Research Article/Araştırma Makalesi

GAZİ
JOURNAL OF ENGINEERING SCIENCES

Storage of Developed Distributed Machine Learning Models on Blockchain

Remzi Gürfidan*[a], Mevlüt Ersoy[b]

Submitted: 10.11.2022 Revised: 21.06.2023 Accepted: 09.08.2023 doi:10.30855/gmbd.0705068

ABSTRACT

In this study, a dataset containing air pollution information collected from different locations is used. This dataset contains air pollution measurement data and result information for 25 different regions of Seoul. The values obtained from similar regions were filtered from the whole dataset and a dataset specific to these regions was created. As the next step, the model was trained with the values obtained in their regions and the model was registered. This model shows very successful results when tested with the data obtained in their own regions. However, when tested with data obtained from different regions and not similar to these regions, the classification success of the model was very low. The aim of this study is to show that more successful results can be obtained with the distributed learning method, which can be a solution to this problem situation, compared to classical machine learning. In addition, security problems during model merging and clustering should not be ignored in Distributed Learning. For this reason, a robust blockchain-based approach is proposed against malicious attacks during model distribution.

Keywords: Distributed learning, machine learning, blockchain, secure storage of model

[a,*] Isparta University of Applied Sciences,
Yalvac Vocational School of Technical Sciences;
Computer Programming
32100 - Isparta, Türkiye
Orcid: 0000-0002-4899-2219
e mail: remzigurfidan@isparta.edu.tr

[b] Süleyman Demirel University,
Engineering Faculty,
Computer Engineering
32500 - Isparta, Türkiye
Orcid: 0000-0003-2963-7729
e mail: mevlutersoy@sdu.edu.tr

*Corresponding author:
remzigurfidan@isparta.edu.tr

Geliştirilen Dağıtılmış Makine Öğrenimi Modellerinin Blok Zincirde Depolanması

ÖZ

Bu çalışmada farklı lokasyonlardan toplanan hava kirliliği bilgilerini içeren bir veri seti kullanılmıştır. Bu veri seti Seul şehrinin 25 farklı ilçesine ait hava kirliliği ölçüm bilgileri ve sonuç bilgilerini içermektedir. Benzer bölgelerden elde edilen değerler tüm veri seti içerisinden filtrelenerek bu bölgelere has veri seti oluşturulmuştur. Sonraki işlem basamağı olarak, kendi bölgelerinde elde edilen değerler ile eğitilip model kaydedilmiştir. Bu model kendi alanlarında elde edilen veriler ile test edildiğinde oldukça başarılı sonuçlar ortaya koymaktadır. Fakat farklı bölgelerden elde edilen ve bu bölgeler ile benzeşmeyen veriler ile test edildiğinde model sınıflama başarısı oldukça düşük kalmıştır. Çalışmanın amacı bu problem durumuna çözüm olabilecek olan dağıtık öğrenme yöntemi ile klasik makine öğrenmesine kıyasla daha başarılı sonuçlar elde edilebileceğini göstermektir. Bunun yanında Dağıtık Öğrenme kısmında model birleştirme ve toplama esnasındaki güvenlik sorunlarının da göz ardı edilmemelidir. Bu sebeple model dağıtımı sırasında kötü niyetli saldırılara karşı blockchain tabanlı sağlam bir yaklaşım önerilmiştir.

Anahtar Kelimeler: Dağıtık öğrenme, makine öğrenmesi, blok zinciri, modelin güvenli depolanması

To cite this article: R. Gürfidan and M. Ersoy, "Storage of Developed Machine Learning Model in Disributed Machine Learning on the Blockchain," *Gazi Journal of Engineering Sciences,* vol. 9, no. 2, pp. 248-263, 2023. doi:10.30855/gmbd.0705068

1. Introduction

The latest design is the trend of design students design used all over the world. In an enterprise that does not want to be behind the times, people like a production facility and those who prefer people's plans. With the development attacks of the internet of things (IoT) technology, it has made great innovations such as smart home systems and smart city planning. In addition, the data emerging in these systems or the protection and security of these systems can be mentioned. The emergence of these vulnerabilities depends on cyber security and the path it follows. The latest system, which is the brightest of its content, is the Blockchain technology, which uses the ledger structure, without meeting the manipulation of data [1]. This emerging technology with crypto flamboyance has successfully demonstrated digital experience and experience. The smart intelligences for this system include the trainings of the model. The model that performs learning with this orientation is related to its ability to predict classification in data samples depending on the preferred options [2].

The sources of motivation for this study can be listed as follows.

- In this study, a method has been developed that can ensure the security of artificial intelligence applications performed by ignoring the security parameter.
- More successful results were obtained with the distributed learning method compared to classical machine learning.
- The data privacy problem related to the data used in artificial intelligence model training has been prevented.
- There are security problems during model aggregation and aggregation in the Distributed Learning section. It is necessary to propose a robust approach against malicious attacks during model collection [3]. Storing the weights presented in this study with the blockchain provides a solution to this problem by preventing manipulation.

In this research study, distributed learning, weight files of artificial intelligence model, blockchain technology, Hyperledger Fabric infrastructure are explained as the subtitles of the introduction. In the second part, studies in which blockchain and distributed learning are used together are examined. In the third chapter, methods and techniques are explained about why weight files should be stored in a distributed learning model and how they can be stored with smart contracts. In the fourth part of the study, the performance measurements of the process performed were carried out and the findings were revealed. In the last part of the study, the values obtained in the light of the findings were discussed and the results and future study objectives were listed.

2. Technology and Methods Used in the Study

In this section, the frameworks that will be used in the structure to be realized in this research study are examined. Distributed learning, blockchain and Hyperledger fabric framework, model weights in machine learning are discussed.

2.1. Distributed Learning

Distributed learning is a machine learning technique that trains algorithms on a server or servers, in which data from more than one independent device is stored in an anonymized form.

In distributed learning, each of the various devices that are part of the learning network has a copy of the machine learning model. Different devices train the copy of the machine learning model using the device's local data, and then the parameters in different machine learning models collect the parameters and send them to a master device or server that updates the overall machine learning model accordingly. This training process can be repeated until the desired level of accuracy is achieved. In short, the main idea behind federated learning is that none of the training data is transmitted between devices or parties, only updates to the model are shared [4]. Distributed learning is the darling of an innovative generation of smart networks where smart devices such as mobile hardware, robots and sensors exchange information with each other to collaboratively train machine learning models without offloading raw data information to a central entity for centralized processing. The distributed learning paradigm can reduce the load on centralized processors by using the computational or communication capacity of individual agents and contributes to the protection of users' data privacy

[5]. In distributed learning systems, it is a collaborative learning method that protects the privacy of users' original data and does not share it with third parties. Recently, it has been attracting more and more attention and use. The combination of federation and machine learning is becoming a hot new research topic. All these reasons have been a serious motivation for the choice of distributed learning in this study [6].

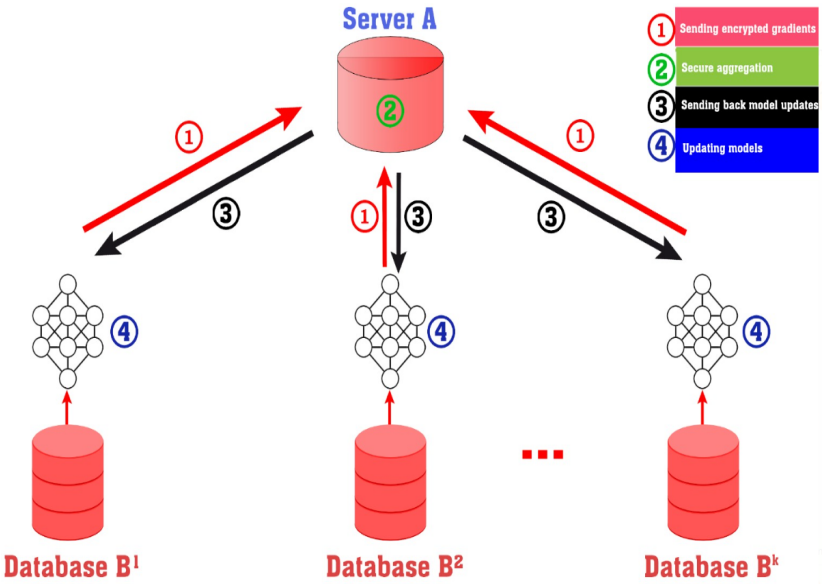


Figure 1. Distributed learning system model

For the appropriate usage scenario of this technique, it is necessary to pay attention to the motivation of choice in a sample application. In the said scenario, it is desired to have more instructive materials for the machine learning model by combining the data obtained by similar purpose systems in different locations. It is known that the more successful learning condition of machine learning models and the increase in the number of heterogeneous data is the most important issue. However, at this very moment, the laws on the protection of personal data and the concepts of data privacy have come into play. In the General Data Protection Regulation (GDPR) criteria, important aspects regarding data privacy and the protection of personal data are guaranteed. According to the GDPR criteria, the data of both parties cannot be transferred or combined without the consent of their respective users [7], [8]. Since the model weights are updated and transferred instead of transferring the data in the federated learning technique, the data is anonymized, thus contributing to the solution of the data privacy problem [3].

2.2. Weight Files in Machine Learning Models

In the developed machine learning models, the progress of the model can be recorded either during the training of the model or after the training is completed. One of the purposes of this process is to ensure that the model can continue its training from where it left off, considering that the training will take a long time. Another purpose is to facilitate the sharing of the saved model. In addition, the time and effort cost of developing machine learning models is very high. Turning off the hardware where the training is performed causes all weights of the model to be lost while the memory is freed. Saving models is critical to optimize reusability to make the most of your time. To better understand the importance of this recorded data file, it is necessary to know the technical infrastructure of how the model developed with machine learning performs the prediction or classification process. The data presented to the machine learning model to be developed are usually labeled as input parameters and output parameter or parameters. Thanks to this labeling, the model that will develop learning discovers meaningful ways to reach the output parameter from the input parameters, depending on the algorithm to be used, and connects these ways to a mathematical model within itself. In this fully created mathematical discovery model, the data we call weight is revealed. These weight data have the value that determines which input parameter will be more dominant and effective in accessing the output parameter. The estimation or classification success of the model that has completed its training is directly related to the correct determination of the created weights and is primarily effective. Within the weight files, there are only meaningful numerical data for the model, completely independent of

the data used in the training. For weight files, it would be correct to say an anonymized dataset in which all the experience that can be obtained from the original data is summarized and packaged.

Research has been carried out on different methods and techniques on saving the weights of the prepared machine learning models and saving the model. This feature, which enables a ready-made model to be used faster and more effectively by saving the time spent in the training process, provides a very time-saving benefit. When the state of the literature in this field is analyzed, a review study on saving, loading and recalling machine learning models [9], research on different methods for saving and loading machine learning models consistently [10] and different methods for saving weights [11] are presented.

2.3. Blockchain and Hyperledger Fabric Framework

Blockchain is a chain structure consisting of interconnected blocks. Blockchain can be thought of as a distributed database. The data on the blockchain system is stored in a ledger technology that can be shared and cannot be changed [12]. Advantages of using the blockchain system [13], [14];

- Transactions to be made can be carried out cheaply and quickly, without being tied to a center or person.
- Manipulation and fraud are prevented as all transactions performed on the distributed ledger technology can be checked and verified.
- Thanks to its distributed data storage capability, it has a resilient and strong structure against the attacks it will be exposed to.
- Since it is based on the mathematical model, there is almost no margin for error.
- It provides transparency since all members will have a peer copy of the records stored in the distributed ledger technology.
- Thanks to smart contracts, transaction security and protocols are realized securely and quickly without the possibility of denial or denial.

The Hyperledger project is an open source, permissioned, distributed ledger established by the Linux Foundation [15]. This project is divided into thirteen projects. These projects have been introduced as Fabric, Sawtooth, Indy, Burrow, Iroha, Grid, Aries, Caliper, Cello, Composer, Explorer, Quilt, Ursa [16], [17].

3. Related Work

Blockchain technology and machine learning technology separately are interesting fields of study for many researchers. When the literature is examined, the number of studies in these subject areas is quite high. The use of these two workspaces together usually includes the keywords security and communication.

Quantitatively large amounts of data and large edge devices present significant security, privacy, service delivery, and network management challenges. The joint evaluation of blockchain and machine learning (ML) provides significant benefits in this regard [18]. Realizing this problem area and lack, Liu et al. carried out research on the integration of blockchain and machine learning. Among the results of their studies, they concluded that it would be efficient to combine this field with larger fields such as IoT (Internet of Things), Big Data and Edge Computing [19]. Adhikari and Song proposed a joint Blockchain-based reputation system to minimize double spend and linear regression-based machine learning to reduce system-induced latency in the field of wireless virtualization. The results obtained in the study showed that the proposed approach outperforms other approaches in terms of expected latency as well as expected utility for wireless virtualization [20]. Shinde and his team propose a technological solution to eliminate the insecurities in road construction processes in India. This recommendation system envisages that the main challenges of traditional road construction contracting systems, which rely heavily on limited available data and assumptions regarding raw materials and traffic behavior, will be effectively resolved. An artificial intelligence model to be developed will perform the time and cost estimation of road construction operations. Blockchain structure, on the other hand, will ensure the integrity of the work-time contract and reconciliation between the stakeholders [21].

Table 1. Detailed and comparative analysis table of existing studies

Work / Year	Type of Algorithm	Type of Blockchain	The Aim of The Study	Contribution of The Study
[22] (2022)	SVM, KNN	No Framework, Self-made blockchain	It provides a framework for analyzing and detecting interference with real-time network data, identifying multiple potentially harmful interference patterns.	SVM outperformed all trained models, achieving 99.05% DA and 0.95% MCR (powered by blockchain technology) to overcome security threats to networking and transaction data
[23] (2018)	Semi-supervised learning	No Framework, Self-made blockchain	Proposes to create a reliable machine learning system using blockchain technology that can store data in a permanent and immutable way.	It establishes a link between machine learning technology and blockchain technology. It proposes a unified analytical framework for reliable machine learning using blockchain technology.
[24] (2020)	Bayes Network based Filtering scheme	No Framework, Self-made blockchain	She proposed a scalable and efficient method to protect the allowed blockchain in the SDN domain from DDoS attack.	It has been seen that the created machine learning model can detect attacks with high accuracy and is safe to protect blockchain applications from DNS Amplification attacks.
[25] (2019)	Similarity Learning	Ethereum	Desired to develop a smart contract-based data trading mode solution and framework using blockchain and machine learning	Authentication in the trade performed, prevention of off-chain verification, resolution of disputes during data trade, automatic payment of the balance, and the creation of dishonest behavior penalties are provided.
[26] (2021)	Deep Learning	No Framework, Self-made blockchain	A model based on integrating edge computing, blockchain technology and machine learning is proposed to support the design of the manufacturing system.	The assignment problem of the system is formulated based on the optimization model.
[27] (2019)	Supervised & Un supervised Learning	No Framework, Self-made blockchain	It is desired to propose a secure model that will be beneficial for institutions and individuals operating in the field of health and focuses on patients' access to data from anywhere.	Data in the field of health are shown by ensuring data privacy for doctors and patient relatives. Patient data is used for disease classification with machine learning
[28] (2021)	Deep Extreme Learning Machine	No Framework, Self-made blockchain	It is aimed to propose a blockchain-based machine learning approach for network security in smart home technologies.	A lightweight yet effective solution is proposed for intrusion detection and prediction in smart home networks.
[29] (2020)	SVM	Hyperledger Fabric	The study plans to propose a smart fitness service with the help of an advanced smart contract on blockchain networks.	A real-time inference engine has been developed. The system also predicts the future diet plan and exercise plan for the patient and guarantees its safety
[30] (2021)	-	Hyperledger Fabric	The study proposed a Hyperledger Fabric Architecture for the secure and efficient management of EHRs systems.	We created a trusted and transparent encyclopedia of patient data in the study, ensuring controlled data access and integrity for the widest stakeholders of the EHR system.
[31] (2021)	-	Hyperledger Fabric	The aim is to increase security in IoT-based health monitoring systems, achieve better storage utilization due to limited storage space, decentralized accountability and eliminate single point of failure.	In this study, a security enhanced model for Hyperledger Fabric based IoT based health monitoring systems is proposed. Alternative solutions to the limitations are presented.
[32] (2020)	-	Hyperledger Fabric	The aim of this paper is to propose an efficient decentralized data integrity checking scheme based on Hyperledger Fabric to address the security and scalability challenges in legacy TPA-based data integrity checking schemes.	It is the first attempt to investigate this area. They formulate the TPA selection model, and two selection algorithms are prepared under complete and incomplete information, which can quickly select TPAs that satisfy users
This Work	Decision Tree	Hyperledger Fabric	Secure storage of the model developed in the distributed learning model and the weights of the model	Creating a secure distributed learning model, increasing machine learning accuracy values by creating a global model

When the academic studies carried out before this study are examined, Table 1 shows the aims of the studies and the innovations they have contributed to the literature. although different algorithms are preferred in many similar studies, studies have generally been carried out on security and data privacy. in this study, in addition to data privacy and security, data anonymization fiction has an important

place. In [R6], a process was followed to create an encyclopedia by storing raw data securely and transparently and sharing it with healthcare organizations in a controlled manner. In [R7], raw data collected from IoT devices are kept in a local chain. Afterwards, the data in the local chain is sent to a global cloud chain with gateways. the aim here is to eliminate the single point error that may occur in the two chains that are kept for verification purposes from the other. in addition, performance measurements in this study do not contain any information about the scalability of the infrastructure created by taking place in future study targets. Unlike [R7], in this work, the data itself is not stored, but the weights kept in the machine learning model. this is perfect for data anonymization in distributed learning machine learning. Since it does not contain meaningful expressions like the original data, there is no threat to personal data. The work carried out in [R8] involves quite extensive and detailed processes. In this study, a new algorithm is proposed to increase efficiency and scalability. Since the TPA process is carried out in a way to include a third verification, it has exceeded the standard security of Hyperledger Fabric. Comparing this study with [R8], it can be said that it has higher level security protocols. In [R8], it is seen that the existing verification processes add additional processing time to the blockchain. Although we have to accept its superiority in terms of security, it is important to establish the optimum performance and security relationship.

In this study, Hyperledger Fabric developed by IBM was used as the blockchain infrastructure. The reason for this choice is to harness the power of a truly distributed storage system that has been proven and tested to be reliable. In addition, the Desicion Tree algorithm, whose classification success is known, is preferred. When the aims of the study and its contribution to the literature are examined, it is seen that it has brought an innovation to the literature. In the next section you will find a detailed section explaining how we did the current work.

4. Proposed Model

In this section, the methods and techniques related to the blockchain structure to be used, its installation, operation, why the weight files should be stored in the distributed learning model and how they can be stored with smart contracts are explained. In the proposed model, Hyperledger Fabric is used as the blockchain infrastructure. To store the data in the weight files created by machine learning, the rough code of the smart contract created is given.

4.1. Hyperledger Fabric Setup

The Hyperledger Fabric will run on an operating system. As the operating system, the Linux-based Ubuntu 22.04.1 version was preferred. After the installation process is completed, the prerequisites for the Hyperledger Fabric infrastructure must be installed. Prerequisites are Curl, Git, Docker and Docker-Compose technologies. Installations from the Ubuntu terminal can be easily performed with codes. When the codes given below are applied in order, the prerequisites for Hyperledger Fabric installation will be completed.

- sudo apt install curl
- sudo apt install git
- sudo apt install docker.io
- sudo apt install docker-compose
- sudo curl -sSL https://bit.ly/2ysbOFE | sudo bash -s

Figure 2. Installation codes of Hyperledger Fabric Dependencies

After the pre-installation process, the network structure to be worked on is opened. Afterwards, the installation of channels through which communication will be carried out on this network is set up. Then, since the coding on the existing structure will be carried out with the JavaScript language, the necessary language settings are made.

- cd fabric-samples/test-network
- sudo ./network.sh up createChannel
- sudo ./network.sh deployCC -ccn basic -ccp ../asset-transfer-basic/chaincode-javascript -ccl JavaScript
- sudo su
- export PATH=${PWD}/../bin:$PATH
- export FABRIC_CFG_PATH=$PWD/../config/

Figure 3. Installation codes of Hyperledger Fabric Channels

After these processes are completed, organization setups should be done. The environmental variables of the first organization are carried out by following the items below in order.

- export CORE_PEER_TLS_ENABLED=true
- export CORE_PEER_LOCALMSPID="Org1MSP"
- export CORE_PEER_TLS_ROOTCERT_FILE=${PWD}/organizations/peerOrganizations/org1.example.com/peers/peer0.org1.example.com/tls/ca.crt
- export CORE_PEER_MSPCONFIGPATH=${PWD}/organizations/peerOrganizations/org1.example.com/users/Admin@org1.example.com/msp
- export CORE_PEER_ADDRESS=localhost:7051

Figure 4. Configuration codes of Hyperledger Fabric

The necessary operational processes and ways for the organization are introduced.

- peer chaincode invoke -o localhost:7050 --ordererTLSHostnameOverride orderer.example.com --tls –cafile ${PWD}/organizations/ordererOrganizations/example.com/orderers/orderer.example.com/msp/tlscacerts/tlsca.example.com-cert.pem -C mychannel -n basic --peerAddresses localhost:7051 --tlsRootCertFiles ${PWD}/organizations/peerOrganizations/org1.example.com/peers/peer0.org1.example.com/tls/ca.crt --peerAddresses localhost:9051 --tlsRootCertFiles
- ${PWD}/organizations/peerOrganizations/org2.example.com/peers/peer0.org2.example.com/tls/ca.crt -c '{"function":"InitLedger","Args":[]}'
- peer chaincode query -C mychannel -n basic -c '{"Args":["GetAllAssets"]}'

Figure 5. Introduction of operational processes and pathways required for Hyperledger Fabric Organization

The same processes are carried out for the second organization. After these processes are completed, wallet definitions are made.

- export CORE_PEER_TLS_ENABLED=true
- export CORE_PEER_LOCALMSPID="Org2MSP"
- export CORE_PEER_TLS_ROOTCERT_FILE=${PWD}/organizations/peerOrganizations/org2.example.com/peers/peer0.org2.example.com/tls/ca.crt
- export CORE_PEER_MSPCONFIGPATH=${PWD}/organizations/peerOrganizations/org2.example.com/users/Admin@org2.example.com/msp
- export CORE_PEER_ADDRESS=localhost:9051
- peer chaincode query -C mychannel -n basic -c '{"Args":["ReadAsset","asset6"]}'
- peer chaincode invoke -o localhost:7050 --ordererTLSHostnameOverride orderer.example.com --tls –cafile ${PWD}/organizations/ordererOrganizations/example.com/orderers/orderer.example.com/msp/tlscacerts/tlsca.example.com-cert.pem -C mychannel -n basic --peerAddresses localhost:7051 –tlsRootCertFiles ${PWD}/organizations/peerOrganizations/org1.example.com/peers/peer0.org1.example.com/tls/ca.crt --peerAddresses localhost:9051 --tlsRootCertFiles ${PWD}/organizations/peerOrganizations/org2.example.com/peers/peer0.org2.example.com/tls/ca.crt -c '{"function":"CreateAsset1","Args":["asset7","192.168.1.1","45","Remzi","100"]}'

Figure 6. Wallet definitions for Hyperledger Fabric Organization

With the following installation codes, Nodejs and npm installations are completed and Hyperledger Fabric is ready to work.

```
• apt install nodejs
• apt install npm
```

Figure 7. Setup Nodejs and npm

There is a need to define an authority that will perform its work on the Hyperledger Fabric infrastructure. This operation can be performed with the code written in the bottom line. With the next line of code, the server is lifted.

```
• node setupadminuser.js
• node server.js
```

Figure 8. Define admin for blockchain system

If you want to reconfigure an established and functioning infrastructure or wallet, you will first need to reset the existing settings. This can get complicated. For this reason, it may be wiser to delete and reinstall. The lines of code listed below first stop the active network and then clear the wallet by deleting it.

```
• cd ../../test-network
• ./network.sh down
• cd ../asset-transfer-basic/application-smartcontract
• rm -r wallet
```

Figure 9. Delete wallet code for Hyperledger Fabric Organization

4.2. Running the Hyperledger Fabric Framework

After the installation is completed, there will be no problem in the operations you will do when you start the system. However, when the system is stopped or closed, some processes need to be repeated from the beginning due to security measures. These sequences of operations are given below, in order, in item. The process, which starts with the authorization received from the terminal, continues with the opening of the relevant files, channel installations, activation of the smart contract, identification of the authorized person, and re-starting the server.

```
• sudo su
• cd fabric-samples/test-network
• ./network.sh up createChannel -ca
• ./network.sh deployCC -ccn basic -ccp ../asset-transfer-basic/chaincode-smartcontract/ -ccl javascript
• cd ../asset-transfer-basic/application-smartcontract
• node setupadminuser.js
• node server.js
```

Figure 10. Start code for Hyperledger Fabric Organization

After these processes are completed, the wallet can be started to listen with the Ip address given in the bash files and the settings made in the contract. The records in the wallet can be tracked.

4.3. Smart Contract Structure

A smart contract is executable code running on an engineered blockchain designed to facilitate the terms of an agreement, execute, and execute transactions between parties in mutual trust issues. It can be thought of as a system that allows digital asset transactions to all or some of the relevant parties after the predefined rules in smart contracts are fulfilled by the parties. Compared to traditional contracts, smart contracts do not allow a trusted third party to operate, resulting in low transaction costs. Prepared smart contracts will be assigned to a unique 20-byte address. Once the smart contract is paired with a blockchain, the code assigned to the contract cannot be changed. For a contract to be executed, it is sufficient for users to send the transaction they want to perform to the address of the contract. The requested transaction is then evaluated by each consensus node in the network to arrive

at a consensus. The status of the contract will then be updated according to the result [30]. The consensus algorithms used, along with variables such as throughput, latency, node scalability, largely determine the performance of the distributed system for blockchains.

Algorithm 1 Smart Contract Pseudo Code

```
function initLedger ()
    config LedgerStandarts ()
function CreateAsset (ctx, params) ←obj
    if exist (ctx) = = true then
        return error
    else
        return (obj ⊃ [params])
function GetAllAsset (ctx, id)
    const allResult = []
    while! result. done then
        allResult.Push → Key: result.value.key, Record: record
        result ← await. iterator. next ()
        return allResult end
```

Figure 11. Smart contract pseudo code

In the prepared smart contract, the initLedger method is run to perform the initial settings that must be made at the beginning of the ledger. Before writing new data to the notebook, it is checked whether there is data with the same id. When a positive response is received, a new object is created and the ledger registration process is started, and a new record is returned at the end of the transaction. The GetAllAsset method can be run to read the records. After the necessary permissions are checked, the data recorded in the ledger can be read and listed with the help of an iterator. Figure 11 shows the pseudocode of the created smart contract.

The Hyperledger Fabric infrastructure uses Raft, a consensus mechanism based on Practical Byzantine Fault Tolerance (PBFT) as a consensus algorithm. Raft is a consensus algorithm that works by participants in the network electing a leader and approving blocks proposed by the leader. Raft is a consensus mechanism that requires most participants in the network to agree. In this way, transactions continue even when there are problems between some participants in the network. Raft focuses on delivering a higher transaction speed and lower latency.

4.4. Decision Tree Model

Decision trees are an easy and interpretable algorithm from supervised learning algorithms. Regression and classification operations can be performed with the decision tree algorithm. The algorithm works in the inductive structure. In a developed decision tree algorithm, it is very effective in defining a method that resembles a tree and its branches to pass the data set and to achieve the expected results from the data set. The branches in the tree structure are based on the values of the input parameters and are divided depending on the value of a particular property.

The information acquisition attribute minimizes the information required to classify the data points into the relevant segments, thus ensuring the least randomness in that segment. Knowledge gain is calculated with the formulas shown in Equation (1) and Equation (2).

$$Info\ (D) = -\sum_{1}^{m} p_i log_2(p_i) \tag{1}$$

$$Info\ _A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|}\, x\ Info_{(D_j)} \tag{2}$$

p_i is the probability that a random tuple in dataset D belongs to a class. Info(D) is the average amount of information required to define the class, category of a data point in D. m and is the amount of data to be retrieved. In most cases, the binary function is used since the information is encoded in bits. Info(D) is also known as the entropy of dataset D. Equation (3) is used to calculate the information gain.

$$Gain(A) = Info(D) - Info_A(D) \quad (3)$$

The information gain measure is used to select attributes with many values. Earnings ratios are processes for improving results. It is calculated by the formula shown in Equation (5).

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{D} x\ log_2\left(\frac{|D_j|}{|D|}\right) \quad (4)$$

$$GainRatio(A) = \frac{GainRatio(A)}{SplitInfo(A)} \quad (5)$$

The Gini index considers a binary split for each attribute. No matter how many values an input parameter has, possible subsets that can occur are formed by grouping binary pairs. This situation is calculated with the formula shown in Equation (6).

$$Gini_A(D) = \frac{|D_1|}{D} Gini(D_1) + \frac{|D_2|}{D} Gini(D_2) \quad (6)$$

When a decision tree is built, most of the branches will reflect anomalies in the training data due to noise or outliers. Small changes in values can cause completely different results. Various methods are used to eliminate this inconsistency caused by overfitting. Tree pruning is a method that typically uses statistical measurements to remove the least reliable branches or branches supported by a few samples. Pruned trees tend to be smaller and less complex and therefore easier to understand. They are also often faster and better at accurately classifying independent test data. The pruning is carried out by the method shown in Equation (7).

$$\Delta Gini(A) = Gini\ (D) - Gini_A(D) \quad (7)$$

The above-mentioned mathematical models can be coded in Python using the necessary libraries. Pseudo codes of the codes performed for the necessary operations are shown in Figure 3 and Figure 4.

Algorithm 2 Decision_Tree_Classifier Pseudo Code

```
Import libraries
Load dataset ()
    Split dataset (Input Values)
    Split dataset (Output Values)
function Create Model (params) ←values
function Model. Fit (entropy, depth)
dump (model, "destination.h5")
```

Figure 12. Decision Tree pseudo code

A model is developed with the code, decision trees algorithm shown in Figure 12, and the developed model is saved. Thanks to this recording process, the trained model can be shared with the desired point. When this sharing is realized, the node that will use the model can directly use the ready model without having training data.

Algorithm 3 Model_Client_Using Pseudo Code

```
Import libraries
Model_Load ("destination.h5")
Test_data = numpy. Array ("Input Values")
Result = Model.Classify ("Test_data")
Print (Result)
```

Figure 13. Client model pseudo code

Figure 4 shows the structured code block to directly use a trained model. Structures at the nodes will be able to use the model directly without entering the training process by using this code block. This will save on the training process.

The model success obtained by collecting and training the data belonging to its own region, which is one of the critical points of the study, remains the local model, while the model that is collected and trained from all regions is described as a global model. Thanks to the code seen in Figure 13, local models will be able to serve as global models.

5. Findings

In this study, the data set containing the Air Pollution Measurement Information of the city of Seoul, Korea was used. This dataset contains air pollution measurement information and result information for 25 different districts of Seoul city. The dataset is open sourced on the Kaggle dataset platform with license number CC BY-SA 4.0. Air pollution values obtained in certain regions vary between certain limits. The values obtained in this region were filtered from the entire data set and a data set specific to these regions was created. Afterwards, these regions were trained with the values obtained in their own regions and the model was recorded. When this model is tested with the data obtained in its own fields, it shows very successful results. However, when tested with data obtained from different regions and dissimilar to these regions, the model classification success is very low. To provide this information, we have listed the test results we performed in Table 3. In Table 3, SO2, NO2, O3, CO, PM10, PM25 data in the data set are shown to the model as input values. It is the classification of one of the "Good", "Normal", "Bad", "Very Bad" results determined as the expected output values from the model. Each of the output values is expressed with different colors in Table 2. In the results section, the classification value of the developed model and the actual result value are shown.

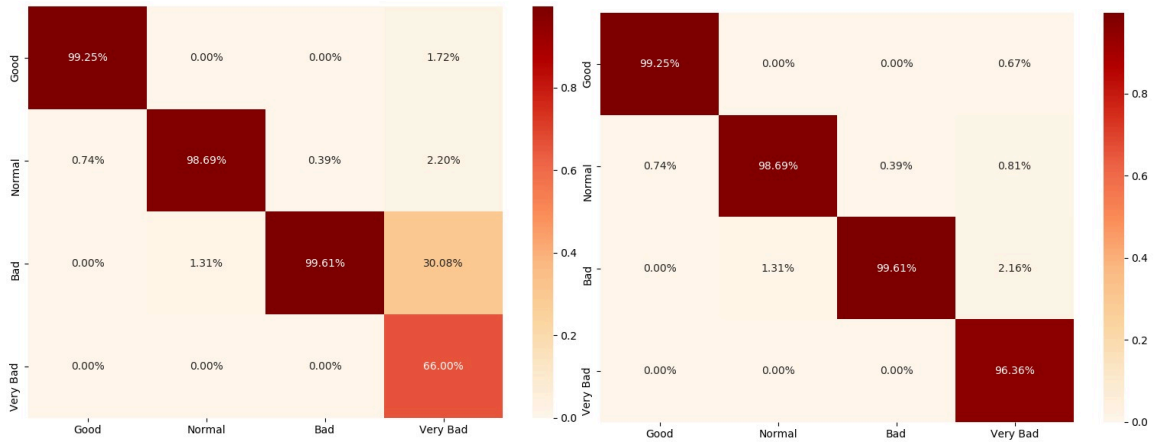
Table2. Model success and error status tested with values obtained from different fields

INPUT VALUES						OUTPUT VALUES				RESULTS	
SO_2	NO_2	O_3	CO	PM10	PM25	Good	Normal	Bad	Very Bad	Model Result	Actual Result
3.736	38.445	12.455	0,4	35	17	**1**	2	3	4	**2**	**4**
0,276	0,135	0,031	26,9	985	985	**1**	2	3	4	**3**	**4**
0,095	0,006	0,007	10,7	985	985	**1**	2	3	4	**2**	**4**
0,194	0,002	0,188	16,9	57	50	**1**	2	3	4	**2**	**4**
0,106	0,01	0,084	2,2	1985	985	**1**	2	3	4	**2**	**4**
0,134	0,063	0,059	13,3	171	178	**1**	2	3	4	**3**	**4**
0,112	0,005	0,006	11,9	985	985	**1**	2	3	4	**2**	**4**
0,121	0,006	0,027	12,1	985	985	**1**	2	3	4	**2**	**4**
0,104	0,008	0,045	3,7	622	610	**1**	2	3	4	**2**	**4**
0,266	0,002	0,125	27,8	268	263	**1**	2	3	4	**2**	**4**
0,007	0,074	0,003	1,4	122	102	**1**	2	3	4	**3**	**3**
0,006	0,07	0,003	1,5	122	103	**1**	2	3	4	**3**	**3**
0,005	0,028	0,031	0,9	146	111	**1**	2	3	4	**2**	**2**
0,005	0,019	0,041	0,9	145	114	**1**	2	3	4	**2**	**2**
0,004	0,039	0,013	0,5	22	10	**1**	2	3	4	**1**	**1**
0,004	0,034	0,018	0,4	22	7	**1**	2	3	4	**1**	**1**
0,005	0,033	0,019	0,4	19	9	**1**	2	3	4	**1**	**1**
0,004	0,035	0,017	0,5	19	15	**1**	2	3	4	**2**	**2**
0,004	0,041	0,014	0,5	25	13	**1**	2	3	4	**1**	**1**

The result value predicted by the model and the actual value are expressed with the colors in the output values columns in the Table 2. In these test results of an accurate and sufficiently advanced global model, the actual values column and the predicted values column would be expected to have the same colors. However, it is seen in Table 2 that there is a failed model in all the test data it encounters, except for the data in which it is trained.

The success rate in the training, which was made by collecting all the data in the data set, was determined as 98.87%. In addition, the error matrix obtained in the test processes performed with the

model at the end of the training is shown in Figure 14/a. According to this error matrix, 99.25% accuracy is obtained in the data that will be included in the "Good" class, while there is a 0.47% margin of error. While 98.69% accuracy is achieved in the data that will be included in the "Normal" class, there is a 1.31% margin of error. Data that will be included in the "Bad" class has an accuracy of 99.61% and a margin of error of 0.39%. The data to be included in the "Very Bad" class has 66% accuracy and 34% margin of error.



a)Local Model b)Global Model

Figure 14. Confusion matrix for decision tree algorithm

The error matrix values obtained from sharing the model with the system stakeholders after the data were combined under a single roof and the training was carried out in a single center are shown in Figure 14/b. These results ensure that the distributed learning model is limited to serving a specific region and is a global model. In addition, it is seen that the success achieved with the global model has reached quite high values compared to the previous value. The model success, which showed 66% accuracy, reached 96.36%.

To determine the time taken to save data to the blockchain structure created in another module of the study, performance measurement was carried out with different parameters. The obtained value is given graphically in Figure 15. Different scenario situations were used in the creation of the graph obtained in Figure 15. According to this scenario situation, the desired amount of work per unit time, that is, the throughput value, was changed and the measurement was made by running 10 different threads for 20 seconds. Throughput times gradually increased from 25 to 400. In the column on the right side of the graph, each throughput value is represented by a color and in the body of the graph it is indicated by this color. Each measurement was increased by 25 units. The final measurement value has been increased by 100 units. Average delay times in measurement processes, number of requests and error percentages encountered during the process were recorded.

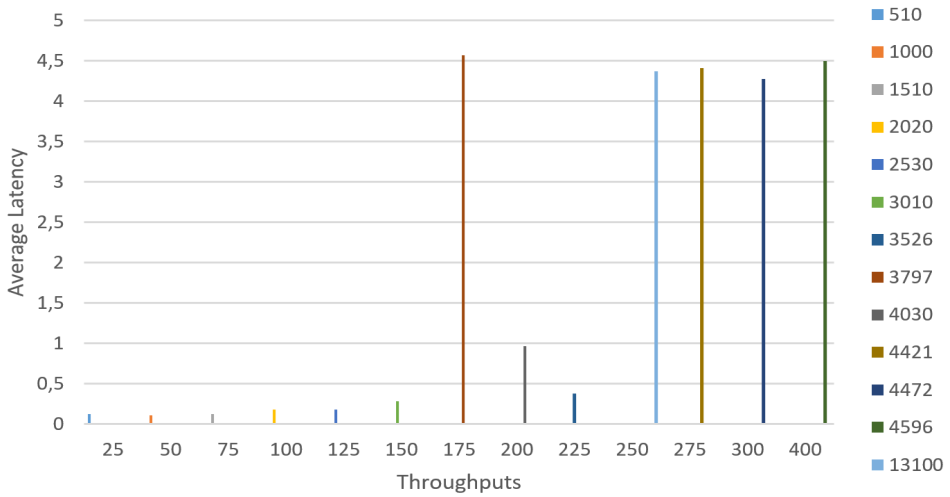
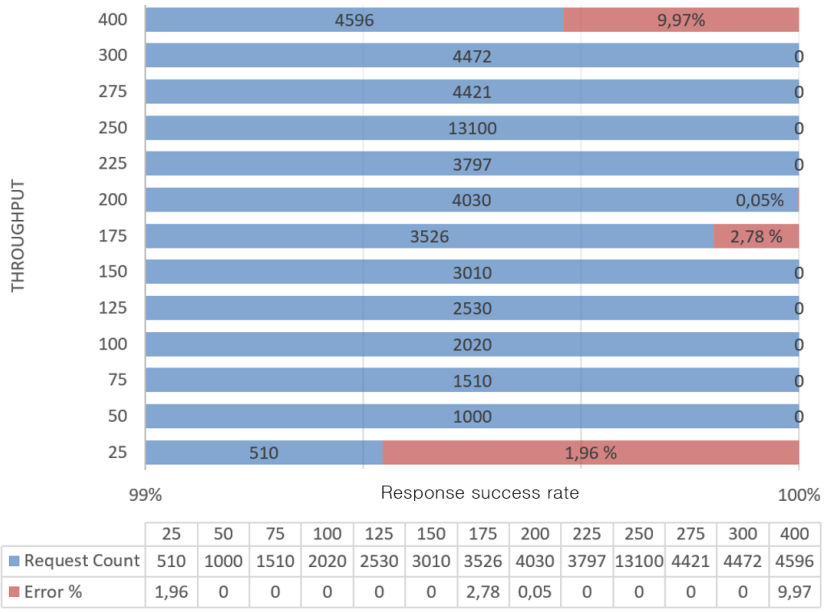


Figure 15. Blockchain performance evaluation

While recording data to the blockchain structure, an abnormal increase was noticed in the average delay time of the measurement, where the value throughout was set to 175. When the reason for this was investigated, it was determined that the connection with the blockchain server was broken during this test and an attempt was made to reconnect. The reason why the average delay time is abnormal can be verified from the graph showing the error value shown in Figure 16. This abnormality is also partially present in the test where the throughput value is set to 200. Similarly, it can be verified from the graph that the error value shown in Figure 126 is found. In addition, if the throughput value is set to 250 or more, it is seen that the delay time of the system starts to increase suddenly. This shows that the blockchain structure is approaching the upper limits of performance. When the graphs in Figure 15 and Figure 16 are examined together, it has been determined that the recording processes of the data started to be queued after the throughput value was set to 400. This data determines the upper limit of the system's performance.



	25	50	75	100	125	150	175	200	225	250	275	300	400
Request Count	510	1000	1510	2020	2530	3010	3526	4030	3797	13100	4421	4472	4596
Error %	1,96	0	0	0	0	0	2,78	0,05	0	0	0	0	9,97

Figure 16. Different throughput values for the percentage values of the error

The number of requests made according to different throughput values and the percentage values of the error rates encountered are shown in Figure 16 in graphic and tabular form. The number of requests made varies between 510 and 13100 in response to the throughput values that change gradually between 25-400. On the other hand, error percentages vary between 1.96% and 9.97%. When the error rates are examined, an error rate of 1.96% was encountered when the first record was tried to be created. The reason for this error is recorded as no response from the server. As soon as the data record requests are started, it is thought that the record requests are unanswered, and an error is made because the blockchain structure raises the ledger structure for the first record and performs the verification of the source from which the data comes from. In addition, small error values were observed when the throughput value was 175 and 200. The reason for this error is recorded in the records as the connection with the server was lost, so no response could be received, and the connection was reset again. Although no error message is encountered for the last measurement criterion, 400 throughputs, the reason why it is passed as an error is because the recording speed is higher than the system performance. It is thought that the error rate is high because some records are queued and held. This measurement value shows the performance upper limit of the constructed blockchain system.

The blockchain's average latency and throughput information can help to make some security interpretations. The latency of a blockchain network represents how long it takes for transactions to propagate through the network and be verified. Faster latency means that transactions are confirmed faster, and the network runs faster. This means that users can process transactions faster and the network can operate more efficiently. In terms of security, fast latency makes it less likely for attackers to manipulate transactions or carry out attacks such as double spending. Throughput refers to the number of transactions that a blockchain network can process in each period. High throughput means that the network can process more transactions at the same time and run faster. This means that the

network can scale better and operate efficiently to meet user demands. From a security perspective, high throughput values make it harder for attackers to force the network to congest or degrade the network's performance.

The Hyperledger Fabric infrastructure is an open-source platform that provides secure and scalable blockchain solutions between its users. Security is a key feature of Hyperledger Fabric. It executes different security mechanisms within itself. Hyperledger Fabric enables users to securely execute transactions on the network through authentication processes and authorization mechanisms. Each participant on the network is given a unique digital identity and transaction ownership is verified by using this identity in transactions. It provides transaction-level privacy through private channels and smart contracts to protect the confidentiality of transaction data. This means that only interested parties can see the transaction data, making it inaccessible to other participants in the network. Participants on the network provide secure communication between themselves using their digital certificates. Provides role-based access control between users. This enables participants to define the roles and permissions required to perform certain operations. In this way, only participants with the relevant permissions can perform operations for which they are authorized. Smart contracts include various security controls and validation mechanisms to prevent unauthorized access and attacks from any part of the code. Hyperledger Fabric includes security controls and algorithms to detect and prevent attacks on the network. When attacks are detected, notifications are sent to network administrators and appropriate actions are taken.

6. Conclusions

In this study, a data set containing air pollution information collected from different locations was used. To serve the purpose of the study, the values obtained from similar regions in the data set were filtered from the entire data set and a dataset specific to these regions was created. With this dataset, model training was carried out for the classification process using the DT algorithm. Testing of the trained model was carried out with heterogeneous data from the entire data set. The success rate of the obtained result is not at the desired level. In this situation, a distributed learning solution was applied by considering data privacy and data security principles. In this solution, the model can be saved and transmitted to different nodes via smart contracts prepared in the Hyperledger Fabric infrastructure. After this process was performed, it was retested on the same nodes with heterogeneous data from the entire data set, and the success rates increased significantly and ranged from 98% to 99%. To perform the performance tests of the proposed system, different scenarios were constructed, and measurements were made. According to this scenario, the required amount of work per unit time, that is, the throughput value, was changed and 10 different threads were run for 20 seconds, and the throughput times were gradually increased between 25 and 400, and measurements were made. According to the results obtained, process performance and error rates were obtained at the desired level.

In future studies for the implementation of this system, it is aimed to secure and automate the air-conditioning processes of greenhouses located in different locations. In the targeted future study, it is planned to develop an IoT member capable of collecting data from the environment in real time and managing air conditioning systems. Machine learning method and blockchain platform setup will be realized with the inspiration obtained from this study.

Acknowledgment

The study was supported by Süleyman Demirel University Scientific Research Projects Coordination Unit with project number FDK-2022-8719. We would like to thank Süleyman Demirel University Scientific Research Projects Coordination Unit for their support.

This article was produced from the first author's doctoral thesis, "Blokzincir Tabanli Dağitik Öğrenme Modelleri İçin Hesaplama Altyapilarinin Gerçekleştirilmesi", which was accepted at Süleyman Demirel University, Institute of Science and Technology.

Conflict of Interest Statement

The authors declare that there is no conflict of interest.

References

[1] O. Güler and S. Savaş, "All aspects of metaverse studies, technologies and future," *Gazi Journal of Engineering Sciences*, vol. 8, no. 2, pp. 292–319, Sep. 2022. doi:10.30855/gmbd.0705011

[2] E. A. Çubukçu, V. Demir and M. F. Sevimli, "Estimating streamflow data with machine learning techniques," *Gazi Journal of Engineering Sciences*, vol. 8, no. 2, pp. 257–272, Sep. 2022. doi:10.30855/gmbd.0705009

[3] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans Wirel Commun*, vol. 19, no. 3, pp. 2022–2035, Mar. 2020. doi:10.1109/TWC.2019.2961673

[4] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: challenges, methods, and future directions," *IEEE Signal Process Mag*, vol. 37, no. 3, pp. 50–60, May 2020. doi:10.1109/MSP.2020.2975749

[5] Cao, X., Başar, T., Diggavi, S., Eldar, Y. C., Letaief, K. B., Poor, H. V., & Zhang, J. ., "Communication-efficient distributed learning: an overview," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 4, pp. 851–873, Apr. 2023. doi:10.1109/JSAC.2023.3242710

[6] Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X., He, B., "A survey on federated learning systems: vision, hype and reality for data privacy and protection," *IEEE Trans Knowl Data Eng*, Apr. 2021. doi:10.1109/TKDE.2021.3124599

[7] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowl Based Syst*, vol. 216, p. 106775, Mar. 2021. doi:10.1016/J.KNOSYS.2021.106775

[8] J. P. Albrecht, "How the GDPR Will Change the World," *European Data Protection Law Review (EDPL)*, vol. 2, 2016, Accessed: Jun. 20, 2023. [Online]. Available: https://heinonline.org/HOL/Page?handle=hein.journals/edpl2&id=313&div=&collection=

[9] X. Wang, W. Qin, F. Jiao, L. Dong, J. Guo, J. Zhang and C. Yang, "Review of tungsten resource reserves, tungsten concentrate production and tungsten beneficiation technology in China," *Transactions of Nonferrous Metals Society of China*, vol. 32, no. 7, pp. 2318–2338, Jul. 2022. doi:10.1016/S1003-6326(22)65950-8

[10] S. Raschka, "MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack," *J Open Source Softw*, vol. 3, no. 24, p. 638, Apr. 2018. doi:10.21105/joss.00638

[11] A. Gaier and D. Ha, "Weight agnostic neural networks," *Adv Neural Inf Process Syst*, vol. 32, 2019, Accessed: Jun. 20, 2023. [Online]. Available: https://weightagnostic.github.io/

[12] E. Tijan, S. Aksentijević, K. Ivanić and M. Jardas, "Blockchain technology ımplementation in logistics," *Sustainability 2019, Vol. 11, Page 1185*, vol. 11, no. 4, p. 1185, Feb. 2019. doi:10.3390/SU11041185

[13] J. Golosova and A. Romanovs, "The advantages and disadvantages of the blockchain technology," *2018 IEEE 6th Workshop on Advances in Information, Electronic and Electrical Engineering, AIEEE 2018 - Proceedings*, Dec. 2018. doi:10.1109/AIEEE.2018.8592253

[14] Z. Fauziah, H. Latifah, X. Omar, A. Khoirunisa and S. Millah, "Application of blockchain technology in smart contracts: a systematic literature review," *Aptisi Transactions on Technopreneurship (ATT)*, vol. 2, no. 2, pp. 160–166, Aug. 2020. doi:10.34306/ATT.V2I2.97

[15] "A Blockchain Platform for the Enterprise — hyperledger-fabricdocs main documentation." https://hyperledger-fabric.readthedocs.io/en/release-2.5/ (accessed Jun. 20, 2023).

[16] "Blockchain Technology Projects – Hyperledger Foundation." https://www.hyperledger.org/use (accessed Jun. 20, 2023).

[17] Q. Nasir, I. A. Qasse, M. Abu Talib and A. B. Nassif, "Performance analysis of hyperledger fabric platforms," *Security and Communication Networks*, vol. 2018, 2018. doi:10.1155/2018/3976093

[18] S. Tanwar, Q. Bhatia, P. Patel, A. Kumari, P. K. Singh and W. C. Hong, "Machine learning adoption in blockchain-based smart applications: the challenges, and a way forward," *IEEE Access*, vol. 8, pp. 474–448, 2020. doi:10.1109/ACCESS.2019.2961372

[19] Y. Liu, F. R. Yu, X. Li, H. Ji and V. C. M. Leung, "Blockchain and machine learning for communications and networking systems," *IEEE Communications Surveys and Tutorials*, vol. 22, no. 2, pp. 1392–1431, Apr. 2020. doi:10.1109/COMST.2020.2975911

[20] A. Adhikari, D. B. Rawat and M. Song, "Wireless network virtualization by leveraging blockchain technology and machine learning," *WiseML 2019 - Proceedings of the 2019 ACM Workshop on Wireless Security and Machine Learning*, pp. 61–66, May 2019. doi:10.1145/3324921.3328790

[21] R. Shinde, O. Nilakhe, P. Pondkule, D. Karche and P. Shendage, "Enhanced Road Construction Process with Machine Learning and Blockchain Technology," *2020 International Conference on Industry 4.0 Technology, I4Tech 2020*, pp. 207–210, Feb. 2020.

doi:10.1109/I4TECH48345.2020.9102669

[22] M. U. Nasir, S. Khan, S. Mehmood, M. A. Khan, M. Zubair and S. O. Hwang, "Network meddling detection using machine learning empowered with blockchain technology," *Sensors 2022, Vol. 22, Page 6755*, vol. 22, no. 18, p. 6755, Sep. 2022. doi:10.3390/S22186755

[23] T. Wang, "A unified analytical framework for trustable machine learning and automation running with blockchain," *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pp. 4974–4983, Jan. 2019. doi:10.1109/BIGDATA.2018.8622262

[24] Z. A. El Houda, A. Hafid and L. Khoukhi, "BrainChain - A machine learning approach for protecting blockchain applications using SDN," *IEEE International Conference on Communications*, vol. 2020-June, Jun. 2020. doi:10.1109/ICC40277.2020.9148808

[25] W. Xiong and L. Xiong, "Smart contract based data trading mode using blockchain and machine learning," *IEEE Access*, vol. 7, pp. 102331–102344, 2019. doi:10.1109/ACCESS.2019.2928325

[26] Z. Shahbazi and Y. C. Byun, "Improving transactional data system based on an edge computing–blockchain–machine learning ıntegrated framework," *Processes 2021, Vol. 9, Page 92*, vol. 9, no. 1, p. 92, Jan. 2021. doi:10.3390/PR9010092

[27] N. V. Pardakhe and V. M. Deshmukh, "Machine learning and blockchain techniques used in healthcare system," *2019 IEEE Pune Section International Conference, PuneCon 2019*, Dec. 2019. doi:10.1109/PUNECON46936.2019.9105710

[28] Khan, M. A., Abbas, S., Rehman, A., Saeed, Y., Zeb, A., Uddin, M. I., Nasser, N., Ali, A, "A machine learning approach for blockchain-based smart home networks security," *IEEE Netw*, vol. 35, no. 3, pp. 223–229, May 2021. doi:10.1109/MNET.011.2000514

[29] F. Jamil, H. K. Kahng, S. Kim and D. H. Kim, "Towards secure fitness framework based on ıot-enabled blockchain network ıntegrated with machine learning algorithms," *Sensors 2021, Vol. 21, Page 1640*, vol. 21, no. 5, p. 1640, Feb. 2021. doi:10.3390/S21051640

[30] Uddin, M., Memon, M. S., Memon, I., Ali, I., Memon, J., Abdelhaq, M., & Alsaqour, R., "Hyperledger fabric blockchain: secure and efficient solution for electronic health records," *Computers, Materials & Continua*, vol. 68, no. 2, pp. 2377–2397, 2021. doi:10.32604/cmc.2021.015354

[31] F. P. Oikonomou, J. Ribeiro, G. Mantas, J. M. C. S. Bastos and J. Rodriguez, "A hyperledger fabric-based blockchain architecture to secure ıot-based health monitoring systems," *2021 IEEE International Mediterranean Conference on Communications and Networking, MeditCom 2021*, pp. 186–190, 2021. doi:10.1109/MEDITCOM49071.2021.9647521

[32] N. Lu, Y. Zhang, W. Shi, S. Kumari and K. K. R. Choo, "A secure and scalable data integrity auditing scheme based on hyperledger fabric," *Comput Secur*, vol. 92, p. 101741, May 2020. doi:10.1016/J.COSE.2020.101741