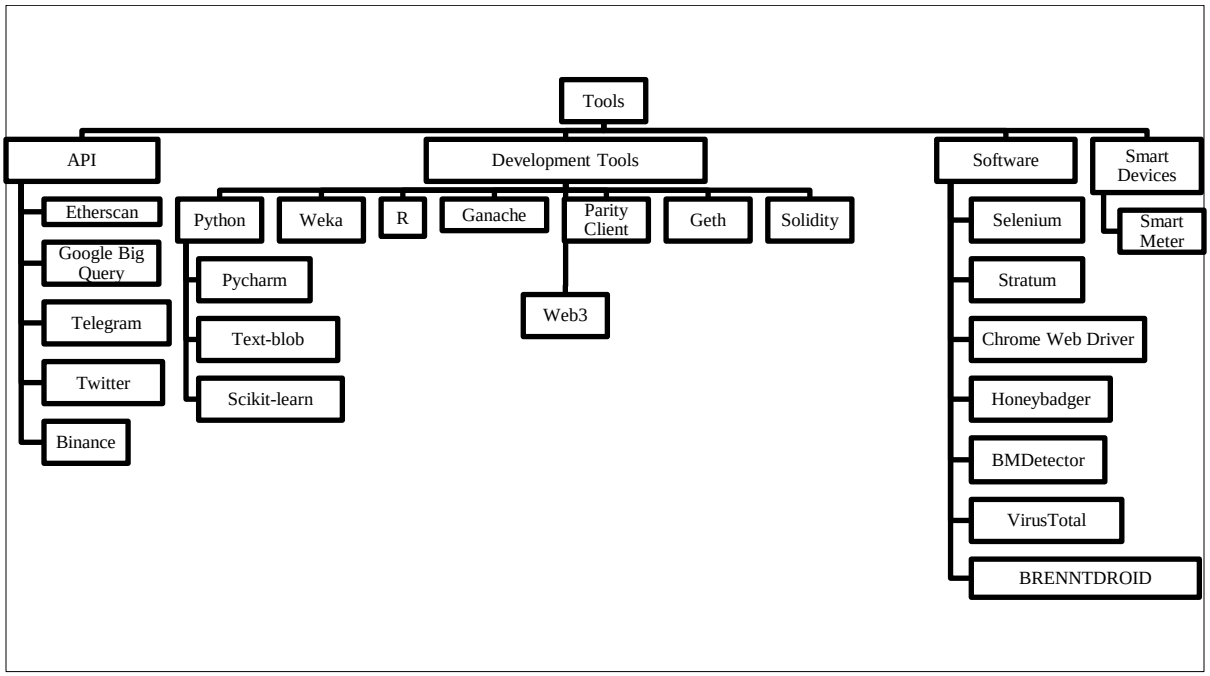

Figure 15 The tool classification for dataset readiness

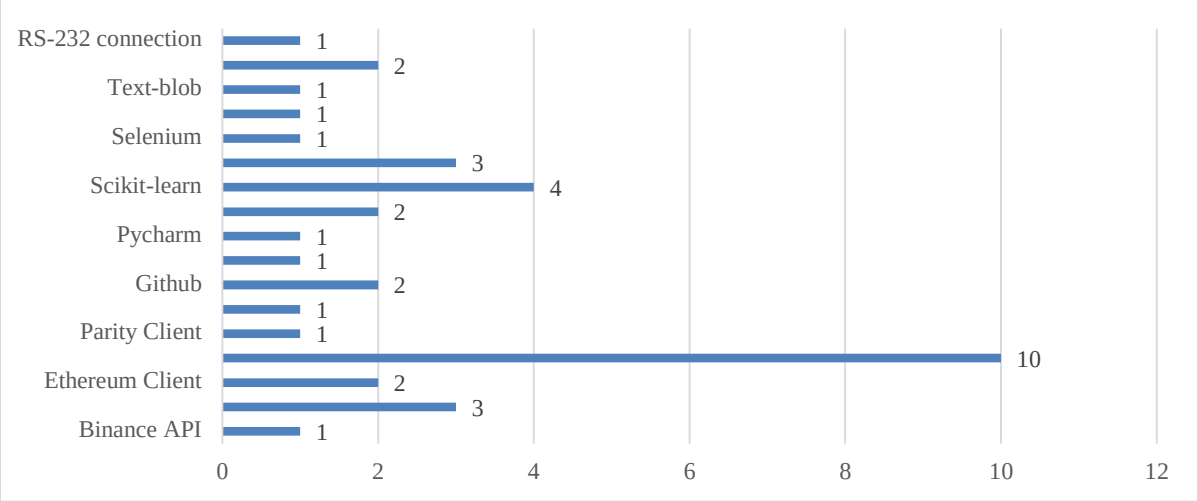

Figure 16 Tools utilized in collected research articles

5.2.2 Performance metrics

The evaluation metrics are used in ML to calculate the accuracy and performance of your trained ML models. This can assist you in figuring out how much better your ML model will perform on a dataset it has never seen before. A performance evaluation metric in ML is crucial for determining how well the proposed ML model performs on a dataset it has never seen before. In real analysis, the ML model's performance was measured in more than one way to make sure that the analysis was as accurate as possible. In total, we found 41 publications in the selected research articles that explicitly reported the performance measures of their suggested models.

Figure 17 illustrates that accuracy and F-score were the most regularly utilized performance metrics (23 papers). Furthermore, the researcher's favoured approach to determining the performance parameter is recalled, with 22 studies using recall and 14 publications utilizing accuracy. It calculates the number of anomalies that have been accurately classified. Furthermore, we noticed that 3 of the 41 articles utilized only one performance metric, and the bulk of those publications only employed accuracy,F-score and Area under the ROC Curve (AUC), which is insufficient to verify the ML model's performance.

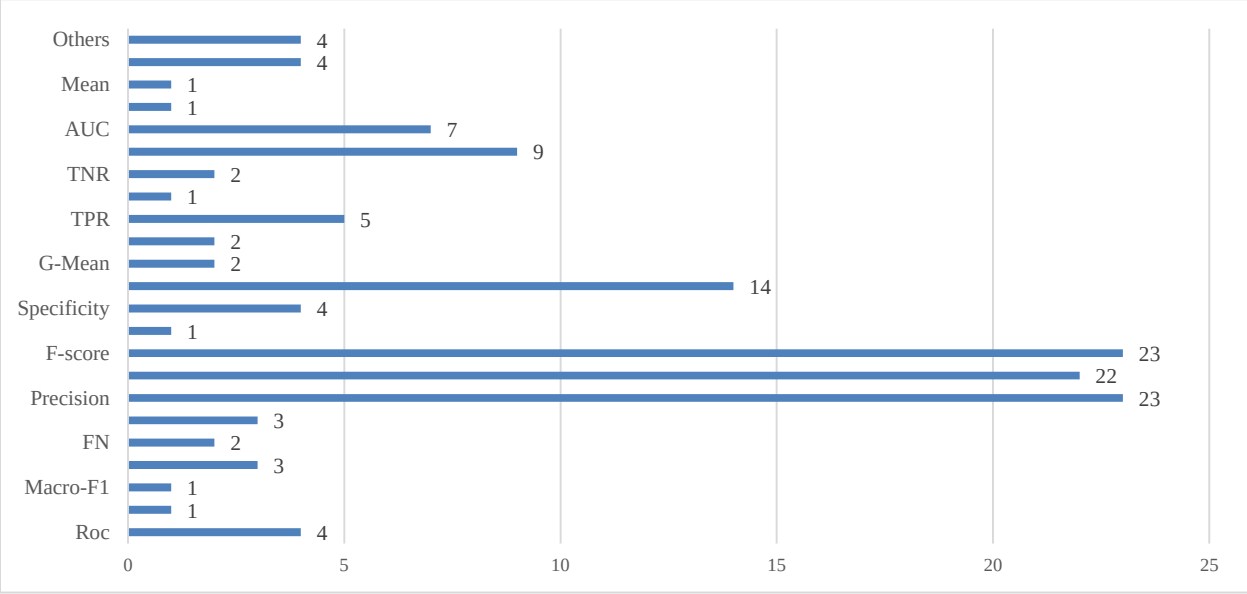

Figure 17 Frequency of performance metrics among the research papers

6.Discussion

In this study, research information was collected from 2017 to 2022. After the analysis was done, it was found that 21 ML models were used in previous research by researchers. Among the very widely used models is RF. Nevertheless, we can witness a new developing tendency, and most researchers have sought to apply the ensemble method. In terms of datasets, it was found that the entire experiment conducted in the previous research used 32 different datasets. From these observations, the live dataset category is widely used as training data and testing data in research. The most widely used approach or instrument in research for anomaly discovery is access to Etherscan via API, which is documented in 10 research articles. In addition, 7 research articles leveraged the Python library as tools for ML model construction

In general, performance measurement is an important element of looking at the performance of a final model that has been developed. The previous analysis found that 3 out of 41 studies used only one measurement method. As a result, the use of fewer measurement methods will produce less accurate experimental results. An analysis of the frequency of use of anomaly detection types was also performed. Referring to *Figure 10*, a total of 59 research studies adapted the supervised learning approach and made it the most dominant use in research. Meanwhile, unsupervised learning was used in 22 research studies. From the perspective of the model analysis, a total of 22 research publications uses RF, making it the most popular used in research from 2017 to 2022.Although the XGBoost, RF, and AdaBoost models show high usage trends in research, studies to improve performance need to be done from time to time. There has been researched on the use of ensemble strategies to improve learning performance on algorithms.

A complete list of abbreviations is shown in *Appendix I.*

7.Conclusion and future work

Blockchain is a P2P technology that uses a distributed ledger to store data in blocks. Thus, the feature makes the ledger on the Blockchain unchangeable. Although blockchain has many benefits in terms of technology, it is still prone to various security issues, operations, data management, fraud and so on. Therefore, the adaptation of ML technology to the blockchain has a positive impact through its ability to learn from past data. Furthermore, anomaly detection methods are very popular in ML approaches, which are responsible for detecting, from the outset, suspicious transactions. This indirectly helps predict strange and unusual transactions within the blockchain network. In this review, some subjects are explained, particularly the background of BT, an overview of ML technology, integration of ML with blockchain and implementation of anomaly detection methods in the distributed blockchain network. The analysis of the ML model is carried out through four main perspectives, namely the type of anomaly detection, dataset, tools, and measurement methods.Thus, the study concludes that supervised machine learning is the most commonly employed technique in previous research. Over the years, however, the usage of ensemble approaches has increased. In terms of models, XGBoost, RF, and AdaBoost are frequently employed in research. While the data source (live dataset), the Etherscan blockchain explorer platform (API), and the Python library are also the most adapted from prior studies.The selection of feature adjustments as well as extraction impacts performance improvements. In addition, research needs to use more than one measurement metric to produce more accurate results.

Based on this study's analysis, there are a number of challenges that researchers must overcome. Among them is the difficulty of immediately acquiring the most recent raw data pulled from the blockchain network, as it needs an excessive amount of storage space and a high level of technical expertise. Consequently, the majority of researchers rely on outdated datasets posted on community websites or in public repositories. The only way to create models that perform better and are more accurate is by utilising current data. This is supported by the authors [58], who advised academics and scholars to use the latest databases in their research.

In line with the development of the ML approach in the world of data science, among the potential future research by researchers is the use of auto-ML, ML Operations (MLOps) and ML designers. These are suitable for data preprocessing techniques, training data for models, evaluating model performance, experiments, and model monitoring in cloud-based. Besides, performance optimization produced by the ML approach using optimization methods such as ant colony, metaheuristic algorithms, etc., is a potential topic of study in the future.

Acknowledgment
The Universiti Sultan Zainal Abidin's Research Management and Innovation Centre funded this study.

Conflicts of interest
The authors have no conflicts of interest to declare.

Author's contribution statement
Sabri Hisham: Choosing an analysis method, searching a journal database, reading, writing, making changes, writing the final draft, checking for plagiarism, and proofreading. **Mokhairi Makhtar:** Supervision, input on draught revision, and final revision.**Azwa Abdul Aziz:** Supervision, exchange of sample manuscripts, and draught evaluation comments.

References
[1] Abd RNH. Blockchain technology in e-voting: comparative study. 2020.
[2] Shen M, Tang X, Zhu L, Du X, Guizani M. Privacy-preserving support vector machine training over blockchain-based encrypted IoT data in smart cities. IEEE Internet of Things Journal. 2019; 6(5):7702-12.
[3] Warraich J, Singh C, Thapa P. Blockchain-based intelligent monitored security system for detection of replication attack in the wireless healthcare network. European Journal of Engineering and Technology Research. 2021; 6(6):160-70.
[4] Boughaci D, Alkhawaldeh AA. Enhancing the security of financial transactions in Blockchain by using machine learning techniques: towards a sophisticated security tool for banking and finance. In 2020 first international conference of smart systems and emerging technologies (SMARTTECH) 2020 (pp. 110-5). IEEE.
[5] Wang Z, Luo N, Zhou P. GuardHealth: Blockchain empowered secure data management and graph convolutional network enabled anomaly detection in smart healthcare. Journal of Parallel and Distributed Computing. 2020; 142:1-12.
[6] Dhieb N, Ghazzai H, Besbes H, Massoud Y. A secure ai-driven architecture for automated insurance systems: fraud detection and risk measurement. IEEE Access. 2020; 8:58546-58.
[7] Nakamoto S. Bitcoin: a peer-to-peer electronic cash system. Decentralized Business Review. 2008.
[8] Moubarak J, Filiol E, Chamoun M. On Blockchain security and relevant attacks. In IEEE middle east and north Africa communications conference (MENACOMM) 2018 (pp. 1-6). IEEE.
[9] Wang W, Song J, Xu G, Li Y, Wang H, Su C. Contractward: automated vulnerability detection models for smart contracts. IEEE Transactions on Network Science and Engineering. 2020; 8(2):1133-44.
[10] Irwin AS, Turner AB. Illicit bitcoin transactions: challenges in getting to the who, what, when and where. Journal of Money Laundering Control. 2018.
[11] Chen L, Peng J, Liu Y, Li J, Xie F, Zheng Z. Phishing scams detection in ethereum transaction network. ACM Transactions on Internet Technology (TOIT). 2020; 21(1):1-16.
[12] Conti M, Kumar ES, Lal C, Ruj S. A survey on security and privacy issues of bitcoin. IEEE Communications Surveys & Tutorials. 2018; 20(4):3416-52.
[13] Chalapathy R, Chawla S. Deep learning for anomaly detection: a survey. arXiv preprint arXiv:1901.03407. 2019.
[14] Khan AA, Khan MM, Khan KM, Arshad J, Ahmad F. A blockchain-based decentralized machine learning framework for collaborative intrusion detection within UAVs. Computer Networks. 2021.
[15] Liu J, Zhao Z, Cui X, Wang Z, Liu Q. A novel approach for detecting browser-based silent miner. In IEEE third international conference on data science in cyberspace (DSC) 2018 (pp. 490-7). IEEE.
[16] Wu J, Yuan Q, Lin D, You W, Chen W, Chen C, et al. Who are the phishers? phishing scam detection on ethereum via network embedding. IEEE Transactions on Systems, Man, and Cybernetics: Systems. 2020; 52(2):1156-66.
[17] Mirtaheri M, Abu-el-haija S, Morstatter F, Ver SG, Galstyan A. Identifying and analyzing cryptocurrency manipulations in social media. IEEE Transactions on Computational Social Systems. 2021; 8(3):607-17.
[18] Alarab I, Prakoonwit S, Nacer MI. Comparative analysis using supervised learning methods for anti-money laundering in bitcoin. In proceedings of the 2020 5th international conference on machine learning technologies 2020 (pp. 11-7).
[19] Kok SH, Abdullah A, Jhanjhi NZ, Supramaniam M. Prevention of crypto-ransomware using a pre-encryption detection algorithm. Computers. 2019; 8(4):1-15.
[20] Jung E, Le TM, Gehani A, Ge Y. Data mining-based ethereum fraud detection. In 2019 IEEE international conference on blockchain (Blockchain) 2019 (pp. 266-73). IEEE.
[21] Iyer S, Thakur S, Dixit M, Katkam R, Agrawal A, Kazi F. Blockchain and anomaly detection based monitoring system for enforcing wastewater reuse. In 2019 10th international conference on computing, communication and networking technologies (ICCCNT) 2019 (pp. 1-7). IEEE.
[22] Li M, Zhang K, Liu J, Gong H, Zhang Z. Blockchain-based anomaly detection of electricity consumption in smart grids. Pattern Recognition Letters. 2020; 138:476-82.
[23] Bhutta MN, Khwaja AA, Nadeem A, Ahmad HF, Khan MK, Hanif MA, et al. A survey on blockchain technology: evolution, architecture and security. IEEE Access. 2021; 9:61048-73.
[24] Merkle RC. Protocols for public key cryptosystems. In secure communications and asymmetric cryptosystems 2019 (pp. 73-104). Routledge.
[25] Tschorsch F, Scheuermann B. Bitcoin and beyond: a technical survey on decentralized digital currencies. IEEE Communications Surveys & Tutorials. 2016; 18(3):2084-123.

[26] Casino F, Dasaklis TK, Patsakis C. A systematic literature review of blockchain-based applications: current status, classification and open issues. Telematics and Informatics. 2019; 36:55-81.
[27] Arya J, Kumar A, Singh AP, Mishra TK, Chong PH. Blockchain: basics, applications, challenges and opportunities.
[28] Schollmeier R. A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications. In proceedings first international conference on peer-to-peer computing 2001 (pp. 101-2). IEEE.
[29] Ozisik AP, Levine BN. An explanation of Nakamoto's analysis of double-spend attacks. arXiv preprint arXiv:1701.03977. 2017.
[30] Baliga A. Understanding blockchain consensus models. Persistent. 2017; 4(1).
[31] Zheng Z, Xie S, Dai H, Chen X, Wang H. An overview of blockchain technology: architecture, consensus, and future trends. In international congress on big data (BigData congress) 2017 (pp. 557-64). IEEE.
[32] Back A. Hashcash-a denial of service counter-measure. 2002.
[33] Nguyen CT, Hoang DT, Nguyen DN, Niyato D, Nguyen HT, Dutkiewicz E. Proof-of-stake consensus mechanisms for future blockchain networks: fundamentals, applications and opportunities. IEEE Access. 2019; 7:85727-45.
[34] Yao W, Ye J, Murimi R, Wang G. A survey on consortium blockchain consensus mechanisms. arXiv preprint arXiv:2102.12058. 2021.
[35] Cai W, Wang Z, Ernst JB, Hong Z, Feng C, Leung VC. Decentralized applications: the blockchain-empowered software system. IEEE Access. 2018; 6:53019-33.
[36] Ali MS, Vecchio M, Pincheira M, Dolui K, Antonelli F, Rehmani MH. Applications of blockchains in the internet of things: a comprehensive survey. IEEE Communications Surveys & Tutorials. 2018; 21(2):1676-717.
[37] Xie J, Tang H, Huang T, Yu FR, Xie R, Liu J, et al. A survey of blockchain technology applied to smart cities: research issues and challenges. IEEE Communications Surveys & Tutorials. 2019; 21(3):2794-830.
[38] Lu Y. Blockchain: a survey on functions, applications and open issues. Journal of Industrial Integration and Management. 2018; 3(4).
[39] Xinyi Y, Yi Z, He Y. Technical characteristics and model of blockchain. In international conference on communication software and networks (ICCSN) 2018 (pp. 562-6). IEEE.
[40] Lin IC, Liao TC. A survey of blockchain security issues and challenges. International Journal of Network Security. 2017; 19(5):653-9.
[41] Puthal D, Malik N, Mohanty SP, Kougianos E, Yang C. The blockchain as a decentralized security framework [future directions]. IEEE Consumer Electronics Magazine. 2018; 7(2):18-21.
[42] Yang R, Yu FR, Si P, Yang Z, Zhang Y. Integrated blockchain and edge computing systems: a survey, some research issues and challenges. IEEE Communications Surveys & Tutorials. 2019; 21(2):1508-32.
[43] Nakamoto WS. A next generation smart contract & decentralized application platform. Etherum. 2014.
[44] Androulaki E, Barger A, Bortnikov V, Cachin C, Christidis K, De CA, et al. Hyperledger fabric: a distributed operating system for permissioned blockchains. In proceedings of the thirteenth Eurosys conference 2018 (pp. 1-15).
[45] https://en.bitcoin.it/wiki/Genesis_block. Accessed 15 May 2022.
[46] Szabo N. Smart contracts: building blocks for digital markets. EXTROPY: The Journal of Transhumanist Thought,(16). 1996; 18(2).
[47] Zheng Z, Xie S, Dai HN, Chen W, Chen X, Weng J, et al. An overview on smart contracts: challenges, advances and platforms. Future Generation Computer Systems. 2020; 105:475-91.
[48] Chen W, Guo X, Chen Z, Zheng Z, Lu Y, Li Y. Honeypot contract risk warning on ethereum smart contracts. In international conference on joint cloud computing 2020 (pp. 1-8). IEEE.
[49] Tikhomirov S, Voskresenskaya E, Ivanitskiy I, Takhaviev R, Marchenko E, Alexandrov Y. Smartcheck: static analysis of ethereum smart contracts. In proceedings of the 1st international workshop on emerging trends in software engineering for blockchain 2018 (pp. 9-16).
[50] Chen W, Zheng Z, Ngai EC, Zheng P, Zhou Y. Exploiting blockchain data to detect smart ponzi schemes on ethereum. IEEE Access. 2019; 7:37575-86.
[51] Salah K, Rehman MH, Nizamuddin N, Al-fuqaha A. Blockchain for AI: review and open research challenges. IEEE Access. 2019; 7:10127-49.
[52] Chen F, Wan H, Cai H, Cheng G. Machine learning in/for blockchain: future and challenges. Canadian Journal of Statistics. 2021; 49(4):1364-82.
[53] Wang T. A unified analytical framework for trustable machine learning and automation running with blockchain. In 2018 IEEE international conference on big data (Big Data) 2018 (pp. 4974-83). IEEE.
[54] Azaria A, Ekblaw A, Vieira T, Lippman A. Medrec: using blockchain for medical data access and permission management. In 2016 2nd international conference on open and big data (OBD) 2016 (pp. 25-30). IEEE.
[55] Liu J, Peng S, Long C, Wei L, Liu Y, Tian Z. Blockchain for data science. In proceedings of the international conference on blockchain technology 2020 (pp. 24-8).
[56] Zhang D. Big data security and privacy protection. In 8th international conference on management and computer science (ICMCS 2018) 2018 (pp. 275-8). Atlantis Press.

[57] Rahouti M, Xiong K, Ghani N. Bitcoin concepts, threats, and machine-learning security solutions. IEEE Access. 2018; 6:67189-205.
[58] Nassif AB, Talib MA, Nasir Q, Dakalbab FM. Machine learning for anomaly detection: a systematic review. IEEE Access. 2021; 9:78658-700.
[59] Bulusu S, Kailkhura B, Li B, Varshney PK, Song D. Anomalous example detection in deep learning: a survey. IEEE Access. 2020; 8:132330-47.
[60] Bartoletti M, Pes B, Serusi S. Data mining for detecting bitcoin ponzi schemes. In 2018 crypto valley conference on blockchain technology (CVCBT) 2018 (pp. 75-84). IEEE.
[61] Ostapowicz M, Żbikowski K. Detecting fraudulent accounts on blockchain: a supervised approach. In international conference on web information systems engineering 2020 (pp. 18-31). Springer, Cham.
[62] Bamhdi AM, Abrar I, Masoodi F. An ensemble based approach for effective intrusion detection using majority voting. Telkomnika (Telecommunication Computing Electronics and Control). 2021; 19(2):664-71.
[63] Awang MK, Makhtar M, Udin N, Mansor NF. Improving customer churn classification with ensemble stacking method. International Journal of Advanced Computer Science and Applications. 2021; 12(11).
[64] Awang MK, Makhtar M, Mamat AR. Ensemble selection and combination based on cost function for UCI datasets. Journal of Theoretical and Applied Information Technology. 2021; 99(16): 4015-25.
[65] Rosly R, Makhtar M, Awang MK, Hassan H, Rose AN. Deep multi-classifier learning for medical data sets. International Journal of Engineering Trends and Technology (IJETT). 2020.
[66] Mendes-moreira J, Soares C, Jorge AM, Sousa JF. Ensemble approaches for regression: a survey. ACM Computing Surveys. 2012; 45(1):1-40.
[67] Monamo P, Marivate V, Twala B. Unsupervised learning for robust Bitcoin fraud detection. In 2016 information security for South Africa (ISSA) 2016 (pp. 129-34). IEEE.
[68] Kumar P, Gupta GP, Tripathi R. TP2SF: a trustworthy privacy-preserving secured framework for sustainable smart cities by leveraging blockchain and machine learning. Journal of Systems Architecture. 2021.
[69] Toyoda K, Ohtsuki T, Mathiopoulos PT. Identification of high yielding investment programs in bitcoin via transactions pattern analysis. In GLOBECOM 2017 (pp. 1-6). IEEE.
[70] Camino RD, State R, Montero L, Valtchev P. Finding suspicious activities in financial transactions and distributed ledgers. In international conference on data mining workshops (ICDMW) 2017 (pp. 787-96). IEEE.
[71] Monamo PM, Marivate V, Twala B. A multifaceted approach to bitcoin fraud detection: global and local outliers. In 2016 15th IEEE international conference on machine learning and applications (ICMLA) 2016 (pp. 188-94). IEEE.
[72] Harlev MA, Sun YH, Langenheldt KC, Mukkamala R, Vatrapu R. Breaking bad: de-anonymising entity types on the bitcoin blockchain using supervised machine learning. In proceedings of the 51st hawaii international conference on system sciences 2018.
[73] Chen W, Zheng Z, Cui J, Ngai E, Zheng P, Zhou Y. Detecting ponzi schemes on ethereum: towards healthier blockchain technology. In proceedings of the 2018 world wide web conference 2018 (pp. 1409-18).
[74] Toyoda K, Mathiopoulos PT, Ohtsuki T. A novel methodology for hyip operators' bitcoin addresses identification. IEEE Access. 2019; 7:74835-48.
[75] Zola F, Bruse JL, Eguimendia M, Galar M, Orduna UR. Bitcoin and cybersecurity: temporal dissection of blockchain data to unveil changes in entity behavioral patterns. Applied Sciences. 2019; 9(23):1-20.
[76] Kanemura K, Toyoda K, Ohtsuki T. Identification of darknet markets' bitcoin addresses by voting per-address classification results. In 2019 IEEE international conference on blockchain and cryptocurrency (ICBC) 2019 (pp. 154-8). IEEE.
[77] Lin YJ, Wu PW, Hsu CH, Tu IP, Liao SW. An evaluation of bitcoin address classification based on transaction history summarization. In international conference on blockchain and cryptocurrency (ICBC) 2019 (pp. 302-10). IEEE.
[78] Song J, He H, Lv Z, Su C, Xu G, Wang W. An efficient vulnerability detection model for ethereum smart contracts. In international conference on network and system security 2019 (pp. 433-42). Springer, Cham.
[79] Momeni P, Wang Y, Samavi R. Machine learning model for smart contracts security analysis. In 2019 17th international conference on privacy, security and trust (PST) 2019 (pp. 1-6). IEEE.
[80] I MJZ, Suárez-varela J, Barlet-ros P. Detecting cryptocurrency miners with NetFlow/IPFIX network measurements. In international symposium on measurements & networking (M&N) 2019 (pp. 1-6). IEEE.
[81] Victor F, Hagemann T. Cryptocurrency pump and dump schemes: quantification and detection. In international conference on data mining workshops (ICDMW) 2019 (pp. 244-51). IEEE.
[82] Baek H, Oh J, Kim CY, Lee K. A model for detecting cryptocurrency transactions with discernible purpose. In eleventh international conference on ubiquitous and future networks (ICUFN) 2019 (pp. 713-7). IEEE.
[83] La MM, Mei A, Sassi F, Stefa J. Pump and dumps in the bitcoin era: real time detection of cryptocurrency market manipulations. In 2020 29th international conference on computer communications and networks (ICCCN) 2020 (pp. 1-9). IEEE.
[84] Camino R, Torres CF, Baden M, State R. A data science approach for detecting honeypots in ethereum. In 2020 IEEE international conference on blockchain and cryptocurrency (ICBC) 2020 (pp. 1-9). IEEE.
[85] Liao Q, Gu Y, Liao J, Li W. Abnormal transaction detection of bitcoin network based on feature fusion. In 2020 IEEE 9th joint international information

technology and artificial intelligence conference (ITAIC) 2020 (pp. 542-9). IEEE.
[86] Ngo QD, Nguyen HT, Tran HA, Nguyen DH. IoT botnet detection based on the integration of static and dynamic vector features. In 2020 IEEE eighth international conference on communications and electronics (ICCE) 2021(pp. 540-5). IEEE.
[87] Cheema MA, Qureshi HK, Chrysostomou C, Lestas M. Utilizing blockchain for distributed machine learning based intrusion detection in internet of things. In 16th international conference on distributed computing in sensor systems (DCOSS) 2020 (pp. 429-35). IEEE.
[88] Fan S, Fu S, Xu H, Zhu C. Expose your mask: smart ponzi schemes detection on blockchain. In 2020 international joint conference on neural networks (IJCNN) 2020 (pp. 1-7). IEEE.
[89] Dashevskyi S, Zhauniarovich Y, Gadyatskaya O, Pilgun A, Ouhssain H. Dissecting android cryptocurrency miners. In proceedings of the tenth ACM conference on data and application security and Privacy 2020 (pp. 191-202).
[90] Huang D, Chen B, Li L, Ding Y. Anomaly detection for consortium blockchains based on machine learning classification algorithm. In international conference on computational data and social networks 2020 (pp. 307-18). Springer, Cham.
[91] Kumar N, Singh A, Handa A, Shukla SK. Detecting malicious accounts on the Ethereum blockchain with supervised learning. In international symposium on cyber security cryptography and machine learning 2020 (pp. 94-109). Springer, Cham.
[92] Alarab I, Prakoonwit S, Nacer MI. Competence of graph convolutional networks for anti-money laundering in bitcoin blockchain. In proceedings of the 2020 5th international conference on machine learning technologies 2020 (pp. 23-7).
[93] Lašas K, Kasputytė G, Užupytė R, Krilavičius T. Fraudulent behaviour identification in ethereum blockchain. In CEUR workshop proceedings [Electronic Resource]: IVUS 2020, information society and university studies, kaunas, lithuania: proceedings. Aachen: CEUR-WS 2020.
[94] Lorenz J, Silva MI, Aparício D, Ascensão JT, Bizarro P. Machine learning methods to detect money laundering in the bitcoin blockchain in the presence of label scarcity. In proceedings of the first ACM international conference on AI in Finance 2020 (pp. 1-8).
[95] Poursafaei F, Hamad GB, Zilic Z. Detecting malicious Ethereum entities via application of machine learning classification. In 2020 2nd conference on blockchain research & applications for innovative networks and services (BRAINS) 2020 (pp. 120-7). IEEE.
[96] Farrugia S, Ellul J, Azzopardi G. Detection of illicit accounts over the Ethereum blockchain. Expert Systems with Applications. 2020.
[97] Nerurkar P, Busnel Y, Ludinard R, Shah K, Bhirud S, Patel D. Detecting illicit entities in bitcoin using supervised learning of ensemble decision trees. In proceedings of the 2020 10th international conference on information communication and management 2020 (pp. 25-30).
[98] Badawi E, Jourdan GV, Bochmann G, Onut IV. An automatic detection and analysis of the bitcoin generator scam. In 2020 IEEE european symposium on security and privacy workshops (EuroS&PW) 2020 (pp. 407-16). IEEE.
[99] Bhowmik M, Chandana TS, Rudra B. Comparative study of machine learning algorithms for fraud detection in blockchain. In 2021 5th international conference on computing methodologies and communication (ICCMC) 2021 (pp. 539-41). IEEE.
[100] Wen H, Fang J, Wu J, Zheng Z. Transaction-based hidden strategies against general phishing detection framework on ethereum. In 2021 IEEE international symposium on circuits and systems (ISCAS) 2021 (pp. 1-5). IEEE.
[101] Gouda DK, Jolly S, Kapoor K. Design and validation of blockeval, a blockchain simulator. In 2021 international conference on communication systems & networks (COMSNETS) 2021 (pp. 281-9). IEEE.
[102] Balagolla EM, Fernando WP, Rathnayake RM, Wijesekera MJ, Senarathne AN, Abeywardhana KY. Credit card fraud prevention using blockchain. In 2021 6th international conference for convergence in technology (I2CT) 2021 (pp. 1-8). IEEE.
[103] Sousa JE, Oliveira VC, Valadares JA, Vieira AB, Bernardino HS, Villela SM, et al. Fighting under-price DoS attack in ethereum with machine learning techniques. SIGMETRICS Perform. Eval. Rev. 2021; 48(4):24-7.
[104] Al-e'mari S, Anbar M, Sanjalawe Y, Manickam S. A labeled transactions-based dataset on the ethereum network. In international conference on advances in cyber security 2020 (pp. 61-79). Springer, Singapore.
[105] Mittal R, Bhatia MP. Detection of suspicious or un-trusted users in crypto-currency financial trading applications. International Journal of Digital Crime and Forensics (IJDCF). 2021; 13(1):79-93.
[106] Agarwal R, Barve S, Shukla SK. Detecting malicious accounts in permissionless blockchains using temporal graph properties. Applied Network Science. 2021; 6(1):1-30.
[107] Yang X, Chen Y, Qian X, Li T, Lv X. BCEAD: a blockchain-empowered ensemble anomaly detection for wireless sensor network via isolation forest. Security and Communication Networks. 2021.
[108] Chen B, Wei F, Gu C. Bitcoin theft detection based on supervised machine learning algorithms. Security and Communication Networks. 2021.
[109] Lacruz F, Saniie J. Applications of machine learning in fintech credit card fraud detection. In international conference on electro information technology (EIT) 2021 (pp. 1-6). IEEE.
[110] Ibrahim RF, Elian AM, Ababneh M. Illicit account detection in the ethereum blockchain using machine learning. In 2021 international conference on information technology (ICIT) 2021 (pp. 488-93). IEEE.

[111]Liu X, Jiang F, Zhang R. A new social user anomaly behavior detection system based on blockchain and smart contract. In 2020 IEEE international conference on networking, sensing and control (ICNSC) 2020 (pp. 1-5). IEEE.
[112]Shahin R, Sabri KE. A secure IoT framework based on blockchain and machine learning. International Journal of Computing and Digital System. 2021.
[113]Jatoth C, Jain R, Fiore U, Chatharasupalli S. Improved classification of blockchain transactions using feature engineering and ensemble learning. Future Internet. 2021; 14(1):1-12.
[114]Brinckman E, Kuehlkamp A, Nabrzyski J, Taylor IJ. Techniques and applications for crawling, ingesting and analyzing blockchain data. In international conference on information and communication technology convergence (ICTC) 2019 (pp. 717-22). IEEE.
[115]Sayadi S, Rejeb SB, Choukair Z. Anomaly detection model over blockchain electronic transactions. In 2019 15th international wireless communications & mobile computing conference (IWCMC) 2019 (pp. 895-900). IEEE.
[116]Yang L, Dong X, Xing S, Zheng J, Gu X, Song X. An abnormal transaction detection mechanim on bitcoin. In international conference on networking and network applications (NaNA) 2019 (pp. 452-7). IEEE.
[117]Keshk M, Turnbull B, Moustafa N, Vatsalan D, Choo KK. A privacy-preserving-framework-based blockchain and deep learning for protecting smart power networks. IEEE Transactions on Industrial Informatics. 2019; 16(8):5110-8.
[118]Sureshbhai PN, Bhattacharya P, Tanwar S. KaRuNa: a blockchain-based sentiment analysis framework for fraud cryptocurrency schemes. In international conference on communications workshops (ICC Workshops) 2020 (pp. 1-6). IEEE.
[119]Bhargavi MS, Katti SM, Shilpa M, Kulkarni VP, Prasad S. Transactional data analytics for inferring behavioural traits in ethereum blockchain network. In 2020 IEEE 16th international conference on intelligent computer communication and processing (ICCP) 2020 (pp. 485-90). IEEE.
[120]Fan S, Fu S, Xu H, Cheng X. Al-SPSD: anti-leakage smart ponzi schemes detection in blockchain. Information Processing & Management. 2021; 58(4).
[121]Yilmaz I, Kapoor K, Siraj A, Abouyoussef M. Privacy protection of grid users data with blockchain and adversarial machine learning. In proceedings of the 2021 ACM workshop on secure and trustworthy cyber-physical systems 2021 (pp. 33-8).
[122]Hu T, Liu X, Chen T, Zhang X, Huang X, Niu W, et al. Transaction-based classification and detection approach for Ethereum smart contract. Information Processing & Management. 2021; 58(2).
[123]Zarpelão BB, Miani RS, Rajarajan M. Detection of bitcoin-based botnets using a one-class classifier. In IFIP international conference on information security theory and practice 2018 (pp. 174-89). Springer, Cham.
[124]Bach LM, Mihaljevic B, Zagar M. Comparative analysis of blockchain consensus algorithms. In 2018 41st international convention on information and communication technology, electronics and microelectronics (MIPRO) 2018 (pp. 1545-50). IEEE.

Sabri Hisham earned his Bachelor's degree in computer science (Industrial Computing) from Universiti Teknologi Malaysia (UTM) in 2001 and his Master's degree in software engineering from Universiti Malaysia Pahang (UMP) in 2014. He is currently a PhD student at Universiti Sultan Zainal Abidin's Department of Computer Science in the Faculty of Computing and Informatics at Terengganu, Malaysia. He is also the Head of Infostructure at Universiti Malaysia Pahang's Information and Technology Department. He is also a Blockchain Solidity Smart Contract and Ethereum Expert and certified professional. Blockchain, Bitcoin, ML, IoT, Mobile App, Web Application, SCADA and Telemetry System are among his current research interests.
Email: sabrihisham@ump.edu.my

Prof. Mokhairi Makhtar earned his PhD in 2012 from the University of Bradford in the United Kingdom. He is currently a Professor at Universiti Sultan Zainal Abidin in Terengganu, Malaysia, in the Department of Computer Science. Machine Learning, Ensemble Method, Data Mining, Soft Computing, Timetabling and Optimisation, Natural Language Processing, E-Learning, and Deep Learning are some of his current research interests.
Email: mokhairi@unisza.edu.my

Mr. Azwa Abdul Aziz earned a degree in computer science in 2002 from Malaysia's Universiti Teknologi Mara. He completed his studies at the Bachelor's level and graduated from Universiti Teknologi Mara, Malaysia, in 2004. Then, in 2010, he earned a master's degree in computer science from Malaysia's University of Malaysia Terengganu. He is recently a lecturer at the Department of Computer Science at Sultan Zainal Abidin University in Terengganu, Malaysia. Included in his research interests are Big Data Analytics, Text Mining, Business Intelligence, and Machine Learning.
Email: azwaaziz@ unisza.edu.my

Appendix I

S. No.	Abbreviations	Descriptions
1	AI	Artificial Intelligence
2	API	Application Programming Interface
3	AUC	Area Under the ROC Curve
4	BT	Blockchain Technology
5	CA	Centralized Authority
6	DApps	Decentralized Applications
7	DeFi	Decentralized Finance (DeFi)
8	DL	Deep Learning
9	DLT	Distributed Ledger Technology
10	DoS	Denial of Service
11	DPOS	Delegated proof of stake
12	DT	Decision Trees
13	ELM	Ensemble Learning Methods
14	EM	Ensemble Method
15	EVM	Ethereum Virtual Machine
16	FN	False Negative
17	FNR	False Negative Rate
18	FP	False Positive
19	FPR	False Positive Rate
20	GCN	Graph Convolutional Network
21	IF	Isolation forest
22	IoT	Internet Of Things
23	KNN	K-Nearest Neighbour
24	LR	Logistic Regression
25	LSTM	Long short-term memory
26	ML	Machine Learning
27	MLOps	Machine Learning Operations
28	MSP	Membership Service Provider
29	NB	Naïve Bayes
30	NFT	Non-fungible token
31	NN	Neural Network
32	OCVM	One-Class Support Vector Machine
33	P2P	Peer to Peer
33	PBFT	Practical Byzantine Fault Tolerance
34	PCA	Principal Component Analysis
35	PoA	Proof of Authority
36	PRISMA	Reporting Items for Systematic Reviews and Meta-Analysis
37	PoS	Proof of Stake
38	PoW	Proof of Work
39	RF	Random Forest
40	RL	Reinforcement Learning
41	SD	Standard Deviation
42	SCM	Supply Chain Management
43	SVM	Support Vector Machines
44	TNR	True Negative Rate
45	TPR	True Positive Rate
46	UAV	Unmanned Aerial Vehicle
47	WoS	Web of Science