Hindawi
Security and Communication Networks
Volume 2020, Article ID 6647562, 12 pages
https://doi.org/10.1155/2020/6647562

Research Article

A Privacy-Protection Model for Patients

Wenzhi Cheng, Wei Ou, Xiangdong Yin, Wanqin Yan, Dingwan Liu, and Chunyan Liu

School of Electronics and Information Engineering, Hunan University of Science and Engineering, Yongzhou 425199, Hunan, China

Correspondence should be addressed to Wei Ou; ouwei@huse.edu.cn

Received 2 October 2020; Revised 12 November 2020; Accepted 25 November 2020; Published 10 December 2020

Academic Editor: Honghao Gao

The collection and analysis of patient cases can effectively help researchers to extract case feature and to achieve the objectives of precision medicine, but it may cause privacy issues for patients. Although encryption is a good way to protect privacy, it is not conducive to the sharing and analysis of medical cases. In order to address this problem, this paper proposes a federated learning verification model, which combines blockchain technology, homomorphic encryption, and federated learning technology to effectively solve privacy issues. Moreover, we present a FL-EM-GMM Algorithm (Federated Learning Expectation Maximization Gaussian Mixture Model Algorithm), which can make model training without data exchange for protecting patient's privacy. Finally, we conducted experiments on the federated task of datasets from two organizations in our model system, where the data has the same sample ID with different subset features, and this system is capable of handling privacy and security issues. The results show that the model was trained by our system with better usability, security, and higher efficiency, which is compared with the model trained by traditional machine learning methods.

1. Introduction

Medical information contains personal information, which includes name, gender, age, and address, and the leakage of these information may cause serious risks to patients. Hathaliya and Tanwar [1] stated that patients' data is stored on cloud servers, and attackers can launch various security attacks, such as data confidentiality, privacy, and integrity attacks, which will pose a serious threat to patients. Prakash and Singaravel [2] stated that many organizations or companies collect personal privacy or sensitive data on the Internet, which means that personal information of sensitive data must be protected, and the individual has the right of management to access data or information for himself or herself. Papaioannou et al. [3] introduced that Internet of medical things (IoMT) provides e-health service for patients to help them improve their quality of life, but there are some attacks in the IoMT system (e.g., eavesdropping attacks, spoofing attacks, masquerading attacks, and Denial-of-Service attacks), which cause serious issues for patients. Bernd et al. [4] introduced that the advanced medical technology improves the efficiency of medical system and brings convenience to patients, but it also poses huge threats to the privacy of patients.

The collection and analysis of patient cases can effectively help researchers to extract case feature and to achieve the objectives of precision medicine, but it may cause privacy issues for patients. Hulsen et al. [5] informed that machine learning can identify patterns in biomedical data which can provide information for the development of clinical biomarkers or indicate unsuspected treatment targets with accelerating the goals of precision medicine, and patient information must be protected. Azencott [6] indicated that machine learning plays an important role in the development of precision medicine, which can tailor treatment plans based on the clinical or genetic features of patients, but these developments require the collection and sharing of a large amount of patient data or information, which will cause a crisis of patient privacy leakage. Gao et al. [7] and Yang et al. [8] proposed a sparsity alleviation recommendation approach that can achieve a better user information sharing and recommendation. Price et al. [9] stated that big data in health can help to measure hospital quality, to develop scientific hypothesis, and to monitor drug and device safety, but there is little to protect patients from health privacy threats.

The medical data has also the risk of privacy data leakage in the communication channel, and attackers may obtain or tamper with user information from the channel. Gao et al. [10, 11] presented a Manhattan mobility model, which can improve and optimize the channel transmission quality to achieve the expected information transmission. Ma et al. [12] solved communication problems by optimized scheduling of servers. Gnad et al. [13] stated that vulnerabilities may cause channel attacks in Internet of Things, which will pose a serious threat to users.

Encryption is a good way to protect privacy, but it is not conducive to the sharing and analysis of medical cases. Tian et al. [14] explained that if a patient dies during treatment and the secret key is lost, it will inevitably lead to the unavailability of the data, which is not conducive to the further study of the case. Meanwhile, if the patient falls into a coma and cannot share the secret key to access the case, it will also affect the patient's treatment, which may delay treatment and cause death. In this situation, encryption will limit the rescue work or medical research of doctors, which will lead to serious issues.

In order to protect patient information privacy, many researchers have taken different measures to solve privacy issues. Sun et al. [15] discussed that existing solutions of privacy use data encryption, access control, and trusted third-party auditing in Internet of Medical Things (IoMT), but there is still Information Island of sharing. Li et al. [16] stated that a patient can encrypt and transmit medical information to a third-party trusted cloud service to authorize doctors to access it, and they presented two Secure and Efficient Dynamic Searchable Symmetric Encryption (SEDSSE) schemes for solving patient privacy. Cano and Cañavate-Sanchez [17] addressed an approach with a dual signature in ECDSA (Elliptic Curve Digital Signature Algorithm), which can protect the privacy of patient data transmitted from IoMT (Internet of Medical Things) devices to cloud servers. Liu et al. [18] presented two distinct RSSs (redactable signature schemes) with flexible release control (RSSs-FRC) to protect the privacy of the release of authenticated medical documents, which proves to have great advantages in security and efficiency. Hamza et al. [19] proposed a chaotic encryption password system based on privacy protection to protect the privacy of patients, which can protect the patient's image from being compromised by the broker.

Although encryption can solve the issues of patient information privacy, it is not conducive to the sharing of case information and its disadvantage of promoting the development of medical technology. In addition, if the patient's secret key is lost and the medical data cannot be shared, it will be detrimental to the research of treatment plans. Therefore, we propose a privacy-protection model with technology of blockchain, homomorphic encryption, and federated learning, which will be discussed in Section 2.

The rest structure of this article is as follows. Section 2 describes related work, which includes blockchain, homomorphic encryption, and federated learning. Section 3 proposes a privacy model and describes its algorithm. Section 4 validates the model with experiments and discusses the results. Section 5 concludes this paper and proposes directions for future research.

2. Related Work

2.1. Blockchain. Blockchain technology originated from Bitcoin, and Satoshi Nakamoto uses blockchain technology to solve the problem of maintaining the order of transactions and avoid double spending in Bitcoin [20]. Crosby et al. [21] stated that the underlying blockchain technology of the digital currency Bitcoin has worked perfectly and is widely used in the financial and nonfinancial fields, and the core of the blockchain is the distributed ledger, decentralization, smart contracts, and consensus mechanism. Bashir [22] defined distributed ledger, decentralization, smart contracts, and consensus mechanism; distributed ledger means that the ledger is distributed across the entire network among all peers in the network, and each peer has a copy of the complete ledger; the basic idea of decentralization is to assign control and authority to the periphery of the organization, rather than having a central agency completely control the organization; the smart contract is a secure and uninterruptible computer program that represents an agreement that can be automatically executed and enforced; the consensus mechanism is a set of steps taken by most or all nodes in the blockchain to reach a consensus on the state or value of the proposal.

Blockchain has advantages in solving privacy security, so some researchers use blockchain to solve medical privacy issues. Liu et al. [23] presented a blockchain-based privacy-preserving data sharing (BPDS) for EMRs (Electronic Medical Record), and the patient predefined access authorization to share secure data automatically through the smart contract of the blockchain. Zhang and Lin [24] proposed a blockchain-based secure and privacy-preserving PHI (Personal Health Information) sharing (BSPP) scheme which based on blockchain to improve the diagnosis of e-health systems. Ji et al. [25] presented a blockchain-based multilevel privacy-preserving location sharing (BMPLS) scheme, which can realize the sharing of exclusive protected locations based on blockchain in the telemedicine information system. Al Omar et al. [26] proposed a patient-centric medical data management system by using blockchain as storage to obtain privacy, and they can ensure security by using encryption to protect patient data.

Although blockchain can solve the privacy of patients, there are still some problems in the sharing of medical records between different hospitals in different regions. Meanwhile, some medical institutions even strictly prohibit patient record sharing in order to protect patient information, which is not conducive to the development of medical technology. Therefore, this paper intends to use blockchain technology as the bottom layer and introduce homomorphic encryption and federated learning to facilitate data sharing between different hospitals in different regions, which can guarantee patient privacy.

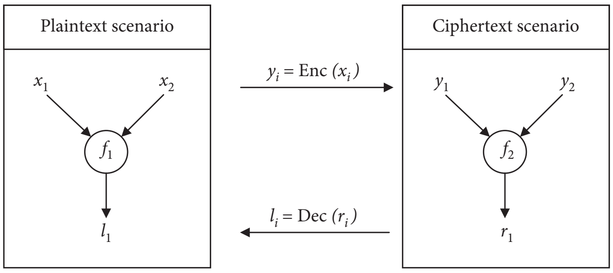

FIGURE 1: The process of homomorphic encryption. The function operations f_1 and f_2 are equivalent.

2.2. Homomorphic Encryption. Homomorphic encryption (HE) can perform meaningful calculations on encrypted data without decrypting the data, and the homomorphic encryption system can be used to encrypt the data to solve the privacy problem in a satisfactory manner [27]. Alloghani et al. [28] stated that homomorphic encryption can allow encrypted data to be calculated in the cloud server to avoid user privacy leakage. Acar et al. [29] discussed that the servers or the users with the key have high authority in the legacy encryption system, which leads to privacy and security issues, and homomorphic encryption (HE) is a solution of a special encryption scheme that can solve these issues for allowing any third party to operate on encrypted data without prior decryption.

Homomorphic encryption has good privacy-protection features, which is extremely suitable for sharing medical records, and some researchers apply homomorphic encryption to the medical field for privacy protection. Vengadapurvaja et al. [30] proposed an effective homomorphic encryption algorithm to encrypt medical images and perform useful operations on them without breaking confidentiality. Yang et al. [31] presented a safe and high-vision medical image framework to protect the privacy and security of medical images, and they adopted a novel Reversible Data Hiding (RDH) technology to embed private data into medical images and used homomorphic encryption which based on chaotic maps to ensure the security of medical images.

According to Acar et al. [29] and Gentry and Boneh [32], this paper describes the process of homomorphic encryption, which is shown in Figure 1.

As shown in Figure 1, both x_1 and x_2 are plaintext, and they can perform operation f_1 to get result l_1 on the client, which is no encryption operation in this process. In addition, $x_1$and x_2 are encrypted by function $y_i = \mathrm{Enc}(x_i)$ (encryption process) to obtain ciphertexts $y_1$and y_2, and the ciphertexts can be sent to third-party application scenarios (e.g., cloud computing) for operations (such as calculations and search), which the result is ciphertext r_1 after $c_1$and c_2 are operated by function f_2. Finally, after the ciphertext R_1 passes through the decryption function $l_i = \mathrm{Dec}(r_i)$(decryption process), the result is equal to the plaintext l_1, which is $l_1 = \mathrm{Dec}(r_1)$.

We assume that an encryption scheme G is represented as (M, C, K, E, D), where M is the plaintext space; C is the ciphertext space; K is the key space; E is the encryption algorithm; D is the decryption algorithm; and $\oplus$ is the ciphertext-related operator. Therefore, we define some definitions of homomorphic encryption as follows.

Definition 1. Assuming P and L are operations, when the plaintext dataset $M = \{m_1, m_2, \ldots, m_n\}$, $k \in K$ if it satisfies$P(E_k(m_1), E_K (m_2), \ldots, E_K(m_n)) = E_k(L(m_1, m_2, \ldots, m_n))$, then the encryption scheme is homomorphic for operation L. The basic idea of homomorphic encryption is to achieve the same effect as the corresponding plaintext operation by performing certain operations on multiple ciphertexts.

Definition 2. For any plaintext $m_i, m_j \in M$, the corresponding ciphertext is $c_i = E(m_i)$, $c_j = E(m_j)$ and $c_i, c_j \in C$. If $E(m_i + m_j) = E(m_i) \oplus E(m_j)$ or $D(E(m_i) \oplus E(m_j)) = m_i + m_j$ is true, then the encryption scheme G has additive homomorphism.

Definition 3. For any plaintext $m_i, m_j \in M$, the corresponding ciphertext is $c_i = E(m_i)$, $c_j = E(m_j)$, and $c_i, c_j \in C$. If $E(m_i \cdot m_j) = E(m_i) \oplus E(m_j)$ or $D(E(m_i) \oplus E(m_j)) = m_i \cdot m_j$ is true, then the encryption scheme G has multiplicative homomorphism.

Definition 4. For any plaintext $m_i, m_j \in M$, the corresponding ciphertext is $c_i = E(m_i)$, $c_j = E(m_j)$, and $c_i, c_j \in C$. If $E(m_i \cdot m_j) = E(m_i) \oplus m_j$ or $D(E(m_i) \oplus m_j) = m_i \cdot m_j$ is true, then the encryption scheme G has a homomorphic property with mixed multiplication.

Definition 5. If scheme G has both additive homomorphism and multiplication homomorphism properties and it can satisfy the finite numbers of addition and multiplication ciphertext operations, then the encryption scheme G is called a somewhat homomorphic encryption scheme.

Definition 6. If scheme G has both additive homomorphism and multiplication homomorphism properties and it can satisfy the any numbers of addition and multiplication ciphertext operations, then the encryption scheme G is called a fully homomorphic encryption scheme.

The homomorphic encryption process allows two or more different organizations to update the algorithm model without contact, which means that it can be performed on plaintext on encrypted data without decryption, and it can well protect user privacy date. Moreover, the above definitions will be applied in subsequent encryption algorithms, which are shown in Section 3.2.

With the technical support of the blockchain in the previous section, it can ensure that data can be safely and accurately transmitted in the communication channel. Meanwhile, homomorphic encryption can guarantee that users' private data participate in calculations or statistics in encrypted form, which avoids the risk of data leakage. In subsequent sections, this article will briefly introduce federated learning, which guides user privacy data from different institutions or organization on how to participate in training.

2.3. Federated Learning. Federated learning (FL) involves training statistical models on remote devices or isolated data centres with keeping data localized, which can protect user privacy as much as possible [33]. Li et al. [34] stated that federated learning enables multiple parties to jointly train machine learning models without exchanging local data, which covers technologies from multiple research fields, such as distributed systems, machine learning, and privacy. According to Li et al. [33, 34] and Yang et al. [35], we could draw an illustration of federated learning system, which is shown in Figure 2.

In Figure 2, federated learning is divided into central training and local training. The federated learning system identifies k clients, and each client trains its private dataset ρ_k in L-SGD (Local Stochastic Gradient Descent). Then, the final results of the entire dataset are aggregated to obtain the final training results in the central server. Next, the w_i is the trained parameter, which is calculated and generated by the client, and the central server is responsible for collecting parameter w_i and coordinating processing. Finally, the server will obtain the latest model parameters and assign them to each client for completing the parameter update of the entire federated learning system after all rounds of data training.

In addition, federated learning can be applied in the medical field to solve the privacy problem of patients. Xu et al. [36] surveyed the applications of federated learning in the medical field, and federated learning keeps sensitive data locally, which provides an effective solution for the data source and privacy protection of medical institutions. Sheller et al. [37] stated that federated learning can promote multiagency and cross-field cooperation without sharing medical data, which can well protect the privacy of patients and improve the development of precision medicine. Yang et al. [38] described federated learning into the following three categories, which are shown in Figure 3.

As shown in Figure 3, federated learning is divided into three categories by Yang et al., and the detailed explanation is as follows:

(1) *Horizontal Federated Learning*. When the features of different users of the two datasets more overlap, horizontal federated learning can extract the same feature data of different users for training. For instance, the user groups of the two banks in different regions have less intersections, but their businesses are very similar; therefore, the recorded user characteristics are the same, and horizontal federated learning can be used to construct a federated model. In this case, the bank can get a better training set or data result without uploading user information, which can protect user privacy.

(2) *Vertical Federated Learning*. Vertical federated learning is similar to horizontal federated learning, but is different in data feature extraction, where vertical federated learning extracts data of the same user in different dimensions. For example, two organizations with different businesses conduct data analysis on users in the same area, which can analyse the characteristic values of the same users to achieve better data results.

(3) *Federated Transfer Learning*. When the user characteristics and samples overlap in the two datasets are less, federated transfer learning can be used to solve the lack of data for analysis. For example, banks in two different countries have different user characteristics and samples, and federated transfer learning can be used for effective data analysis.

To sum up, federated learning can easily solve user data analysis problems without uploading user information, which can protect user privacy. In order to understand the application of federated learning in this article, we have made the following definition.

The training process can be regarded as an optimization problem in machine learning, and its formula can be expressed as follows:

$$\min_{\omega \in R^{d}} f(\omega), \tag{1}$$

$$f(\omega) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(\omega). \tag{2}$$

In formulas (1) and (2), $f(\omega)$ stands for the loss function and $f_i(\omega)$ is the loss corresponding to the prediction of the data point with index i. For client k, there are n_k training points in the dataset on client k, where $n_k = |\rho_k|$. Therefore, the optimization problem of federated learning can be reexpressed as

$$\begin{aligned} f(\omega) &= \sum_{k=1}^{k} \frac{n_k}{n} \cdot F_k(\omega), \\ F_k(\omega) &= \frac{1}{n_k} \sum_{i \in \rho_k} f_i(\omega). \end{aligned} \tag{3}$$

In addition, it is difficult for federated learning to satisfy the independent and identical distribution assumption (IID Assumption), in which the training data is uniformly and randomly distributed on each client.

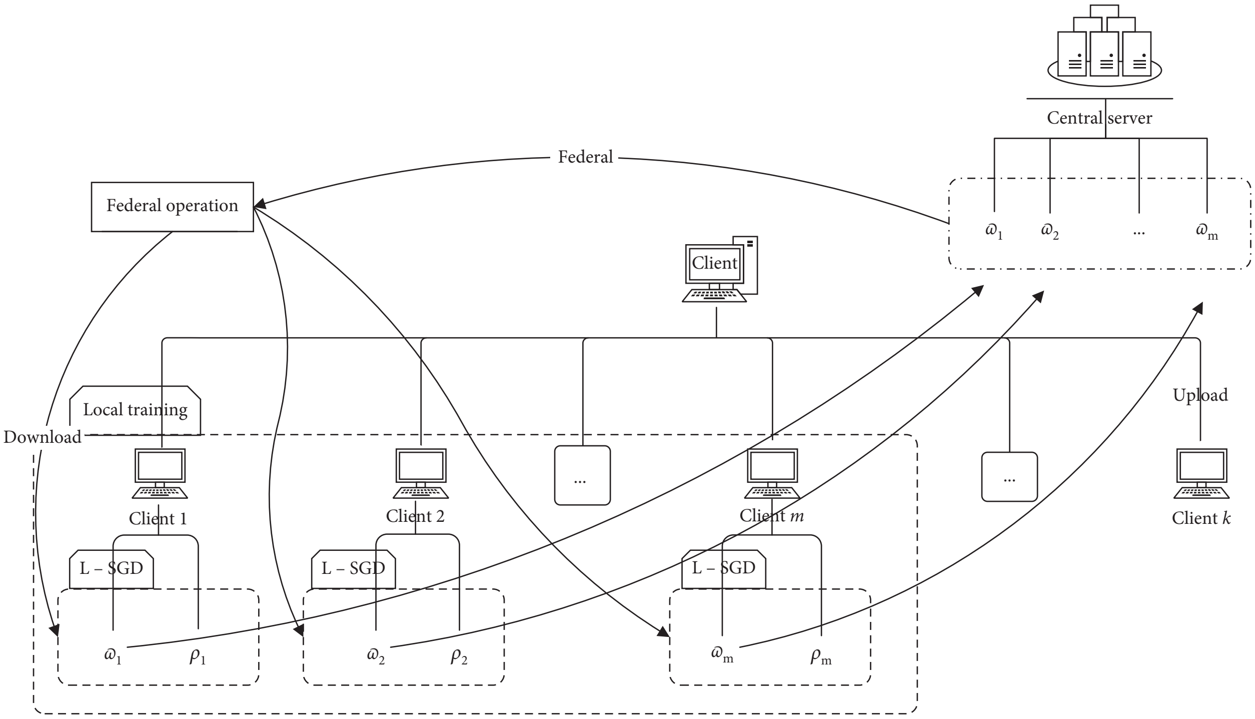

FIGURE 2: Federated Learning System. L-SGD means Local Stochastic Gradient Descent.

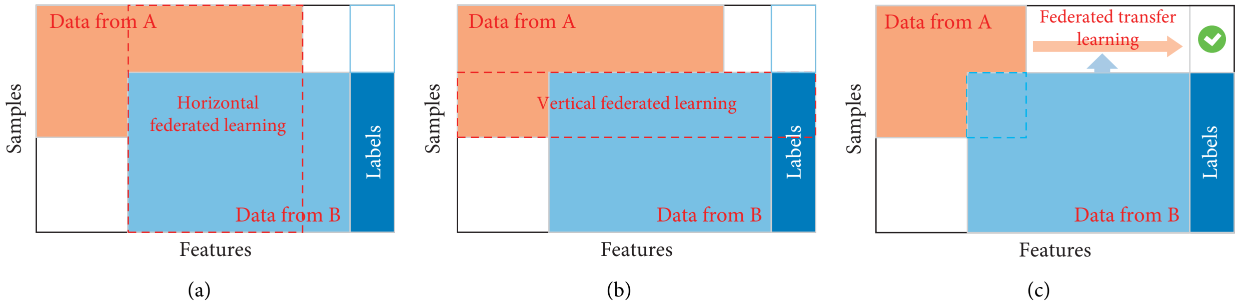

FIGURE 3: Categorization of federated learning. (a) Horizontal federated learning. (b) Vertical federated learning. (c) Federated transfer learning.

3. Models and Algorithms

3.1. Federated Learning Verification Model. This part will explain the design and deployment of the model and gives the following two assumptions.

Normal Distribution Assumption. Assume that each assumption sample is a D-dimensional normal distribution drawn from k samples, and each is $\text{Normal}\,(\mu_k, \sigma_k^2)$

Independence Assumption. Assuming that the two datasets are independent, for each $\text{Normal}\,(\mu_k, \sigma_k^2)$, satisfy $\text{Normal}\,(\mu_k, \sigma_k^2) \propto \text{Normal}\,(\mu 1_k, \sigma 1_k^2) \cdot \text{Normal}\,(\mu 2_k, \sigma 2_k^2)$

We established the following model scheme, which is based on the above assumptions, and the model scheme obeys the following steps:

(1) Establish a Gaussian mixture model based on the clustering problem, and use the EM (Expectation Maximization) algorithm to update the parameters.

(2) Build FL (federated learning) server. The main job of the server is to send the FL scheme to the clients, to receive the trained parameters, and to integrate the federated distribution tasks to the engineers for analysis.

(3) The clients preprocess the data. Each client preprocesses its own data, which uses encryption

algorithms to encrypt sensitive data and stores it in a fixed area.

(4) The server assigns tasks to the clients. At this stage, the server sends a signal to the client to propose the conditions required for the training plan, such as memory, container, and size of the collected data.

(5) The server receives the response from the clients. After the client receives the signal from the server, it responds to the server and returns information (e.g., data size, time, etc.) to the server for collecting data.

(6) The server initializes the parameters to the client. The server initializes the parameters: probability matrix ϕ, category probability θ, and $\mu 1_0$, $\sigma 1_0^1$, $\mu 2_0$, $\sigma 2_0^1$, ... $\mu 1_n$, $\sigma 2_0^1$.

(7) The server sends the training plan to the clients. For example, the server sends probability matrices ϕ, $\mu 1_0$, and $\sigma 1_0^1$ to client A and sends probability matrices ϕ, $\mu 2_0$, and $\sigma 2_0^1$ to client B.

(8) The clients start to train the data based on the requirement of the server. At the first step, client A locally updates $\mu 1_{t-1}, \sigma 1_{t-1}^2$ and the given probability matrix ϕ, which obtains an updated D1-dimensional normal distribution $\mathrm{Normal}(\mu 1_t, \sigma 1_t^2)$. Meanwhile, client B will get an updated D2-dimensional normal distribution $\mathrm{Normal}(\mu 2_t, \sigma 2_t^2)$ with the similar steps.

(9) After finishing the training, the result will be sent to the server (e.g., client A sent $\mu 1_t$ and $\sigma 1_t^2$, and client B sent $\mu 2_t$ and $\sigma 2_t^2$), and data is encrypted by Paillier, which is a homomorphic encryption algorithm.

(10) The server updates the parameters from the client and starts a new cycle towards the end.

We built a federated learning verification system for conducting subsequent simulation experiments, which is shown in Figure 4.

In Figure 4(a), it is a federated model of data alignment, which is a local client training model, and this model makes sure that privacy data cannot be exchanged or updated. In Figure 4(b), we have designed a vertical federated learning model, which conducted "noninteractive" connection for data platforms with similar samples and different indicator dimensions, and the model can realize the collaborative calculation of the expanded sample size with similar index data, which can improve the overall safety of the system and the integrity and comprehensiveness of the analysis results. Next, the system is divided into the user layer and edge service layer. The user layer is composed of Internet of Things (IoT) equipment, mobile terminals, etc., and the server layer is equipped with mobile edge computing servers, which are composed of base stations with storage and computing capabilities. Meanwhile, the federated learning local training runs on the client of user side (e.g., Agency A and Agency B), which learns the local model parameters with based on the data of user side.

In addition, we have designed an operation flowchart of the system, and it helps the following experimental work, which is shown in Figure 5. As shown in Figure 5, it is an operation flowchart of the system, which is from Figure 4. According to Figures 4 and 5, the operation of the model system is divided into two parts: client and server.

(1) The client participates in local data training, which uses an algorithm based on gradient descent to find model parameters for minimizing the loss function.

(2) The server collects the model parameters with being trained by the client, which collects parameters from various clients and federated model parameters and updates the overall model, and the server sends the new model to the clients for starting a new round of training and learning.

As shown in Figure 5, the client will select a model and an agency from the server for starting, which follows the system in Figure 4. Next, if the client contains an existing model, it will update the model from the selected model, or it will use the selected model. Moreover, if the client needs to update the model, it will request communication to train the model, or it will send the existing model to the central server for obtaining results.

In the process of data training, the user privacy data will not be shared or calculated in the form of plain text, and they will be securely encrypted in the communication channel by using blockchain and homomorphic encryption technology. In addition, the client does not need to upload or share the private data of the local user and gets the final result after the training of the model. Therefore, our model shows that it can protect user privacy data.

3.2. A FL-EM-GMM Algorithm.

In the previous section, we designed a federated learning verification model, which will be made an experiment in the next section. In this section, we have completed the model algorithm, which is FL-EM-GMM Algorithm (Federated Learning Expectation Maximization Gaussian Mixture Model Algorithm), and it is shown in Algorithm 1.

As shown in Algorithm 1, it is a FL-EM-GMM algorithm, which displays the detail of federated learning process. At the server steps, the algorithm will complete the initialization work which is based on the client's data, and it will assign task parameters to the client. Next, at the client steps, the algorithm will finish the training process which is based on server parameters and local model parameters to obtain the best training dataset, and it will share the trained parameters to the server. Finally, the clients will update the latest model according to the server's parameters after the

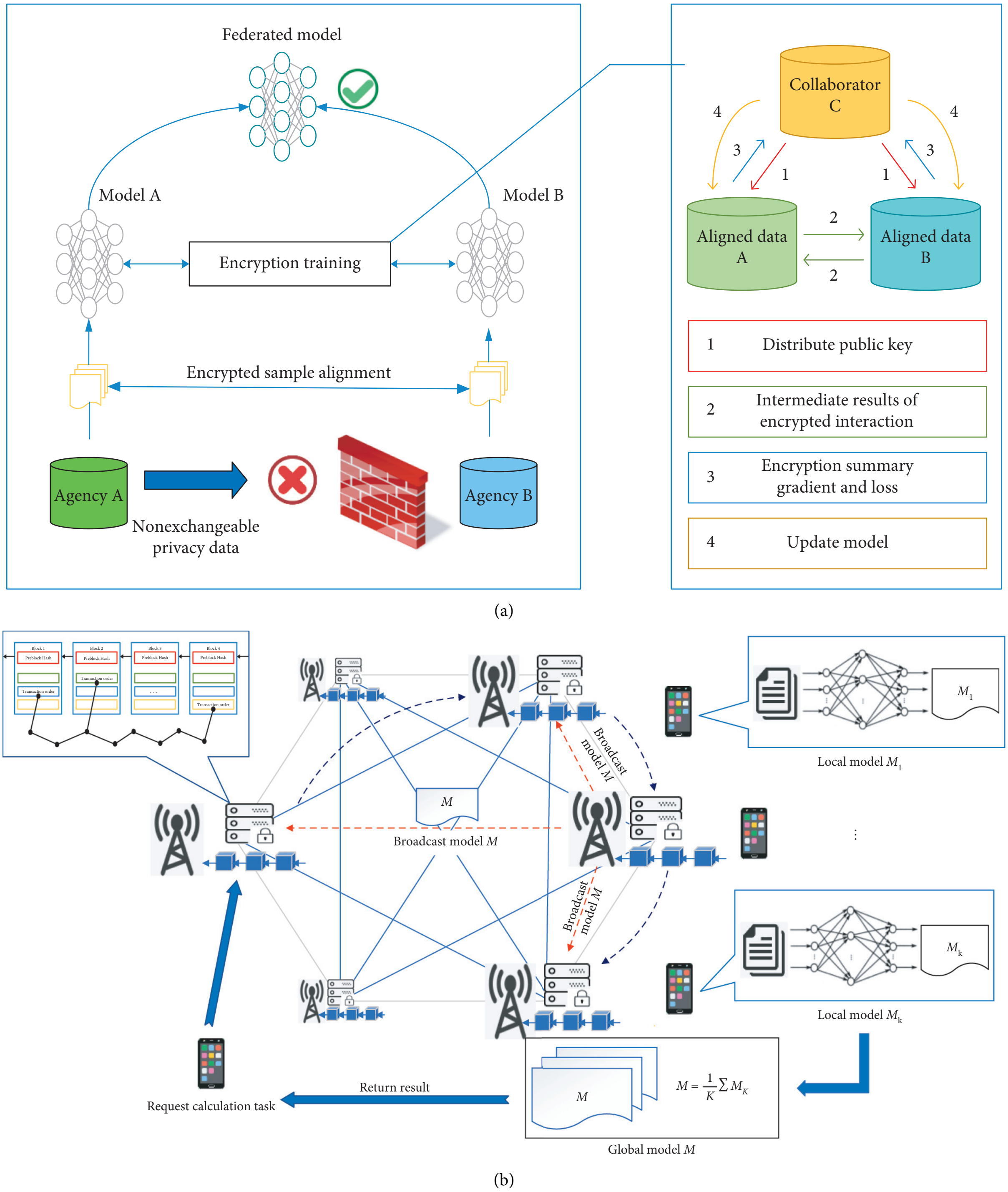

Figure 4: (a) Federated model of data alignment. (b) Federated learning verification model system.

cycle ends, and the parameters are trained in local, which avoids the risk of privacy data leakage.

4. Experiment and Simulation

In experiment, we used FATE official components [39] and three servers for testing, and the data source is provided by FATE official, which is breast cancer patients' data (<569, 30>). Next, we use the previously designed model and system to ensure that privacy data is not leaked, which achieves open patients' records and collaborative diagnosis. The experiment deploys FATE on two machines, and each machine runs a FATE instance. In order to judge whether the patient is benign or malignant, the basic information of the patient is obtained from the database, such as ID number, standard deviation of grey values, and average size

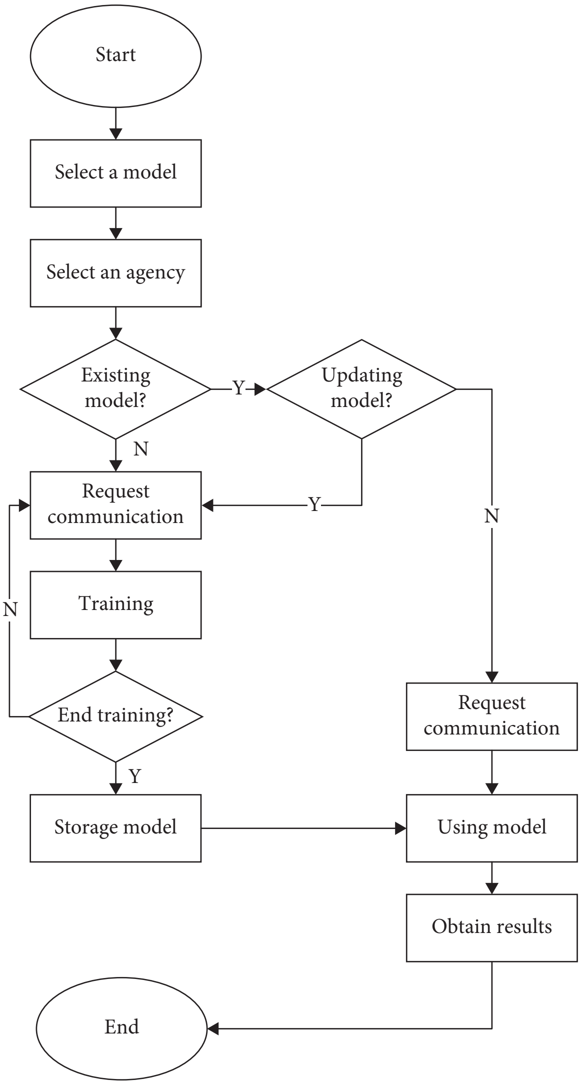

Figure 5: An operation flowchart of system.

of core tumour. While uploading these source data to the server, the server will normalize and process them to desensitize the data and enter it into the system. In data analysis, some key terms are as follows:

(1) *Feature selection*. Feature selection is to find the correlation between X and Y labels, which is called IV (Information Value). Since not every X and Y are correlated, we exclude the X with lower correlation for reducing noise.

(2) *Feature binning*. As the normalized data, most of them become very small values, so the regressed model is too sensitive to the requirements of parameter accuracy. Therefore, we perform simple binning, which change the continuous value into several segment types and make the data into 1–10 (10 discrete values of labels), and it helps to improve variable expression ability and model discrimination.

The feature binning parameters are shown in Table 1.

In Table 1, WoE means Weight of Evidence and IV means Information Value. Meanwhile, according Table 1, we draw Figures 6 and 7.

In Figure 6, we have differentiated and compared the data of Event Count and Nonevent Count, and it shows that the value is related to the choice of feature binning parameters. Figure 6 shows that, as the number of training increases, the Event Count has decreased, which means the parameters are effectively trained.

Meanwhile, we draw Figure 7 based on WoE (Weight of Evidence) in Algorithm 1. The greater the WoE value is, the greater the possibility that the cancer is benign. Conversely, the smaller the WoE value is, the greater the possibility that the cancer is malignant. Figure 7 indicates that cancer patients can be screened by the value of WoE for better disease prevention and achieve the goal of precision medicine.

```
Input: Data $x$ and $y$ from dataset A and dataset B, number of clusters K.
Output: GMM parameters $\theta$, $\mu$ and $\sigma^2$ clusters assignment distribution $\phi_{\mathrm{i}}$.
The server initializes parameters$\theta$, $\mu_0$ and$\sigma_0^2$.
for $t = 1, \ldots, T$//Iteration
  for $i = 1, \ldots, N$
    for $j = 1, \ldots, K$
      //#The server:
      //At E-step, calculate the formula below.
      $\phi_i(j)_{(t)} = \theta_{j_{(t-1)}} \cdot \mathrm{Normal}(x_i|\mu 1_{j_{(t-1)}}, \sigma 1^2_{j_{(t-1)}}) \cdot \mathrm{Normal}(y_i|\mu 2_{j_{(t-1)}}, \sigma 2^2_{j_{(t-1)}}) / \sum_{k=1}^{K} \pi_{k_{(t-1)}} \mathrm{Normal}(x_i|\mu 1_{k_{(t-1)}}, \sigma 1^2_{k_{(t-1)}}) \cdot$
      $\mathrm{Normal}(y_i|\mu 2_{k_{(t-1)}}, \sigma 2^2_{k_{(t-1)}})$
      //At $M$-step, calculate the formula below.
      $n_{j_{(t)}} = \sum_{i=1}^{n} \phi_i(j)_{(t)}$
      $\theta_{j_{(t)}} = n_{j_{(t)}}/n$
      //# Client A: At $M$-step, calculate the formula below, the output parameter data is encrypted with partial homomorphic
      cryptographic algorithm paillier.
      $A(\mu_{j_{(t)}}) = 1/n_{j_{(t)}} \sum_{i=1}^{n} \phi_i(j)_{(t)} x_i$
      $A(\sigma^2_{j_{(t)}}) = (1/n_{j_{(t)}} \sum_{i=1}^{n} \phi_i(j)_{(t)} (x_i - \mu_{i(t)})(x_i - \mu_{i(t)})^T)^{-1}$
      //# Client B: At $M$-step, calculate the formula below, the output parameter data is encrypted with partial homomorphic
      cryptographic algorithm paillier.
      $B(\mu_{j_{(t)}}) = 1/n_{j_{(t)}} \sum_{i=1}^{n} \phi_i(j)_{(t)} y_i$
      $B(\sigma^2_{j_{(t)}}) = (1/n_{j_{(t)}} \sum_{i=1}^{n} \phi_i(j)_{(t)} (y_i - \mu_{i(t)})(y_i - \mu_{i(t)})^{T-1})$
    end for
  end for
end for
```

Algorithm 1: FL-EM-GMM algorithm.

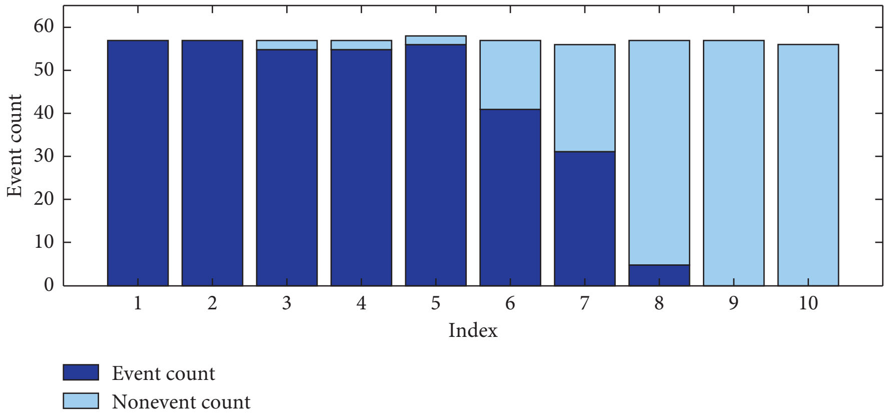

Figure 6: Data difference of event count and nonevent count.

We have divided the machines into guest and host for data privacy protection, and guest machine is the data application side, while host machine is the data provider. The guest machine uses the IV (Information Value) to check the binning effect, but it cannot determine the characteristic values X and Y, which can protect patient's privacy. We obtained a set of values after training, and we only displayed the first 10 results for the reason that there are too many training results, which are shown in Table 2.

In Table 2, "predict_score" is the probability of logistic regression prediction output. When the probability of "predict_score" is greater than 0.5, "predict_result" is 1, or it is 0. In addition, when "predict_result" is 0, the result is benign, and when it is 1, the result is malignant. According

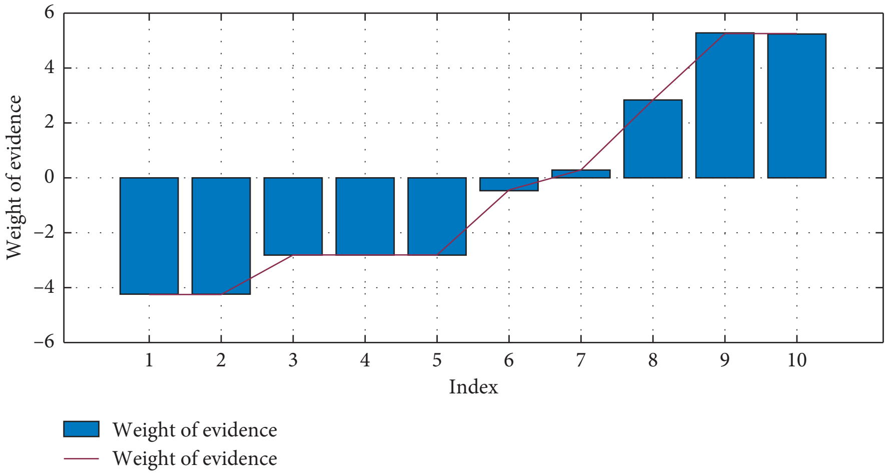

Figure 7: Weight of evidence.

Table 1: The result of feature binning parameters.

Index	Binning	IV	WoE	Event_Count	Event_Ratio	Non_Event_Count	Non_Event_Ratio
1	$x0 \leq -1.047670$	0.670339	−4.223783	57	0.161064	0	0.002358
2	$-1.047670 < x0 \leq -0.784675$	0.670339	−4.223783	57	0.161064	0	0.002358
3	$-0.784675 < x0 \leq -0.612797$	0.403950	−2.793036	55	0.154062	2	0.009434
4	$-0.612797 < x0 \leq -0.469910$	0.403950	−2.793036	55	0.154062	2	0.009434
5	$-0.469910 < x0 \leq -0.269040$	0.414430	−2.811055	56	0.156863	2	0.009434
6	$-0.269040 < x0 \leq -0.053674$	0.016531	−0.419834	41	0.114846	16	0.075472
7	$-0.053674 < x0 \leq 0.232100$	0.009515	0.306038	31	0.086835	25	0.117925
8	$0.232100 < x0 \leq 0.840923$	0.662137	2.862955	5	0.014006	52	0.245283
9	$0.840923 < x0 \leq 1.536720$	1.420925	5.266082	0	0.001401	57	0.271226
10	$X0 > 1.536720$	1.391434	5.248537	0	0.001401	56	0.266509

Table 2: The result of training.

Index	Label	Predict_result	Predict_score	Predict_detail	Type	Other
1	0	0	0.078641	{‘0’: 0.9213587838889536	‘1’: 0.07864121611104637}	Train
2	0	0	0.242046	{‘0’: 0.7579539730347816	‘1’: 0.2420460269652184}	Train
3	0	0	0.295851	{‘0’: 0.7041491549790799	‘1’: 0.29585084502092}	Train
4	0	1	0.953959	{‘0’: 0.046041144486783373	‘1’: 0.9539588551321663}	Train
5	0	1	0.598822	{‘0’: 0.401177819540372	‘1’: 0.598822180459628}	Train
6	0	1	0.980817	{‘0’: 0.019183466307048036	‘1’: 0.980816533692952}	Train
7	0	0	0.049818	{‘0’: 0.95018217382329	‘1’: 0.04981782617671007}	Train
8	1	0	0.003503	{‘0’: 0.9964965574508085	‘1’: 0.00350344254919155}	Train
9	1	0	0.02129	{‘0’: 0.978709874597958	‘1’: 0.021290125402041987}	Train
10	0	0	0.35347	{‘0’: 0.6465295143293885	‘1’: 0.3534704856706114}	Train

to statistics, the proportion of “1” is 60.61%, and the proportion of “0” is 39.39% during this experiment.

5. Conclusions

The emergence of federated learning provides new ideas for artificial intelligence to break the data barrier and further develop, and it allows data owners to jointly establish a common model with the premise of protecting local data, which solves the privacy and data security issues of medical institutions. In order to solve the problem of privacy leakage of medical patients, we designed a federate learning verification model and an FL-EM-GMM Algorithm (Federated Learning Expectation Maximization Gaussian Mixture Model Algorithm), and we have verified the algorithm, which shows that medical privacy data training can be completed in local to avoid the risk of data leakage. Moreover, we conducted experiments on the federated task of datasets from two organizations in our model system, in which the data has the same sample ID with different subset features, and this system is capable of handling privacy and security issues. The results show that the model trained by our system has better usability, security, and higher efficiency, compared with the model trained by traditional machine learning methods.

In the future, we will build an improved federated learning system with customizable password algorithms, which will make the federated learning system more flexible and easier to implement. Meanwhile, we will improve the compatibility of homomorphic encryption algorithms and federated learning systems to make them more efficient and stable.

Data Availability

In the experiment, we used FATE official components and three servers for testing, and the data source is provided by FATE official, which is breast cancer patients' data, in https://github.com/FederatedAI/FATE.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to express appreciation for the support of the Construct Program of Applied Characteristic Discipline and the Research Project (Project No. 19XKY053) in Hunan University of Science and Engineering, Key Research and Development Project of Hunan Provincial Science & Technology Department (Project No. 2017NK2390), and Yongzhou Guiding Science and Technology Program (Project No. 2019-yzkj-10).

References

[1] J. J. Hathaliya and S. Tanwar, "An exhaustive survey on security and privacy issues in healthcare 4.0," *Computer Communications*, vol. 153, pp. 311–335, 2020.

[2] M. Prakash and G. Singaravel, "An approach for prevention of privacy breach and information leakage in sensitive data mining," *Computers & Electrical Engineering*, vol. 45, pp. 134–140, 2015.

[3] M. Papaioannou, M. Karageorgou, G. Mantas et al., "A survey on security threats and countermeasures in internet of medical things (IoMT)," *Transactions on Emerging Telecommunications Technologies*, vol. e4049, 2020.

[4] B. Bernd, D. M. Lopez, and C. Gonzalez, "Patient privacy and security concerns on big data for personalized medicine," *Health and Technology*, vol. 6, no. 1, pp. 75–81, 2016.

[5] T. Hulsen, S. S. Jamuar, A. R. Moody et al., "From big data to precision medicine," *Frontiers in Medicine*, vol. 6, no. 34, 2019.

[6] C.-A. Azencott, "Machine learning and genomics: precision medicine versus patient privacy," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2128, Article ID 20170350, 2018.

[7] H. Gao, L. Kuang, Y. Yin, B. Guo et al., "Mining consuming behaviors with temporal evolution for personalized recommendation in mobile marketing apps," *ACM/Springer Mobile Networks and Applications (MONET)*, vol. 25, no. 4, pp. 1233–1248, 2020.

[8] X. Yang, S. Zhou, and M. Cao, "An approach to alleviate the sparsity problem of hybrid collaborative filtering based recommendations: the product-attribute perspective from user reviews," *Mobile Networks and Applications*, vol. 25, no. 2, pp. 376–390, 2020.

[9] W. N. Price, I. G. Cohen, and I. Glenn Cohen, "Privacy in the age of medical big data," *Nature Medicine*, vol. 25, no. 1, pp. 37–43, 2019.

[10] H. Gao, C. Liu, Y. Li et al., "V2VR: reliable hybrid-network-oriented V2V data transmission and routing considering RSUs and connectivity probability," *IEEE Transactions on Intelligent Transportation Systems*, vol. 99, pp. 1–14, 2020.

[11] H. Gao, W. Huang, and Y. Duan, "The cloud-edge based dynamic reconfiguration to service workflow for mobile ecommerce environments: a QoS prediction perspective," *ACM Transactions on Internet Technology*, vol. 2020, Article ID 8843584, 14 pages, 2020.

[12] X. Ma, H. Gao, H. Xu et al., "An IoT-based task scheduling optimization scheme considering the deadline and cost-aware scientific workflow for cloud computing," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019249 pages, 2019.

[13] D. R. E. Gnad, J. Krautter, and M. B. Tahoori, "Leaky noise: new side-channel attack vectors in mixed-signal IoT devices," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, vol. 3, pp. 305–339, 2019.

[14] H. Tian, J. He, and Y. Ding, "Medical data management on blockchain with privacy," *Journal of Medical Systems*, vol. 43, no. 2, p. 26, 2019.

[15] W. Sun, Z. Cai, Y. Li et al., "Security and privacy in the medical internet of things: a review," *Security and Communication Networks*, vol. 2018, Article ID 5978636, 9 pages, 2018.

[16] H. Li, Y. Yang, Y. Dai et al., "Achieving secure and efficient dynamic searchable symmetric encryption over medical cloud data," *IEEE Transactions on Cloud Computing*, vol. 991 page, 2017.

[17] M. D. Cano and A. Cañavate-Sanchez, "Preserving data privacy in the internet of medical things using dual signature ECDSA," *Security and Communication Networks*, vol. 2020, Article ID 4960964, 9 pages, 2020.

[18] J. Liu, J. Ma, Y. Xiang et al., "Authenticated medical documents releasing with privacy protection and release control," *IEEE Transactions on Dependable and Secure Computing*1 page, 2019, https://ieeexplore.ieee.org/document/8611220.

[19] R. Hamza, Z. Yan, K. Muhammad, P. Bellavista, and F. Titouna, "A privacy-preserving cryptosystem for IoT E-healthcare," *Information Sciences*, vol. 527, pp. 493–510, 2020.

[20] S. Nakamoto, "Bitcoin: a peer-to-peer electronic cash system," *Manubot*, 2019, https://git.dhimmel.com/bitcoin-whitepaper/.

[21] M. Crosby, P. Pattanayak, S. Verma et al., "Blockchain technology: beyond bitcoin," *Applied Innovation*, vol. 2, no. 6–10, 2016.

[22] I. Bashir, *Mastering Blockchain: Distributed Ledger Technology, Decentralization, and Smart Contracts Explained*, Packt Publishing Ltd, Birmingham, UK, 2018.

[23] J. Liu, X. Li, L. Ye et al., "BPDS: a blockchain based privacy-preserving data sharing for electronic medical records," in *Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM)*, December 2018.

[24] A. Zhang and X. Lin, "Towards secure and privacy-preserving data sharing in e-health systems via consortium blockchain," *Journal of Medical Systems*, vol. 42, no. 8, p. 140, 2018.

[25] Y. Ji, J. Zhang, J. Ma et al., "BMPLS: blockchain-based multi-level privacy-preserving location sharing scheme for telecare medical information systems," *Journal of Medical Systems*, vol. 42, no. 8, p. 147, 2018.

[26] A. Al Omar, M. S. Rahman, A. Basu et al., "Medibchain: a blockchain based privacy preserving platform for healthcare data," in *Proceedings of the International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage*, pp. 534–543, Springer, Cham, Switzerland, December 2017.

[27] L. Zhang, Y. Zheng, and R. Kantoa, "A review of homomorphic encryption and its applications," in *Proceedings of the 9th EAI International Conference on Mobile Multimedia Communications*, pp. 97–106, Xi'an China, June 2016.

[28] M. Alloghani, M. Alani, D. Al-Jumeily et al., "A systematic review on the status and progress of homomorphic encryption technologies," *Journal of Information Security and Applications*, vol. 48, Article ID 102362, 2019.

[29] A. Acar, H. Aksu, A. S. Uluagac, and M. Conti, "A survey on homomorphic encryption schemes," *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–35, 2018.

[30] A. M. Vengadapurvaja, G. Nisha, R. Aarthy, and N. Sasikaladevi, "An efficient homomorphic medical image encryption algorithm for cloud storage security," *Procedia Computer Science*, vol. 115, pp. 643–650, 2017.

[31] Y. Yang, X. Xiao, X. Cai, and W. Zhang, "A secure and high visual-quality framework for medical images by contrast-enhancement reversible data hiding and homomorphic encryption," *IEEE Access*, vol. 7, pp. 96900–96911, 2019.

[32] C. Gentry and D. Boneh, *A fully homomorphic encryption scheme*, Stanford University, Stanford, CA, USA, 2009.

[33] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[34] Q. Li, Z. Wen, Z. Wu et al., "A survey on federated learning systems: vision, hype and reality for data privacy and protection," 2019, https://arxiv.org/abs/1907.09693.

[35] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, 2019.

[36] J. Xu, S. Benjamin, S. Chang et al., "Federated learning for healthcare informatics," 2019, https://arxiv.org/abs/1911.06270.

[37] M. J. Sheller, B. Edwards, G. A. Reina et al., "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[38] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, pp. 1–207, 2019.

[39] https://github.com/FederatedAI/FATE.