Research on the Privacy-Preserving Method of Medical Data Based on Blockchain

Kaili Lu, Haiyan Kang

Abstract— **With the continuous development and progress of big data technology, the storage and sharing of information-based medical data has greatly promoted the development of medical treatment. However, the information systems of most medical institutions use their own independent centralized storage methods, which makes it difficult to carry out efficient data interconnection between medical institutions and brings inconvenience to patients' cross-institution treatment. On the other hand, when the server is attacked by criminals, some important data is likely to be leaked, and then a series of chain problems such as private information being illegally sold are produced. Blockchain technology has the characteristics of decentralization, tamper-proof, traceability and so on, which can ensure that the system has good robustness and data security, and can solve the problem of data privacy security. Therefore, the medical data privacy protection platform based on blockchain technology will play a huge role in medical data security.**

***Index Terms*—blockchain; Privacy-Preserving; Digital healthcare; Medical Data**

I. INTRODUCTION

With the continuous improvement of the level of social information, the digitization of the medical industry has become an inevitable development trend. The concept of digital medical has been widely recognized by all walks of life and has made fruitful development results. However, many of the current HER systems neglect the protection of patient privacy, and data leakage problems are common. At the same time, the EHR systems of medical institutions are independent of each other, and the data is not interconnected, which makes it difficult to obtain electronic health records when patients are transferred to hospitals. In addition, with the continuous development of EHR systems, the data volume of relevant medical data will increase significantly, and distributed storage is essential in order to ensure data security. If efficient and reliable data sharing can be realized, it will facilitate the cross-hospital treatment of patients, which will further promote the development of the healthcare industry.

There are still some limitations in current electronic health record systems. For example, a large number of medical institutions independently maintain a set of electronic health record systems, and the management style, data format and type of each system are not the same, which is difficult to interconnect, resulting in a serious "data island" problem. At the same time, it is obviously impossible for a patient to visit the same hospital all his life, especially in the context of the increasingly frequent population mobility in China, patients will have to carry one or more paper medical records to new medical institutions to provide doctors with complete medical record information. This not only brings inconvenience to patients, but also leads to the maintenance of a large number of redundant data in the e-health system of various medical institutions. A few medical institutions have realized medical data sharing based on cloud storage within a certain range. Although this cloud storage scheme does not occupy local storage space and data transmission speed is fast, there are still security risks when using cloud servers, such as data read and write rate is affected by network speed and is vulnerable to attacks. The resulting loss of data is unacceptable for the risk-sensitive medical domain. Especially in the case of multiple medical institutions, once such a centralized storage system is damaged, all nodes will be unable to store and use data, that is, there is a "single point of failure" problem.

The blockchain is composed of chained data blocks, and the chain structure is encrypted based on cryptography technology. Strict encryption technology protects the security of stored data. The data block can not only verify the rationality and validity of the transaction information, but also generate the next block. At the same time, each independent data block also stores the transaction data in the network. Blockchain has the characteristics of ensuring that transaction information cannot be tampered with, the system is highly open, the decentralization of distributed computing and storage, and the anonymity of transaction parties without public identity. Therefore, in the medical related digital field, blockchain has very obvious application scene advantages. Its main advantages are :(1) blockchain can effectively achieve value transfer and protect customer data, making data tamper-resistant. (2) Data participants jointly maintain the same information ledger to ensure the accuracy of the information in the ledger. (3) Effectively help users obtain more rights and reconstruct the production relations of the Internet.

II. BASIC CONCEPTS OF MEDICAL DATA

A. Digital Healthcare

Digital medicine is the modern computer technology, information technology applied to the whole medical process of a new type of modern medical treatment, mainly reflected in the digital and network of medical equipment, medical institution management information and personalized medical services [1], is the development direction and management goal of public medical treatment. The emergence of digital

Manuscript received March 01, 2023.

Kaili Lu, School of Computer Science, Beijing Information Science and Technology University, Beijing, China

Haiyan Kang, School of Information Management, Beijing Information Science and Technology University, Beijing, China

medical equipment has greatly enriched the connotation and capacity of medical information. The visualization of electrophysiological information and medical image information greatly enriches the diagnosis technology of doctors, and makes medicine enter a new visual information age. On the basis of the digitization of medical equipment, the staff of medical institutions can easily carry out their work, understand the operation of various departments at any time, and realize the efficient management of medical institutions. With the popularity of wearable devices and the application of big data technology, medical data can be collected and analyzed in large amounts, so as to provide personalized medical services for patients and predict potential future health risks [2].

B. Medical Data Standards

At present, the mainstream medical data standards include International Health Data Standards, International Medical Digital Imaging Communication Standards, Clinical Data Exchange Standards Association, etc. Taking medical Imaging as an example, DICOM (Digital Imaging and Communications in Medicine) is a widely adopted international standard for medical imaging [3]. It is based on the clinical needs for image quality. The format and quality of medical image in the process of data exchange are clearly defined. At present, the number of medical images conforming to DICOM standards in clinical use has reached 10 billion. Files stored according to this standard are called DICOM files.

C. Electronic Health Records

At present, the electronic health record usually refers to the electronic medical record, used to replace the traditional paper medical record, which contains the patient's gender, age and other basic information, symptom description, medical examination results and imaging data, preliminary diagnosis results and medication prescriptions. It basically includes the complete treatment process. In the future, it will gradually move closer to the electronic health record, which will record almost all the content related to medical services, including not only people's medical records, but also various vaccinations, physical examination results, medical insurance protection and so on. Basically cover a person's life. The development of electronic health records is not only beneficial to people to know their health status, but also to reduce medical costs. It has made great contributions to improving the quality of medical services.

D. Privacy Leakage of Medical Data

Medical data, which involves a large number of important user personal information of patients, is different from general important business data, which has high accuracy and social relationship attributes, and often becomes the preferred attack object of attackers. Therefore, data leakage incidents in the medical industry have been common in recent years. The impact of data breaches on patients is very large and may be harassed by the medical insurance industry. The disclosure of health information of patients with special diseases may seriously damage their personal reputation and lead to discrimination of patients and their families. What's worse, an attacker could use this information to commit fraud, blackmail, or tamper with the information. Mass health and medical data leakage is more terrible, easy to cause public opinion, and easy to cause social panic.

The above security risks have led many organizations and individuals to resist medical data sharing due to privacy concerns. Therefore, it is particularly important to design a secure and reliable data sharing scheme.

III. THEORIES RELATED TO BLOCKCHAIN TECHNOLOGY

A. Blockchain Definition

Blockchain technology originates from Bitcoin, which is a very popular information technology on the Internet in recent years. It is essentially a shared database, which is integrated by existing technologies such as cryptography and P2P technology [4], and has the characteristics of tamper-proof, traceability, openness and transparency [5]. Generally speaking, blockchain system consists of data layer, network layer, consensus layer, incentive layer, contract layer and application layer [6]. According to different application scenarios, blockchain is divided into three types: public chain, private chain and consortium chain. The differences between these three blockchain systems are mainly reflected in the degree of centralization and the restrictions on nodes joining the network.

All the nodes in the system have the right to keep accounts, and the nodes that keep accounts can also obtain certain material incentives. The data is completely open and transparent, so its degree of decentralization is the highest, and the typical application project is Ether. However, due to the large number of nodes and unknown identities, the transaction speed of the system is generally not high, and there is a risk of malicious nodes to attack. At the same time, accounting incentives and computing power competition caused by competing for accounting rights between nodes also increase the cost of transactions. In contrast, private chains are closed and maintained by a single individual or organization. Consortium blockchain is generally oriented to multiple specific institutions, and only authorized nodes are allowed to join. The accounting nodes are specified in advance, and the blocks are decided by these accounting nodes alternately or jointly. It can be seen that private blockchains and consortium blockchains have faster transaction speed, very low or even zero transaction costs, and more secure systems due to the small number of nodes and identity authentication. But at the same time, they are also slightly less decentralized and more biased towards weak centralization or multi-centralization. Therefore, the choice of blockchain should be properly weighed according to the characteristics and needs of the actual scene. At present, the development prospect of alliance chain is more promising, and it has been widely used in product traceability, financial transactions, information security and other fields.

B. Blockchain Data Structure

(1) Data structures

At the technical level, the data structure of the blockchain is shown in Fig. 1 The blockchain contains N block nodes arranged in chronological order, and the block structure of each block node is divided into block head and block body. In the block head, there is a pointer to the previous block, and all blocks are connected to form a chain through the pointer [7].

The block header mainly includes the version number, the previous block hash value, the timestamp, the random number, the target hash, and the Merkle root. In the block body is the ledger information recorded through the Merkle tree.

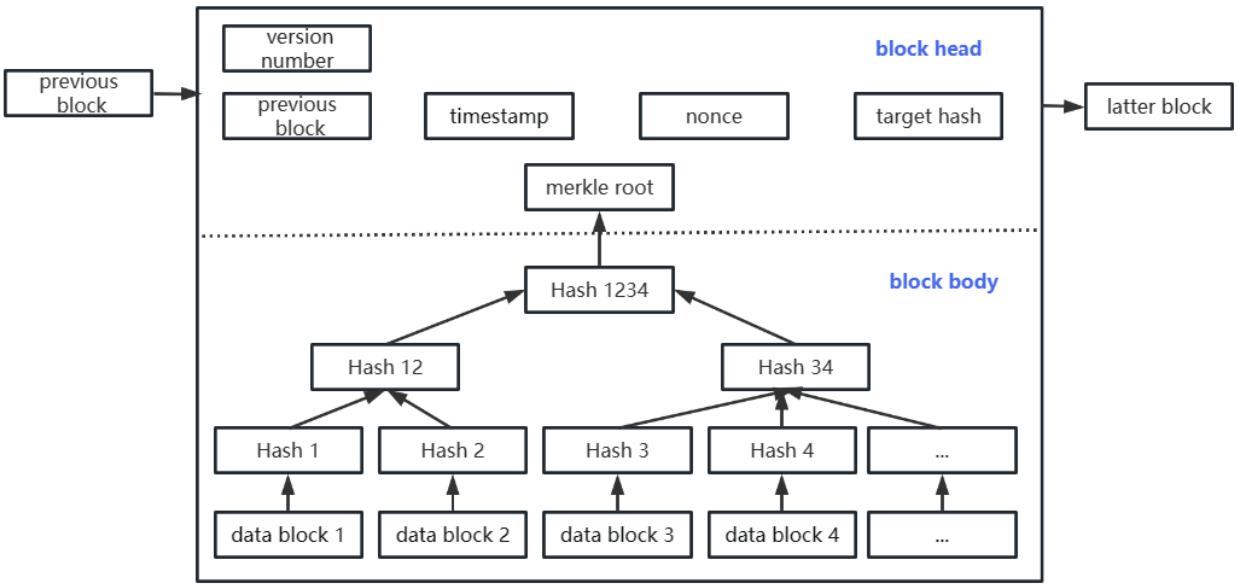

Fig.1 Data structure of the blockchain

The hash value contained in the block header realizes the physical link between different blocks, and the hash value is used as the unique identity of each block to take advantage of the characteristics of hash function. All transaction information is based on the hashing process of Merkle tree. The hash data sequence of the previous parent block contained in the block header is queried to obtain the next block. Through this transitive relationship, since the initial block, the previous and subsequent blocks are connected with each other according to the hash data value, forming an interconnected chain whole in physical structure. It can store Merkle tree roots permanently in a block [8].

(2) Hierarchical structure

When blockchain technology was first proposed, it was usually divided into six layers in the architecture, namely, data layer, network layer, consensus layer, incentive layer, contract layer and application layer. With the development of time, the research of blockchain technology is more and more in-depth, blockchain technology is developing rapidly, and the blockchain architecture is also changing. Many traditional modules are weakened, and the incentive layer mechanism has even been replaced in alliance chain and private chain technology. Now the blockchain is divided into three layers: network layer, transaction layer and application layer. The architecture diagram is shown in Fig. 2.

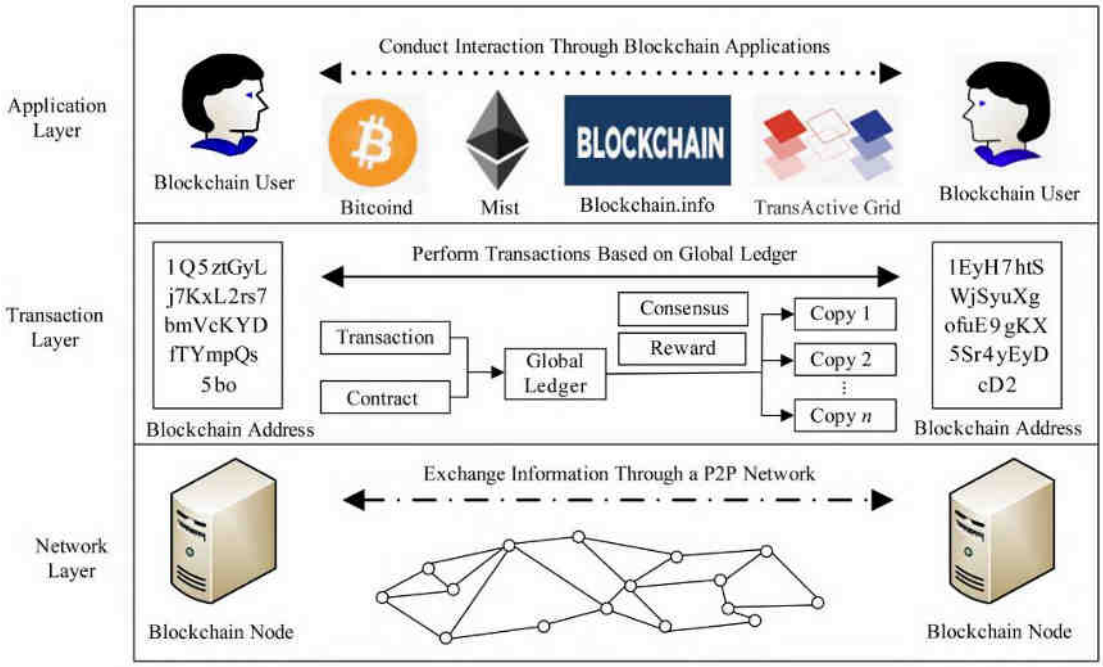

Fig. 2 Diagram of the blockchain architecture

C. Consensus Mechanism

In blockchain, data is maintained by all nodes, but in order to ensure data consistency, only one node can be used for each block of data. So there is competition for the right to keep accounts. The algorithms that solve this problem are called consensus mechanisms. At present, the mainstream blockchain consensus mechanisms mainly include PoW

algorithm, PoS algorithm, DPoS algorithm and PBFT algorithm. Table 1 is the performance comparison of commonly used consensus algorithms.

Dwork C and Naor M first proposed the concept of Proof-of-Work (PoW) in an academic paper in 1993 [9], and Jakobson M and Juels A formally introduced the term proof-of-work in a paper published in 1999 [10]. Later, in his 2008 paper, Satoshi Nakamoto introduced Proof of work as the consensus computation for Bitcoin [11]. The success of Bitcoin has had a huge impact on cryptocurrencies, and PoW has become one of the mainstream consensus algorithms. The core of the PoW algorithm is that the Hash value of the block head calculated by the node is required to be less than the target Hash value given in advance [12].

Table 1 Performance comparison of common consensus algorithms

Index of evaluation	PoW	PoS	DPoS	PBFT
Performance Efficiency	low	high	high	high
Degree of Decentralization	high	high	low	low
Fault-tolerant %	50	50	50	50
Deterministic	probabilistic	probabilistic	probabilistic	probabilistic
Resource Loss	high	low	low	low
Applicable Scene	Public blockchain	Public blockchain	Public blockchain	Consortium chain

Proof-of-Stake (PoS) selects the accounting node through a set of token mechanism [13]. In simple words, nodes can buy tokens, and according to the number of tokens held by nodes and the time, the algorithm will reduce the difficulty of mining nodes, speed up the discovery of Nonce values, and shorten the consensus time [15]. This mechanism encourages users to use the funds originally used to buy mining equipment to buy tokens, which significantly alleviates the waste of computing resources and energy.

Delegated Proof-of-Stak (DPoS) is one of the most high-throughput, decentralized, and flexible consensus mechanisms available in the mainstream [16]. DPoS algorithm uses nodes to exercise voting rights to approve all kinds of transactions, and solves the consensus problem in a democratic and fair way. All network parameters, from cost estimates to block spacing and transaction sizes, can be adjusted by the selected representatives. The principle is to let each node vote to generate a certain number of representatives with equal rights among representatives [17]. These representatives (also known as super-nodes) generate blocks sequentially on behalf of all nodes, and jointly perform block validation and accounting. If a super-node is unable to effectively perform its duties (e.g. generate blocks), it will be removed and a new super-node will be selected to replace it. After successful block generation, all super-nodes will receive about 10% transaction fee as a reward [15]. The advance determination of billing rights allows transactions to be confirmed within 1 second on average. By protecting super nodes from unnecessary logical checks, compared with PoW and PoS algorithms, DPoS greatly reduces the number of nodes involved in verification and accounting, improves efficiency, and can reach the consensus speed of seconds [16]. It is easy to see that the DPoS algorithm is closer to weak centralization or multi-centralization [17] than decentralization, which belongs to a compromise to system throughput.

When the number of consensus nodes is small (anywhere from a few to a few dozen) and assuming there are malicious or failed nodes, the problem of achieving correct and effective consensus can be modeled as Byzantine Failures. The consensus algorithms to solve this kind of problem are collectively called Byzantine Fault Tolerance (BFT). In essence, the Byzantine fault tolerance scheme is that the minority is subordinate to the majority. Practical Byzantine Fault Tolerance (PBFT)[18] reduces the time complexity of the algorithm to polynomial level, greatly improves the efficiency, and can work normally when there are at most 1/3 malicious nodes or faulty nodes in the whole network. It is feasible in practical application. At present, most of the consensus algorithms adopted by various consortium blockchains are developed based on PBFT algorithm. The advantages of PBFT are obvious, it does not need to rely on mining to reach consensus, and is very energy efficient. Blockchains are also less prone to fork. When the number of nodes is small, it has very high throughput and very low transaction delay. The master node and the deputy node are relatively equal, and they do not have too much power. Once they commit evil, the request will be rejected and replaced by the deputy node.

D.Smart Contracts

Smart contract is a contract that runs in cyberspace relying on computers [19]. It is transmitted, verified or executed in an information-based way, read and executed by computers, and has the characteristics of self-service. The decentralization of blockchain and the tamper-proof of data determine that smart contracts are more suitable for implementation on blockchain. Therefore, with the development of blockchain technology, smart contracts have a broader development prospect.

In fact, a smart contract is a program composed of computer code, and its conclusion process is as follows: in the first step, the two or more parties involved in the contracting will agree to formulate a smart contract; In the second step, the smart contract was broadcast and stored to the fulcrums of each blockchain around the world through the blockchain network. The third step is to build a successful smart contract to automatically execute the contract content after the condition is reached.

As a computer protocol, its dissemination, verification and execution of contracts all use information means, which makes smart contracts do not need a trusted third party when conducting transactions, and these transactions can be traced and irrevocable. Compared with traditional contracts, smart contracts reduce the cost of executing contracts in the

transaction process and are more secure. However, due to the limitation of The Times, a secure and reliable execution environment could not be established at that time, so smart contracts could not be put into practical use for a long time and only stayed in the conceptual stage. With the emergence of blockchain, people realize that it coincides with the idea of decentralization of blockchain. Ethereum is a typical project combining blockchain and smart contract. There are two types of smart contracts in blockchain: Turing-complete and non-Turing-complete. The former is more adaptable and supports sequential execution, conditional execution and loop execution, which is suitable for business scenarios with complex logic. For example, Ethereum Virtual Machine (EVM) is the current mainstream Turing-complete smart contract. However, non-Turing-complete smart contracts cannot fully support the above three execution statements, so their applications are very limited. In general, in scenarios such as electronic voting and electronic auction where human cheating may exist, smart contracts can be applied to efficiently and reliably implement pre-formulated rules through computers, so as to ensure the fairness and justice of business. However, as a computer transaction protocol, the protocol content of smart contract is open and transparent to all users in the network, which means that all protocol vulnerabilities are also publicly visible and difficult to repair in a short time. Therefore, once the vulnerability is triggered or maliciously exploited, it will bring systemic risk. This puts forward high requirements for the robustness of smart contracts.

IV. Design of Medical Data Privacy Protection Scheme Based on Blockchain

A. Structural Design of Medical Data Sharing System

In order to ensure the authenticity and security of medical data, this paper designs a blockchain-based medical data privacy protection system, which can protect the privacy of patients and medical institutions without sacrificing efficiency. The immutability of diagnostic records after being linked ensures the integrity of the record, and the traceability ensures that problems can be accurately traced. For data privacy protection, the AES algorithm is used to encrypt the diagnostic data and the SM2 algorithm is used to encrypt the AES encryption, and the medical institutions are used as group members to perform group signature on the ciphertext. While hiding the information of medical institutions, the data privacy is protected by encryption algorithm.

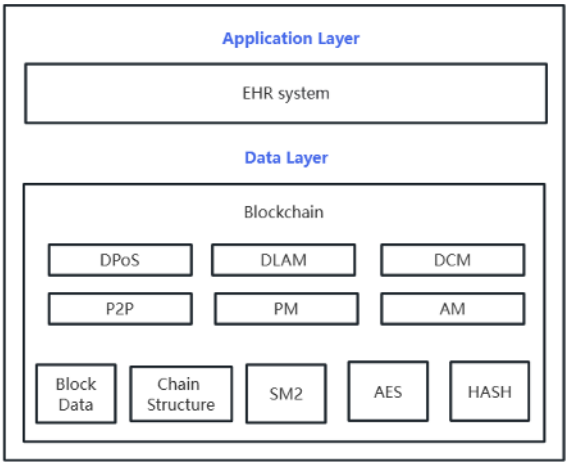

Fig. 3 Structure diagram of medical data sharing system based on blockchain technology

The system architecture of the medical data sharing system based on blockchain technology and cryptography technology is mainly divided into application layer and data layer. The application layer is an electronic health record system, and the data layer is a blockchain network based on PoS consensus algorithm. Fig. 3 shows the system structure of the medical data sharing system based on blockchain technology.

B. Functional Design of Medical Data Sharing System

The medical data sharing system based on blockchain technology mainly includes four modules, which are user management module, data processing module, blockchain consensus module, and data request module.

(1) The user management module includes registration, login, personal information input, password recovery and other functions for physician users and administrator users.

(2) The data processing module includes the functions of medical data encryption, symmetric key encryption, member signature and data upload.

(3) The blockchain consensus module includes transaction collection, accounting right competition, node consensus and other functions.

(4) The data request module includes the functions of administrator browsing data request, request transaction retrieval, symmetric key decryption, transaction decryption, signer tracing, request reply and so on.

C. Security Analysis of Medical Data Sharing System

In the medical data sharing system based on blockchain technology, AES-128 was used to encrypt the medical data, the AES key was encrypted by the public key of SM2 algorithm, and the ciphertext and the encrypted key were stored on the blockchain. Therefore, if an attacker wants to obtain the AES key to decrypt the ciphertext, he must first obtain the SM2 private key, which is secretly kept by the administrator. Therefore, it can be considered that the attacker cannot break the AES key and obtain the plaintext without the administrator leaking the private key. This ensures the security of the data. In addition, the inherent characteristics of blockchain also make transactions impossible to be tampered with.

On the basis of encrypted data storage, the desensitization of transactions is realized by dividing and processing transactions separately. The identity identifier calculated by SHA256 ensures that the patient's identity information is difficult to be deciphered but easy to be verified by the referring physician. At the same time, the medical institution information was further erased by using group signature, so that the transaction did not contain identity information and geographic location information. Therefore, even if the attacker illegally obtains the SM2 private key or uses the quantum computer to brute force to decrypt the plaintext, it can only obtain the medical information without identity direction, which effectively protects the privacy of patients.

V. Conclusion

With the application of medical big data, Internet hospital and other technologies, blockchain technology will bring new

opportunities and challenges for the security protection of medical data. The combination of the medical industry and emerging computing technologies will disrupt some of the traditional medical industry processes and technical measures. This paper comprehensively considers the key links of the data life cycle, such as data storage, data sharing and data traceability, and designs a blockchain-based medical data privacy protection model, in order to ensure data security, enhance the sharing degree of medical data, and further promote the development of the medical industry.

REFERENCES

[1] Jain, N, V. Gupta, and P. Dass. "Blockchain: A novel paradigm for secured data transmission in telemedicine." Wearable Telemedicine Technology for the Healthcare Industry (2022):33-52.

[2] Wu H T, Tsai C W. Toward blockchains for health-care systems: Applying the bilinear pairing technology to ensure privacy protection and accuracy in datasharing[J]. IEEE Consumer Electronics Magazine,2018,7(4): 65-71.

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

[4] Li F, Liu K, Zhang L, et al. EHRChain: A Blockchain-based EHR System Using Attribute-Based and Homomorphic Cryptosystem[J]. IEEE Transactions on Services Computing, 2021,

[5] Ghosh, T., A. Roy, and S. Misra. "B2H: Enabling delay-tolerant blockchain network in healthcare for Society 5.0." Computer networks Jun.19(2022):210.

[6] Gupta, B. B., M. Dabra, and K. E. Psannis. "Blockchain-Assisted Secure Fine-Grained Searchable Encryption for Healthcare Cloud-Based Cyber-Physical System." IEEE/CAA Journal of Automatica Sinica (2021).

[7] Mistry C, Thakker U, Gupta R , et al. MedBlock: An AI-enabled and Blockchain-driven Medical Healthcare System for COVID-19[C]// IEEE International Conference on Communications. IEEE, 2021.

[8] Bhagi, A., et al. "Blockchain technology for immunisation data storage in India: opportunities for population health innovation." BMJ Innovations 8.1(2022):1-3.

[9] Boumezbeur, I., and K. Zarour. "Privacy Preservation and Access Control for Sharing Electronic Health Records Using Blockchain Technology." Acta Informatica Pragensia 2022(2022).

[10] Boumezbeur, I., and K. Zarour. "Privacy Preservation and Access Control for Sharing Electronic Health Records Using Blockchain Technology." Acta Informatica Pragensia 2022(2022).

[11] Saeed, H., et al. "Blockchain technology in healthcare: A systematic review." PLOS ONE 17(2022).

[12] Mark, G., et al. "Blockchain and population health." Journal of Public Health 4(2022):4.

[13] Saranya, R., and A. Murugan. "A systematic review of enabling blockchain in healthcare system: Analysis, current status, challenges and future direction." Materials Today: Proceedings 6(2021).

[14] Alhabidi, A. M. "Blockchain benefits and barriers of implementation in healthcare: A scoping review." Computer Methods and Programs in Biomedicine 205.4(2021):105980.

[15] Gupta, R., et al. "GaRuDa: A Blockchain-Based Delivery Scheme Using Drones for Healthcare 5.0 Applications." IEEE Internet of Things Magazine 4(2021):4.

[16] Kshetri, N. "Healthcare and pharmaceutical industry supply chains - ScienceDirect." Blockchain and Supply Chain Management (2021):115-137.

[17] Hashim, F., K. Shuaib, and F. Sallabi. "MedShard: Electronic Health Record Sharing Using Blockchain Sharding." Sustainability 13.11(2021):5889.

[18] Uddin, M., et al. "Hyperledger Fabric Blockchain: Secure and Efficient Solution for Electronic Health Records." Computers, Materials and Continua 68.2(2021):2377-2397.

[19] Shrimali, B., and H. B. Patel . "Blockchain State-of-the-Art: Architecture, Use Cases, Consensus, Challenges and Opportunities." Journal of King Saud University - Computer and Information Sciences 3(2021).