Federated Diabetes Mellitus Analysis via Homomorphic Encryption

Yun Xie¹, Peng Li², Xiang Zhu³, Qiuling Wu⁴
¹Nanjing University of Science and Technology ZiJin College, Nanjing 210023, China
²College of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
³College of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
⁴School of Software Engineering, Jinling Institute of Technology, Nanjing 211169, China

Abstract—Diabetes is one of the most extensive chronic diseases in the world. Most patients suffer from the disease and its complications for a long time due to the lack of accurate and standardized treatment at early stage. Therefore, by analyzing diabetes data and establishing relevant predictive models, it is very meaningful to give reasonable health advice to high-risk groups. The establishment of an accurate prediction model requires a large number of data sources as support, and multiple medical institutions participate in data contribution and collaborative learning. Federated learning provides a secure general architecture for distributed collaborative learning. However, in the process of federated learning, the data of participants is still subject to the risk of security attacks or indirect information leakage. For example, when a participant uploads the local model parameters to an honest but curious cloud server, the cloud server can obtain relevant information from the participant. In order to solve this problem, this paper proposes a federated forest algorithm based on homomorphic encryption to strengthen the protection of the data privacy of the participants while ensuring that the accuracy of data analysis does not decrease. Analysis proves that our algorithm has good performance in privacy protection and prediction accuracy.

1. INTRODUCTION
Diabetes Mellitus (DM for short) is one of the most common chronic diseases, which is mainly characterized by high blood sugar. At the same time, DM has obvious family genetic characteristics. Nearly half of diabetes patients have a family genetic history, which can be divided into 1 Type 2 diabetes (T1D) and type 2 diabetes (T2D) [1]. Because it is difficult to self-healing and related complications can bring serious effects, the early prevention and mid-term treatment of diabetes has gradually become the focus of widespread and continuous attention. In disease prediction and type diagnosis, many potential factors that affect the results, such as the weight, blood sugar, blood pressure, lifestyle, physical activity, eating habits, and family inheritance of the tested person, have brought technical difficulties to the diagnosis.
In recent years, the combination of machine learning and medical diagnosis has created a hot field of "smart medicine". However, the current level of smart medical care is far from reaching the true "smart" level. This is because obtaining a highly accurate model for the diagnosis of diseases (for example, DM) need a large number of medical data sources firstly. In reality, medical data is generally privately owned by many organizations, and due to the protection of patients' personal privacy information, as well as from legal norms and ethical constraints, medical data cannot be published to the public directly, which is commonly referred to as "data islands", which restricts the performance of the model. If more medical institutions can unite, contribute their data, and collect a sufficiently large data, the training effect of the machine learning model will be a qualitative breakthrough.

Considering how to build a machine learning model based on data sets distributed in various medical institutions, Google first proposed the federated learning [2, 3, 4] to ensure that multiple parties do not need to directly exchange data during model training to prevent data leakage. It provides a relatively safe general framework for collaborative learning. There are two important entities in federated learning, participants and cloud servers. Assuming that both parts are honest and curious, a member attack from a certain participant cannot be ruled out. On the other hand, when participants send local model parameters to cloud servers, local data information may be leaked to honest but curious servers.

Therefore, it is still necessary to further study how to improve the security of federated learning and strengthen the data privacy protection for participants.

The contributions of this paper are as follows:

1) We propose a federated learning security model based on fully homomorphic encryption for predictive diagnosis of diabetes in the context of multi-party collaboration.

2) We proposed the Federal Forest Algorithm to further improve the accuracy of diabetes classification, while protecting the privacy of local data of the participants.

3) We have conducted a comprehensive analysis of the safety and performance of the system.

2.RELATED WORK

In recent years, researchers have gradually and frequently explored the field of diabetes prediction research, and have created and studied many algorithms and toolkits [5, 6], such as traditional machine learning algorithms, ensemble learning methods, and association rule learning methods.

To improve the accuracy of classification, BM Patil [7] proposed a hybrid predictive model (HPM), using Simple K-means clustering algorithm to verify the selected class labels in the given data, and by applying the result set the classification algorithm builds a predictive model. Besides, B.M. Patil used the sensitivity and specificity performance measurement methods commonly used in medical classification to improve the verification process of existing experiments in the WEKA platform. Aliza Ahmad [8] compared the prediction accuracy of multi-layer perceptual neural networks and tree-based algorithms (especially ID3 and J48 algorithms). The experimental results showed that the pruned J48 tree (89.3%) was better than multi-layer perceptual Neural network (81.9%) can achieve higher accuracy. Marcano-Cedeño [9] proposed a multi-layer perceptron (AMMLP) as an artificial metaplasticity of a diabetes prediction model to further improve the accuracy. Vijayan V. [10] found that different pre-processing techniques have a certain degree of influence on the accuracy of predicting DM, and analyzed pre-processing methods such as principal component analysis (PCA) and discretization methods.

Some existed research work was mainly to formulate a series of classification prediction standards for high-risk groups who may have potential risk factors for diabetes. Chandrakar and Saini [11] proposed to use the Indian Weighted Diabetes Risk Score (IWDRS) as a diabetes screening tool to solve the undetected pre- and late-stage diagnosis of diabetes. Longfei Han et al. [12] proposed the paired and size-restricted K-means (PSCKmeans) method to screen high-risk groups of DM, which provides a tool for the risk stratification of clinical diseases. Qian Zhu [13] studied the data set of diabetic patients including potential risk characteristics of re-seeing a doctor, and used three different machine learning algorithms (Logistic regression algorithm, Random forest algorithm and improved
Random forest algorithm) to conduct experimental analysis. The risk characteristic factors are identified and ranked according to importance.

In terms of risk prediction of type 2 diabetes, Dr. Wei Zhe [14] researched the risk prediction analysis of type 2 diabetes based on the frequent pattern growth (FP-growth) algorithm and found that the FP-growth algorithm can mine more representative potential risk factors for diabetes. Guo Yirui [15] explored the application of artificial neural network models in the risk prediction of type 2 diabetes and used the receiver operating characteristic curve (ROC) to evaluate the test performance of the predictive model. The experimental results show that the ANN algorithm can achieve better experimental results than the Logistic regression algorithm. Wang Xidan et al. [16] discussed the specific application value of deep belief networks (DBN) in diabetes risk prediction.

The above-related works have completed some basic research on the algorithm comparison and model establishment of diabetes prediction. However, most of the research work are only for a single specific data set and cannot be adapted to various other similar data sets, nor does it consider the situation of multi-party data sets. Therefore, the prediction accuracy and data validity are not enough for practical applications in larger scenarios, and the diagnosis process cannot enjoy the benefits brought by the joint diagnosis of multiple institutions under the background of big data. In addition, security issues have not received enough attention.

3. KNOWLEGE PREPARATION

3.1. Homomorphic Encryption

Homomorphic encryption [17] was first introduced by Rivest shortly after the invention of RSA. Homomorphic encryption enables operations on plaintexts to be performed on their respective ciphertext without disclosing the plaintexts [18].

Definition 1 (Additively Homomorphic Encryption): Let $E$ be an asymmetric encryption scheme, the public key is $pk$, and the private key is $sk$. If it is possible to achieve any number of homomorphic addition and multiplication operations at the same time without the need for private key decryption, it is called a fully homomorphic encryption scheme.

The public key addition homomorphic encryption (PHE) scheme consists of the following (possibly probabilistic) poly-time algorithms.

- $\text{ParamGen}(\lambda) \rightarrow \text{PP}$: $\lambda$ is the security parameter and the public parameter $\text{PP}$ is implicitly provided in the following algorithm.
- $\text{KeyGen}(\lambda) \rightarrow (pk, sk)$: KeyGen generates the key pairs $(pk, sk)$ in which $pk$ is the public key and $sk$ is the private key, according to the security parameter $\lambda$.
- $\text{Enc}(pk, m) \rightarrow c$: Enc encrypts the given message $m$ with $pk$ and obtain the corresponding ciphertext $c$.
- $\text{Dec}(sk, c) \rightarrow m$ : Dec returns message $m$ encrypted in $c$ using the private key $sk$.
- $\text{Add}(c, c')$: Foroun algorithm when inputting ciphertexts $c$ and $c'$, the output is the encryption of plaintext addition $c_{\text{add}}$.
- $\text{DecA}(sk, c_{\text{add}})$: DecA algorithm uses $sk$ decrypting $c_{\text{add}}$ to obtain an addition of plaintexts.

Definition 2 (CPA Security): Regarding the PHE scheme in Definition 2, we consider a game between an opponent $A$ and a challenger $B$ as follows:

- Setup. The challenger $B$ runs $\text{ParamGen}$ algorithm and obtain key pairs $(pk, sk)$. Then the public parameter $pp$ and the public key $pk$ are sent to $A$. 

• Challenge. \textit{A} chooses two plaintexts \(m_0\) and \(m_1\) of the same length and submits them to the challenger \(B\) who in succession takes \(b \in \{0,1\}\) randomly and computes \(C^* = \text{Enc}(pk,m_b)\). The challenge ciphertext \(C^*\) is fed back to \(A\), who then produces a bit \(b'\).

If the advantage in Equation (1) is negligible in \(\lambda\), the PHE scheme can resist selected plaintext attacks (CPA security) [19].

\[
\text{Adv}^\text{PHE\_CPA}(\lambda) = \left| \text{Pr}[b' = b] - \frac{1}{2} \right|
\]

No bit of information is leaked from cipher texts when the PHE scheme can against chosen plaintext attacks (CPA security).

3.2. Random forest

Random forest [20] is a strong classifier. The essence is to use multiple decision tree algorithms to train the sample data set and complete the prediction of the entire data set through the model. A single decision tree algorithm is equivalent to a weak classifier. The random forest needs to complete the voting selection for the classification results of the weak classifier, and the output of the final result needs to follow the principle of minority to majority.

Therefore, the classification result with the highest voting score in the output category determines the final classification result. In large data sets, the use of a single decision tree algorithm will face a large enough sample size, which leads to overfitting of the predicted classification model.

3.3. Federated Learning

Define multiple local data owners \(L_i (i = 1, \cdots, N)\) whom all wish to train a machine learning model by consolidating their respective data \(D_i (i = 1, \cdots, N)\). A conventional method is to aggregate all the data \(D = \{D_1, \cdots, D_N\}\) and use it to train the global model \(M_{\text{global}}\). However, due to legal issues such as privacy and data security, this solution cannot be implemented.

\textbf{Definition 3(Federated Learning)}[21]: A federated learning system is a learning process in which data owners jointly train a global model \(M_{\text{sum}}\). In this process, no data owner will disclose their data \(D_i\) to others.

In addition, the actual accuracy \(V_{\text{sum}}\) of \(M_{\text{sum}}\) should be very close to the expected accuracy \(V_{\text{global}}\) of \(M_{\text{global}}\). Let \(\delta\) be a non-negative real number, if \(V_{\text{sum}}\) and \(V_{\text{global}}\) have the following relationship:

\[
|V_{\text{sum}} - V_{\text{global}}| < \delta,
\]

the federated learning algorithm has \(\delta\)-accuracy loss.

\textbf{Definition 4(Horizontal Federated Learning)}[21]: Horizontal federated learning (also known as sample-based joint learning) is introduced, in the case of data sets sharing the same feature space but different samples.

4. SYSTEM PRESENTATION

4.1. System Structure

We study distributed data sets and federated learning scenarios involving multiple parties. The federated learning system consists of three parts: participants, cloud servers, and trusted third-party collaborators, as shown in Figure 1. For the sake of clear description, just take two participants as an example to illustrate the system architecture.
Supposed that the medical institutions (hospitals or medical examination institutions) of A and B want to train machine learning models together, and their medical information systems each have data about patient diagnosis records. Medical institution B also has label data that the model needs to predict. Considering data privacy and security reasons, A and B cannot directly exchange data. To ensure the confidentiality of data during the training process, a third-party collaborator C was introduced. We assume that Party C is honest and will not collude with Party A or Party B. However, Party A and Party B are honest but curious. Here, the trusted collaborator C is a reasonable assumption, because C can be assumed by such as the government or a secure computing node.

The division of the three parts is as follows:

- **Part 1 Local Participants**
  
  Each local participant who has its own local data set, is curious about other parties' data information. In order to maintain the patient's privacy information, local participant place their data in their own LAN system, and prohibits the sharing of any information with the outside world. So, in our system any participant trains model and makes prediction with their own island data respectively.

  Participants jointly set up the public key $pk$ and private key $sk$ for an additively homomorphic encryption scheme. And each participant establishes secure channel, different from each other, to communicate and protect the integrity of the homomorphic cipher texts.

- **Part 2 Cloud Server**
  
  Participants use local data to train local model and then upload related parameters to CS to update. Then, the updating model parameters are fed back to the participants. The participants trains data once more according to the new parameters and gain preferences model. The above process is run alternately between CS and participants until an optimal training result is reached.

  Obviously, the CS updates the encrypted weight parameters in each iteration process.

- **Part 3 Collaborator C**
  
  Collaborator C generates the public and secret key pairs $(sk, pk)$, and send public key $pk$ to local Participants.

4.2. Training Process

The basic process of the multi-party collaborative learning model is as follows:

- **Step 1**: Collaborator C chooses security parameter $\lambda$ and runs $KeyGen$ algorithm, then outputs key pairs $(sk, pk)$ and sends the public key $pk$ to local participants.

- **Step 2**: If necessary, the participants align data and exchange intermediate results with each other in an encrypted state.
• **Step 3:** After the participants complete the local model training, they submit the new gradient value to the cloud server in an encrypted state. The cloud server aggregates the encrypted model update values of each participant.

• **Step 4:** The cloud server updates the global model according to the aggregation result, and sends the updated model parameters to each participant. Participants retrain and update the local model with the new global model parameters received as initial values.

**Steps 3 and 4** are repeated until the number of iterations is reached or the global convergence reaches the target value.

5. **SYSTEM APPLICATIONS**

5.1. **Features in diabetes data analysis**

Here we mainly study the problem of diabetes prediction and classification, and further use the random forest algorithm to discover that certain factors are potential factors that cause the disease and the related rules between them. In the classification and analysis of diabetes and related data sets, the main problem is the two-class prediction. The method is to select the independent variable of whether you have diabetes as the categorical attribute, and the other dependent variable attributes as the conditional attributes. Finally, each item of the data is divided into different classification subsets.

5.2. **Algorithm design**

**Algorithm:** FederatedForest via Homomorphic Encryption

**Input:** Parameters that need to be set in advance: Number of trees $T$, Number of random attributes $K$, Maximum tree depth $H$, and Minimum number of leaf node records $Num$, etc.

**Output:** Predict classification results.

**For each participant** $L_i, 1 \leq i \leq N$:

1. Download the encrypted global model weight $E(w_{global})$ stored at the server.
2. Decrypt the received ciphertexts $E(w_{global})$ with the private key $sk$ and obtain $w_{global}$.
3. Take a random replacement method to select a small batch of samples from the local data set $D_i$ to build a decision tree.
4. Using the values of $w_{global}$ and data items at steps 2 and 3, update local model $w_{local}$ and compute the gradients $G_i$.
5. Send back the encrypted $E(-\eta \cdot G_i)$ to the cloud server.
6. Repeat steps 1 to 5 until the termination condition to obtain the best local classification results.

**For cloud server:**

Aggregates local update values submitted by all participants, $E(w_{global}) + \sum_{i=1}^{N} E(-\eta \cdot G_i)$ is under additively homomorphic encryption homomorphic protection.
5.3. Performance analysis

5.3.1. Security:
Our system is safe for cloud servers. Because the basic homomorphic encryption scheme ensures the security of CPA, the system does not leak the information on the dataset to honest but curious cloud servers. On the other hand, there is no direct data exchange between participants. The intermediate result of training is also protected by homomorphic encryption. Therefore, neither party can know the data of other participants.

Therefore, the data privacy of the participants in the system is absolutely protected.

5.3.2. Accuracy:
In the process of collaborative learning, each participant works independently and in parallel, without waiting for any other party. The locally trained model is not only affected by the local data set, but also optimized towards the global model. Therefore, the overall accuracy of the model is improved, compared to isolated local models training.

6. CONCLUSION
In this paper, we propose a federated forest algorithm to preserve the privacy of the training data set. And this machine learning method is applied to the prediction and diagnosis of diabetes classification in large-scale scenarios. We have introduced homomorphic encryption to encrypt data communication between participants and cloud servers to increase data privacy protection. Finally, we analyze the proposed method from the aspects of security and training accuracy, and the results show that the proposed method can achieve high quality privacy protection and accuracy guarantee.

ACKNOWLEDGMENT
This work was supported in part the Major Natural Science Research Projects in Colleges and Universities of Jiangsu Province (No. 18KJA520008), and the Postgraduate Research and Practice Innovation Program of Jiangsu Province (No. SJKY19_0759, No. SJKY19_0761).

REFERENCES
[1] S. Gang, L. Shanshan and Y. Ding. Design and Implementation of Diabetes Risk Assessment Model Based on Mobile Things[C]. 2015 7th International Conference on Information Technology in Medicine and Education (ITME), Huangshan, 2015: 425-428, doi: 10.1109/ITME.2015.169.
[2] Jakub Konecný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. CoRR abs/1610.02527 (2016). arxiv:1610.02527 http://arxiv.org/abs/1610.02527
[3] Jakub Konecný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. CoRR abs/1610.05492 (2016). arxiv:1610.05492 http://arxiv.org/abs/1610.05492
[4] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. 2016. Federated learning of deep networks using model averaging. CoRR abs/1602.05629 (2016). arxiv:1602.05629 http://arxiv.org/abs/1602.05629
[5] Y. Luo, C. Ling, J. Schuurman and R. Petrella. GlucoGuide: An Intelligent Type-2 Diabetes Solution Using Data Mining and Mobile Computing[C]. 2014 IEEE International Conference on Data Mining Workshop, Shenzhen, 2014, 748-752, doi: 10.1109/ICDMW.2014:177.
[6] Kavakiotis I , Tsave O , Salifoglou A , et al. Machine Learning and Data Mining Methods in Diabetes Research[J]. Computational & Structural Biotechnology Journal, 2017, 15:104-116.
[7] B.M. Patil. Hybrid Prediction Model for Type-2 Diabetic Patients[J]. Expert Systems with Applications 37, 2010: 8102–8108.
[8] Aliza Ahmad and Aida Mustapha H. Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus[J]. ICDIPC, Part I, CCIS 188, 2011: 537–545.

[9] Alexis Marcano-Cedeño, Joaquin Torres, and Diego Andina. A Prediction Model to Diabetes Using Artificial Metaplasticity[J]. IWINAC, Part II, LNCS 6687, 2011: 418–425.

[10] V. V. Vijayan and C. Anjali. Decision support systems for predicting diabetes mellitus — A Review[C]. 2015 Global Conference on Communication Technologies (GCCT), Thuckalay, 2015:98-103, doi: 10.1109/GCCT.2015.7342631.

[11] Omprakash Chandrakar, Dr. Jatinderkumar R. Saini. Development of Indian Weighted Diabetic Risk Score (IWDRS) using Machine Learning Techniques for Type-2 Diabetes[C]. ACM COMPUTE '16, October 21-23, 2016, Gandhinagar, India.

[12] L. Han, S. Luo, H. Wang, L. Pan, X. Ma and T. Zhang. An Intelligible Risk Stratification Model Based on Pairwise and Size Constrained Kmeans[J]. IEEE Journal of Biomedical and Health Informatics,2017, 21(5):1288-1296. doi: 10.1109/JBHI.2016.2633403.

[13] Strack B , Deshazo J P , Gennings C , et al. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records[J]. Biomed Research International, 2014, 2014:781670.

[14] Zhe Wei, Guangjian Ye and Nengcai Wang. Analysis for Risk Factors of Type 2 Diabetes Mellitus based on FP-growth Algorithm[J]. China Medical Equipment,2016, 13(5): 45-48.

[15] Yirui Guo. Application of Artificial Neural Network to Predict Individual Risk of Type 2 Diabetes Mellitus[J]. Journal of Zhengzhou University, 2014,49(3): 180-183.

[16] Wang Xidan, Wang Xiaodan, Liang Li. Application of Deep Learning Model in Type 2 Diabetes Risk Prediction[J]. Journal of Clinical Medicine, 2017,4(84):16460-16461.

[17] Rivest R , Adleman L , Dertouzos M L. On databanks and privacy homomorphism[J]. Foundations of Secure Computation, 1978, 4(11): 169-180.

[18] Gentry C. Fully homomorphic encryption using ideal lattices[C]. STOC, 2009, 9: 169-178.

[19] O. Goldreich. Foundations of Cryptography: Basic Applications[J]. Journal of the Acm, 2004, 10(509):359–364.

[20] A. Patil and S. Singh. Differential private random forest[C]. 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), New Delhi, 2014:2623-2630, doi: 10.1109/ICACCI.2014.6968348.

[21] Q. Yang , Y. Liu , T. Chen, Y. Tong. Federated Machine Learning[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2):1-19.