Research on Anti-Alzheimer’s Traditional Chinese Medicine with Data Security: Datasets, Methods, and Evaluation

Peng Tang, Yubin Zheng, Weidong Qiu, Hao Wang, Jie Guo, and Zheng Huang

1School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
2College of Basic Medical Sciences, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Correspondence should be addressed to Weidong Qiu; qiuwd@sjtu.edu.cn

Received 27 January 2022; Accepted 2 March 2022; Published 22 March 2022

Alzheimer’s disease (AD), a growing global health concern, has been posing a significant threat to the health of the aging population. The factors contributing to the occurrence and development of AD are extremely complex, including multiple neural networks and multiple targets, which join together to formulate enormous challenges in AD treatment. Traditional Chinese medicine (TCM) possesses the characteristics to regulate multiple targets at the same time, which is consistent with the pathogenesis of AD, moreover, clinical results in TCM treating AD reveal promising effects. In this paper, we first collected anti-Alzheimer’s prescriptions and their therapeutic effects from commonly used literature databases and expanded the data to form the anti-Alzheimer’s TCM dataset. Next, we combined machine learning models to train and analyze the dataset, which was used to predict the effectiveness of new TCM prescriptions. For the first time, we proposed to use the artificial intelligence method to train the properties of nature, flavor, and channel tropism in TCM prescriptions. The accuracy of the prediction model for the effectiveness of anti-Alzheimer’s can reach up to 85%. The experimental results demonstrated that our method can precisely predict the effectiveness of prescriptions against Alzheimer’s disease, and have great value in providing guidance for the development of new anti-Alzheimer’s drugs. Finally, we built a distributed model training architecture based on federated learning to train and predict the effectiveness of TCM prescriptions under the premise of ensuring data security.

1. Introduction

At present, Alzheimer’s disease (AD) has become the third major disease that seriously threatens the health of the elderly after cardiovascular diseases and malignant tumors. AD, as defined by the World Health Organization, is a chronic or progressive syndrome, usually a cognitive function that is worse than the normal aging process, which implicates memory, thinking, orientation, understanding, calculation, learning, language and judgment ability, but not consciousness. Cognitive impairment is usually accompanied by emotional control, social behavior and motivation decline. According to the statistics of the World Health Organization (WHO), currently, there are nearly 50 million AD patients worldwide, and the number is growing at 10 million per year with about $100 billions of treatment cost. It is estimated that the large AD patient population may reach 150 million in 2050, undoubtedly brings huge burden on individuals and society, therefore, new ideas of treatment is in urgent need [1].

Researchers have conducted multiple scientific researches on AD, however, effective approaches for cure AD are still unavailable, clinically, we can only try to alleviate the symptoms through drug treatment or other means. Currently, the drugs approved by the FDA for the treatment of AD mainly include donepezil, galantamine, rivastigmine, and memantine. Most drugs can simply relieve the symptoms of AD’s early or mid-stage, such as memory loss, but unable to prevent the disease itself.

Due to the complicated etiology and unclear pathogenesis, drugs with single target are difficult to achieve the effect of curing AD. Whereas, TCM has the characteristics of multiple-component and multiple-path, which can make up for the limitations of drugs with single target. TCM has a
long history in treating AD and has accumulated many experiences. Although the “Alzheimer’s disease” were unrecorded in ancient medical books, there are records of symptoms such as “forgetfulness,” “dignity,” and the classic prescriptions to treat these symptoms include “Kaixin San,” “Clever Soup,” “Yizhi Soup,” “Dihuang Yinzi,” and so on.

With the rise of big data analysis and deep learning technology, a new path has been opened up for the development of TCM. Computer technology can transform the clinical experience and related information of TCM into data, and construct a large database of TCM, explain TCM theories in a more scientific and systematic way, thereby promoting the development of TCM. At present, big data has been used in the diagnosis of diseases, clinical treatment, and the inheritance of TCM experience.

TCM has a long history of treating AD. In ancient China, there have been many documents recording the treatment of AD-related diseases. The first chapter of “Jing Yue Quan Shu” records the TCM prescriptions for the treatment of dementia, which is the earliest known herbal medicine in the world [2]. Liu et al. [3] conducted a comprehensive investigation, summary and analysis on the causes of AD, TCM treatment methods, and TCM extracts to treat AD. Wang et al. [4] reviewed the active ingredients in TCM, and concluded that TCM has the advantage of multitarget regulation in the treatment of AD. Acetylcholinesterase inhibitors (AChEI) are the first drugs approved by the FDA to treat mild to moderate AD. Most of these drugs, such as huperzine, was first isolated from TCM. Lin et al. [5] based on the Ellman method to detect the AChE inhibitory activities of TCM extracts for insomnia and brain function disorders, and experimental results show that these TCM have great potential for the treatment of AD. HupA, an effective AChEI, is an extract of Huperzia serrata, which has been widely used in the treatment of AD in China. Jiang et al. [6] comprehensively investigated the clinical, pharmacological, chemical and physical structure aspects of HupA.

With the development of big data analysis and machine learning technology, artificial intelligence (AI) technologies such as machine learning played an important role in various fields, such as network analysis [7–10] and security [11, 12], disease diagnose and treatment [13–17]. Joshi et al. [18] investigated the current status of artificial intelligence in the field of healthcare and discussed its prospect. The authors believe that AI technology can be applied to various medical fields, including early disease detection and diagnosis, treatment and results forecasting and so on, while AI can also be applied to different diseases, including cancer, heart disease, liver disease [19] and so on. There are also many studies on machine learning in the treatment of AD. Christian et al. [20] used machine learning methods to identify and diagnose relevant biomarkers in MRI, so as to detect early AD and provide defense and treatment timely. Joshi et al. [21] established a neural networks (NN) classification model and found that some specific factors have a great impact on AD. Machine learning methods are also used in the study of TCM in the treatment of AD. Pang et al. [22] collected 13 TCM prescriptions for the treatment of AD and 25 AD-related targets, and selected 7 representative TCM for the machine learning model training, through the machine learning model to predict the effective components of TCM prescriptions, this method provides new ideas for the research and development of TCM, but this method lacks the analysis of large amounts of data. This paper only mentioned that the model has high prediction accuracy, but did not report the specific value of accuracy, precision, etc. Chen et al. [23] used deep learning and random forest methods to find the best TCM formula for AD, but this method is only for glycogen synthase kinase 3β (GSK3β) related TCM.

Firstly, this paper collected and constructed a data set of TCM prescriptions for anti-Alzheimer’s disease. Secondly, combined with machine learning algorithm, we proposed an effectiveness prediction model based on drug attributes, which can predict whether the TCM prescriptions are effective against Alzheimer’s disease. Finally, an effectiveness prediction model was constructed based on federated learning to complete the prediction while ensuring the security of user privacy data and prescription data. This paper conducted supplementary research on the basis of conference paper [24], carried out experimental expansion from both data and model, enhanced the data set and built a more reasonable model to model and analyze the attributes of TCM prescriptions, and the accuracy of the model is improved.

2. Methodology

Figure 1 shows the overall framework of this article. First, a dataset of TCM prescriptions against Alzheimer’s disease was constructed through literature search, and then the attributes in the TCM prescriptions were extracted as features to train the machine learning models, and different models were used for experiments and comparisons. Finally, the results and advantages of different models were analyzed.

2.1. Properties of TCM. The nature, flavor and channel tropism are the special properties of TCM. According to the 2015 edition of the Chinese Pharmacopoeia [25], there are five types of nature of drugs: “calm,” “cold,” “hot,” “warm,” and “cool.” Flavor can be divided into six types: “pungent,” “sweet,” “sour,” “astringent,” “bitter,” and “salty.” Channel tropism can be divided into eleven types: “heart channel,” “kidney channel,” “lung channel,” “liver channel,” “spleen channel,” “stomach channel,” “gallbladder channel,” “large intestine channel,” “bladder channel,” and “tri-jiao channel.”

2.2. Data Collection and Preprocessing

2.2.1. Collection and Classification of TCM Prescriptions. This article collected literature on treatment of AD-related diseases from 1988 to the present from databases such as China HowNet, Weipu, Wanfang and other eligible libraries, and a total of 224 TCM prescriptions have been collected. The total number of patients who used each TCM prescription and the number of patients who experienced significant improvement with the TCM prescription were
documented in the literature. The proportion of patients with effective treatment is recorded as the effective rate of the TCM prescription. The prescriptions with an effective rate greater than or equal to 85% are labeled “1,” the prescriptions with an effective rate less than 85% are labeled “0.”

2.2.2. Data Standardization. A TCM prescription contains many drugs, and each drug contains properties of nature, flavor and channel tropism. Nature can be divided into five categories, flavor can be divided into 6 categories and channel tropism can be divided into 11 categories. Therefore, each drug has at most 22 properties, which can be represented by an array of 22 bits. If a drug has one of the above properties, the corresponding position in the array is “1”; otherwise, the corresponding position is “0.” For example, “radix rehmanniae preparata,” a common drug in TCM prescriptions, is “warm,” “sweet” and it belongs to “kidney channel” and “liver channel,” so we set “00010” as nature attributes in the order of “calm, cold, hot, warm, cool,” set “010000” as flavor attributes in the order of “pungent, sweet, sour, astringent, bitter, salty” and set “01010000000” as tropism channel attributes in the order of “heart channel, kidney channel, lung channel, liver channel, spleen channel, stomach channel, gallbladder channel, large intestine channel, bladder channel, tri-jiao channel.” Then the radix rehmanniae preparata can then be converted into a 22-bit array “00010100000101000000.”

According to statistics, among the TCM prescriptions collected, each contains at most 18 kinds of drugs. And each drug can be normalized with a 22-bit array, so each prescription can be normalized with an 18 * 22 matrix. For prescriptions with fewer than 18 drugs, fill the matrix with “0.” We ended up with a dataset of 224 prescriptions. The dataset is recorded as “original dataset.” Figure 2 shows a prescription and its’ standardized data.

Some machine learning algorithms, such as SVM, cannot handle 18 * 22 matrices. In order to observe the performance of these models on predicting the effectiveness of TCM, the two-dimensional feature matrix is further accumulated by columns, so that the 18 * 22 two-dimensional feature matrix is reduced to 1 * 22 one-dimensional feature vector, and a new data set is generated, and recorded as “ori-one-dim dataset.” Figure 2 shows the generation of a 1 * 22 one-dimensional feature vector.

2.2.3. Dataset Expansion. Although the use of TCM has a long history, due to the lack of systematic and complete data records in ancient times, credible clinical data is not enough. It has been pointed out in Section 2.2.1 that only 224 prescriptions can be collected from the literature database, however, in Section 2.2.2, the features after data standardization reach 392 ((18 * 22) dimensions. Therefore, the number of samples in the data set is obviously insufficient. In
which case, we consider using the following three methods for data expansion.

(1) Rearrange the Row Vectors of the Feature Matrix. In the feature matrix of prescriptions obtained in Section 2.2.2, the row vectors represent the drug properties, and the row vectors sequence is the sequence of the drugs. Changing the order of the drugs will not change the properties of the prescription, that is, the label of the prescription does not change. But for the feature matrix, changing the order of the row vectors generates a new feature matrix. Figure 3 demonstrates the generation of new sample data from existing sample data. For each prescription, containing \( n \) kinds of drugs, there is only one feature matrix in the "original dataset," \( 2^n \) new feature matrices are randomly generated using the method of row vector rearrangement, and finally, a dataset containing 5396 feature matrices is obtained based on the "original dataset," which is recorded as "rearrange dataset."

(2) Add Redundant Information. Add one or two drugs to the original prescription data to form new prescription data. On account that the original properties of the prescriptions cannot be affected, the added drugs should appear with low frequency in all the prescriptions and do not have main therapeutic effect. The better additives include jujube, musk, cloves and so on. Figure 4 demonstrates the generation of new prescriptions from existing prescriptions. 224 prescriptions in the "original dataset" are split into two parts. One part contains 154 prescriptions, by adding redundant information to this part of the data, 144 new prescriptions are obtained, after merging the original 154 prescriptions, 298 prescriptions are obtained, as a training set and recorded as "add-red dataset." The other part contains 70 prescriptions as a test set.

(3) Mixed Method. The row vector rearrangement method is used to rearrange each matrix in the "add-red dataset," which contains 298 prescriptions, to generate a new training set containing 7452 feature matrices and recorded as "mix dataset." The same method is used for the test set, which contains 70 prescriptions, to generate a new test set containing 1726 feature matrices.

The result of accumulating the matrix in "rearrange dataset" by column is the same as that of accumulating the matrix in "original dataset," both are "ori-one-dim dataset." The matrix in the "add-red dataset" and the matrix in the "mix dataset" are respectively accumulated by column, and the result is the same one-dimensional matrix, which is an extension of the "ori-one-dim dataset," is recorded as "mix-one-dim dataset." Table 1 summarizes all the datasets used and their descriptions.

2.2.4. Feature Engineering. The total number of TCM prescriptions collected for anti-AD is 224, in which each drug has as much as 22 features, which will lead to overfitting for machine learning algorithms. Therefore, it is necessary to filter and select important features for training.

The data set was divided randomly into a training set and a test set, use the XGBoost algorithm to train the data, then call the feature importance function to output the importance of 22 features. This process will be repeated 8 times, each time the data set is re-divided into a training set and a test set, features with low importance are counted in the following 8 results.

The final result showed that, among the 22 features, eight were of relatively low importance and have less effect to the prediction, such as "hot," "cold," "astringent," "gallbladder channel," "large intestine channel," "bladder channel," "trijiao channel," and "small intestine channel." On the basis of "ori-one-dim dataset," these eight features were deleted to generate new data sets, which were recorded as "fil-one-dim dataset."
2.3. Training Method

2.3.1. Multilayer Perceptron (MLP). MLP is a typical artificial neural network, which has been proved to be effective in various scenarios. MLP consists of an input layer, hidden layers, and an output layer that maps the input to the output by nonlinear transformation.

In this paper, the input of MLP is the data set generated in Section 2.2 (including the original TCM prescription dataset and the expanded dataset), and the output is the label of prescription.

2.3.2. Convolutional Neural Network (CNN). CNN is often used in computer vision (CV) tasks. Using the neural network based on CNN to detect and classify objects in images will have a better effect. The basic CNN includes convolution layer, pooling layer and full connection layer, of which the layer is used to extract local features, reduce feature dimensions and complete the classification task, respectively. The sample in the datasets used in this paper is a two-dimensional matrix, so CNN can be used to complete the classification.

In this paper, the input of CNN is a two-dimensional feature matrix of sample, including “original dataset,” “rearrange dataset,” “add-red dataset,” and “mix dataset” and the output is the label of prescription.

2.3.3. Support Vector Machines (SVM). SVM is a traditional supervised classification algorithm. Its principle is to obtain a hyperplane through training, and use the hyperplane to divide data into two categories. Through kernel function, SVM can also solve nonlinear classification problems. Although deep learning technologies such as MLP and CNN have been widely used in various fields, machine learning technologies such as SVM will also achieve better results when the amount of data is small. The original data in this paper are two-dimensional matrices, which can be converted into one-dimensional vectors by adding columns, so that SVM can be used for training.

In the paper, the input of SVM is a one-dimensional feature vector of sample, including “ori-one-dim dataset,” “mix-one-dim dataset” and “fil-one-dim dataset,” and the output is the label of prescription.

2.3.4. Ensemble Learning (EL). EL algorithm is also a more popular machine learning algorithm in recent years. It obtains a strong classifier by integrating multiple weak classifiers, so it usually has better effect than a single classifier. The EL algorithm used in this paper is XGBoost.

In the paper, the input of XGBoost is a one-dimensional feature vector of sample, including “ori-one-dim dataset,” “mix-one-dim dataset,” and the output is the label of prescription. Since the “fil-one-dim dataset” is a dataset filtered

| Dataset               | Description                                      |
|-----------------------|--------------------------------------------------|
| Original dataset      | The feature matrix of the original prescriptions  |
| Ori-one-dim dataset   | Add the original dataset by column to get the one-dimensional vector |
| Rearrange dataset     | Rearrange row vectors for original dataset       |
| Add-red dataset       | Added redundant drugs to original dataset        |
| Mix dataset           | Rearrange row vectors for add-red dataset        |
| Mix-one-dim dataset   | Add the add-red dataset by column to get the one-dimensional vector |
| Fil-one-dim dataset   | a one-dimensional vector after filtering some features of the ori-one-dim dataset |

Figure 4: Generate new prescriptions from existing prescriptions.

Table 1: Summarizes all the data sets used and their descriptions.
according to XGBoost, and using XGBoost to train "fil-one-dim dataset" will produce overfitting, therefore the "fil-one-dim dataset" is not used to train the XGBoost model.

2.4. Evaluation

2.4.1. MLP. Firstly, this paper trained MLP model to predict the effectiveness of TCM prescriptions in the treatment of AD. We constantly adjust the parameters of the MLP model to get the best model and save it. The model consists of two hidden layers, each of which contains 32 nodes. The hidden layers are connected in a fully connected way. The dropout of the hidden layer is set to 0.8. The dropout is used to prevent overfitting of the model. ReLU is used as the activation function of the hidden layer and softmax is used as the activation function of the output layer. The function of the activation function is to carry out nonlinear transformation. The cross entropy is used as the loss function, which is used to measure the prediction quality of the model. Optimizer algorithm and learning rate are used to update the weight of neural network. In the MLP model of this paper, the optimizer algorithm adopts Adam algorithm, and the learning rate is set to 0.001.

The inputs of the MLP model can be either two-dimensional matrices or one-dimensional vectors, so all extended datasets and original datasets in Section 2.2 can be trained using the MLP model. Figures 5 and 6 show results of the MLP model trained using the different datasets in Section 2.2.

It can be seen from Figure 5 that MLP fits well on the training set of "original dataset" and "add-red dataset," but the effect on the test set indicates that MLP has overfitted. The reason for the poor effect on the "original dataset" may be that the number of samples in the "original dataset" is small, and there are more features that lead to overfitting. The reason for the poor effect of "add-red dataset" may be: the additional drugs added to the prescriptions may be considered as redundant information and interfere with model training.

MLP is not overfitting on the "rearrange dataset," and the accuracy of the test set can reach 0.75. This shows that the data expansion method of row vector rearrangement has a certain effect. Row vector rearrangement may help the MLP model resist the interference of the order of medicinal materials and focus more on the properties of nature, flavor and channel tropism. The "mix dataset" can be regarded as generated on the basis of the "rearrange dataset," so the MLP model has the same performance on the "mix dataset."

As shown in Figure 6, both the "ori-one-dim dataset," "mix-one-dim dataset" and the "fil-one-dim dataset" are one-dimensional feature vectors. The "ori-one-dim dataset" is converting the "original dataset" into one-dimensional vector, reducing the feature dimension and improving the effect of the MLP model. However, due to the small number of samples, the accuracy of the model is still not high. "Fil-one-dim dataset" is a data set generated after feature filtering of "ori-one-dim dataset," the accuracy of MLP on this data set has also improved, which shows that feature filtering is useful. "mix-one-dim dataset" is generated by adding up columns of "add-red dataset," due to the reduction of dimension, the accuracy is also improved.

2.4.2. CNN. The inputs to the CNN model are 18°22 matrices. Similar to MLP, the dropout of CNN model is set to 0.8, the activation function of the hidden layer is ReLU, the activation function of the output layer is softmax, the loss function is cross entropy, the optimization algorithm is Adam, and the learning rate is set to 0.001. Different from MLP, CNN’s hidden layer consists of four convolutional layers and pooling layers, which are used to extract local features of matrix.

Figure 7 shows results of the CNN model trained using the different datasets in Section 2.2. CNN is also overfitting on "original dataset" and "add-red dataset," but is not overfitting on "rearrange dataset" and "mix dataset," which is consistent with the results of the MLP model. The accuracy rate of CNN on the test set of "rearrange dataset" reached 0.78, and the accuracy rate of "mix dataset" reached 0.85, which is a greater improvement than the accuracy of the MLP model.

2.4.3. SVM. NuSVC is a support vector machine based on libsvm implementation. In this paper, NuSVC is used to train the one-dimensional vector datasets. NuSVC supports the use of kernel function to complete nonlinear classification. In this paper, the Gaussian kernel function “rbf” is used as the kernel function, and the coefficient gamma of the kernel function is set to “auto.” NuSVC also provides an upper limit parameter of training error rate, Nu, whose value range is (0, 1), which is set to 0.5 in this paper. We train the NuSVC model using the method of k-fold cross validation, where k is set to 8. The data set was randomly divided into training set and test set, and the training set accounted for 75% and the test set 25%.

The generated data in Section 2.2 contains two kinds of one-dimensional vector datasets, one is a one-dimensional vector with a length of 22, and the other is a one-dimensional vector with a length of 14 after feature engineering. So, the input data for the model is 22 or 14 widths. Figure 8 shows results of the SVM model trained using the different datasets in Section 2.2.

It can be seen from Figure 8 that although the accuracy of the NuSVC model is not the highest, its performance is relatively stable. In the cross-validation method, the average accuracy of the model exceeds 0.6. Experiments show that when the amount of data is sufficient, the SVM model will have a good effect.

Compared with the "ori-one-dim dataset," the accuracy of the SVM model on the "fil-one-dim dataset" has been improved, and meanwhile is more stable, indicating that feature filtering is effective on this problem; moreover, the deleted eight features have a lesser degree of influence on the effectiveness of the prescription. The SVM model had the highest accuracy and was more stable in the "mix-one-dim dataset," indicating that data expansion could effectively improve the effect of SVM model.
Figure 5: Continued.
2.4.4. EL. Experiment with the classical algorithm XGBoost in EL. The input data for the model is 22 or 14 widths. Booster is set to "gbtree," the gamma value is set to 0.1, and the maximum depth of the tree is set to 6. The learning rate is set to 0.1, and the training rounds are 500 rounds. Similar to SVM, we used k-fold cross validation to train the XGBoost model, where $k$ is set to 8. The data set was randomly divided into training set and test set, and the training set accounted for 75% and the test set 25%. Figure 9 shows the experimental results of XGBoost on the datasets.

As can be seen from Figure 9, compared with the MLP and CNN models using the "original dataset," XGBoost has better performance on the corresponding dataset ("ori-one-dim dataset") and is more stable, compared with the MLP using "ori-one-dim dataset," the effect of XGBoost is poor. However, compared with SVM, the performance of XGBoost on the "ori-one-dim dataset" is slightly better than that of SVM, but the performance on the "mix-one-dim Dataset" is worse than that of SVM.

3. Data Security Scheme

3.1. Overview. The experimental results in Section Methodology show that our proposed method for training machine learning models based on drug attributes is effective for predicting the effectiveness of anti-Alzheimer’s TCM prescriptions. However, in a real scenario, the data of TCM prescriptions are issued by doctors in different hospitals. The method in Section Methodology needs to aggregate each hospital’s data for model training, which will cause data security problems. For example, the prescription data is the intellectual property of the hospital, the hospital may be unwilling to share the data with a third party, or the patient’s private information exists in the data, and the sharing of data will lead to the leakage of user privacy and so on. Aiming at the data security problems that may exist in real scenarios, this paper built a scheme that can ensure data security, and conducted model training without sharing the data with a third party. The specific scheme is shown in Figure 10.

As shown in Figure 10, we proposed a model training scheme based on federated learning to ensure the data security of TCM prescriptions in real scenarios. The scheme mainly includes two stages: data processing stage and model training stage. In the data processing stage, each hospital participating in the model training mainly completes the data processing and feature extraction of the original TCM prescription. The method of feature extraction was consistent with the method in Section Methodology, which extracted the nature, flavor and channel tropism attributes of the drugs in the prescription. In the model training stage, unlike the training of a single model in Section Methodology, this scheme used its own data to train a separate model locally in each hospital, and then aggregates the parameters of the local model to the central server for aggregation. Then return the parameter update rules to each local model, and the local model updates the parameters according to the rules, and finally completes the training of the local model.

In this scheme, the same data processing and feature extraction methods in Section Methodology are used, but different models are used for training. The model does not collect data of all hospitals for analysis, but only needs to collect encryption parameters in the training process of each local model, which ensures data security.

3.2. Experimental Results and Analysis. In this paper, the open source federated learning framework FATE was used to conduct experiments. The original training set was randomly divided into two equal parts and stored on two servers with FATE installed, so as to simulate the participation of two hospitals in model training. Finally, the model was tested on the test set. SecureBoost model provided by FATE
Figure 6: The experimental results of MLP on the one-dimensional datasets. (a) Ori-one-dim dataset. (b) Mix-one-dim dataset. (c) Fil-one-dim dataset.
Figure 7: Continued.
is used for training in this paper. SecureBoost is implemented based on XGBoost, which is also an ensemble learning algorithm. During the experiment, we continuously adjusted the parameters of the secureBoost model to make the model achieve the best effect. In the final model, the maximum depth of the trees was set to 10, the number of the trees was set to 100, and the learning rate was set to 0.01.

FATE provides an open source visual interface, FATEBoard, to view the model training results. The model training results in this paper are shown in Figure 11, and the accuracy of the model on the training set reaches 0.71.

The model was used to test the test set. The test results are shown in Table 2. The accuracy rate of the model on the test set reached 0.6, and the F1 value reached 0.64, indicating that the prediction model of effectiveness of anti-Alzheimer’s TCM prescription based on federated learning proposed in this paper is still effective.

4. Discussion

4.1. Principal Results. We are the first to analyze the effectiveness of anti-Alzheimer’s prescription using the property of TCM combined with machine learning method, and we are also the first to build the database of anti-Alzheimer’s prescription.

Table 3 summarizes the accuracy of all models in this paper. It can be seen from the table that when the amount of data is small, machine learning algorithms such as SVM and EL have advantages. The SVM model performs better than the EL model, with an accuracy rate of 69%, but the neural network model is overfitting in these small data sets. When the amount of data is large, the neural network algorithm has a greater advantage. Among which, the performance of the CNN model is better than that of the MLP model, with an accuracy rate of 85%. In the dataset constructed in this article, the CNN model achieves the best results, with an accuracy rate of 85%. On account that CNN has the characteristics of learning high-dimensional features, when the amount of data is large enough, CNN can learn the relationship between drugs’ attributes well so as to predict the effectiveness of TCM composed of multiple drugs. Finally, we train the SecureBoost model in federated learning to ensure data security, and the model accuracy can reach 0.60, indicating that the proposed data security scheme is still effective.

4.2. Conclusions. In this paper, we have collected the data of anti-Alzheimer’s disease TCM prescriptions and their efficacy recorded in the literature so far, converted the TCM prescriptions into a two-dimensional digital matrix according to the drug attributes, and formed the original dataset. By rearranging row vectors and adding redundant information, the original data set was expanded and new datasets were generated. At the same time, this paper proposes a machine learning-based prediction method for the effectiveness of anti-Alzheimer’s TCM prescriptions. The datasets constructed in this paper are used to train MLP, CNN, SVM and EL models. Among them, the CNN model has the highest accuracy, reaching 85%. The experimental results show that the method proposed in this paper can effectively predict the effectiveness of TCM for the treatment of Alzheimer’s disease, and has a guiding role for the production of new drugs. The proposed method is also applied to the real scene, and the joint modeling of each hospital is carried out under the condition of ensuring the data security, and the experimental verification is carried out. The verification results show that the proposed scheme can not only ensure the data security, but also complete the model training and prediction. At the same time, the method proposed in this paper can be extended to predict the therapeutic effect of TCM on other diseases, which will play a positive role in the promotion of TCM.
Figure 8: The experimental results of SVM on the datasets. (a) Ori-one-dim dataset. (b) Mix-one-dim dataset. (c) Fil-one-dim dataset.

Figure 9: The experimental results of XGBoost on the datasets. (a) Ori-one-dim dataset. (b) Mix-one-dim dataset.

Figure 10: Model training scheme based on federated learning.
Data Availability

The csv data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 61972249.

References

[1] D. Bates, A. Barth, and C. Jackson, "Regular expressions considered harmful in client-side XSS filters," in Proceedings of the International Conference on World Wide Web, Raleigh, North Carolina, USA, April 2010.

[2] J. Li and B. Y. Liu, "Evaluation of clinical therapeutic effect in Jing Yue quan shu [Jingyue's Complete Works]," Zhonghua yi shi za zhi, vol. 39, no. 1, pp. 59–61, Beijing, China, 2009.

[3] P. Liu, M. Kong, S. Yuan, J. Liu, and P. Wang, "History and experience: a survey of traditional Chinese medicine treatment for Alzheimer’s disease," Evidence-based Complementary and Alternative Medicine: eCAM, vol. 2014, no. 3, Article ID 642128, 2014.

[4] Z. Y. Wang, J. G. Liu, H. Li, and H. M. Yang, "Pharmacological effects of active components of Chinese herbal medicine in the treatment of alzheimer’s disease: a review," American Journal of Chinese Medicine, vol. 44, pp. 1–17, 2016.

[5] H. Q. Lin, M. T. Ho, L. S. Lau, K. K. Wong, P. C. Shaw, and D. C. Wan, "Anti-acetylcholinesterase activities of traditional Chinese medicine for treating Alzheimer’s disease," Chemistry-Biological Interactions, vol. 175, no. 1, pp. 352–354, 2008.

[6] H. Jiang, X. Luo, and D. Bai, "Progress in clinical, pharmacological, chemical and structural biological studies of huperzine A: a drug of traditional Chinese medicine origin for the treatment of alzheimers disease," Current Medicinal Chemistry, vol. 10, no. 21, pp. 2231–2252, 2003.

[7] D. Zhang, Y. Jie, X. Zhu, and C. Zhang, "Network representation learning: a survey," IEEE Transactions on Big Data, vol. 6, no. 99, p. 1, 2017.

[8] C. Li, B. Feng, S. Li, J. Kurths, and G. Chen, "Dynamic analysis of digital chaotic maps via state-mapping networks," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 66, 2014.

[9] H. Wang, J. Wu, P. Zhang, and Y. Chen, "Learning shapelet patterns from network-based time series," IEEE Transactions on Industrial Informatics, vol. 15, no. 7, pp. 3864–3876, 2019.

[10] C. Johnson, B. Khadka, E. Ruiz, J. Halladay, T. Doleck, and R. B. Basnet, "Application of deep learning on the characterization of tor traffic using time based features," J. Internet Serv. Inf. Secur, vol. 11, no. 1, pp. 44–63, 2021.

[11] G. Kasturi, A. Jain, and J. Singh, "Detection and classification of radio frequency jamming attacks using machine learning," J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl, vol. 11, no. 4, pp. 49–62, 2020.

[12] I. V. Kotenko, I. Saenko, and A. Branitskiy, "Applying big data processing and machine learning methods for mobile internet of things security monitoring," J. Internet Serv. Inf. Secur, vol. 8, no. 3, pp. 54–63, 2018.

[13] Y. Huang, C.-L. Hsu, and V. S. Tseng, "PURL: periodic user representation learning from temporal event records for personalized health management," in Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp), Busan, Korea (South), February 2020.

[14] K. Shankar, M. Elhoseny, M. Ilayaraja, R. M. Vidhyavathi, A. E. Mohamed, and A. Majid, "Optimal feature level fusion based ANFIS classifier for brain MRI image classification," Concurrency and Computation: Practice and Experience, vol. 32, no. 1, 2020.

[15] A. Vaid, S. K. Jadadanki, J. Xu, S. Teng, and B. S. Glicksberg, "Federated learning of electronic health records to improve
mortality prediction in hospitalized patients with COVID-19,” Machine Learning Approach, vol. 9, 2021.

[16] K. Ikemura, “Using automated machine learning to predict the mortality of patients with COVID-19: prediction model development study,” Journal of Medical Internet Research, vol. 23, no. 2, Article ID e23458, 2021.

[17] Y. A. Shichkina, G. V. Kataeva, Y. A. Irishina, and E. S. Stanevich, “The use of mobile phones to monitor the status of patients with Parkinson’s disease,” J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl, vol. 11, no. 2, pp. 55–73, 2020.

[18] F. Jiang, Y. Jiang, H. Zhi et al., “Artificial intelligence in healthcare: past, present and future,” Stroke & Vascular Neurology, vol. 2, no. 4, p. 230, 2017.

[19] J. C. Ahn, A. Connell, D. A. Simonetto, C. Hughes, and V. H. Shah, “The application of artificial intelligence for the diagnosis and treatment of liver diseases,” Hepatology, vol. 73, no. 6, pp. 2546–2563, 2021.

[20] S. Christian, C. Antonio, B. Petronilla, G. M. C. Q. Aldo, and C. Isabella, “Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer’s disease: a machine learning approach,” Frontiers in Neuroscience, vol. 9.

[21] S. Joshi, D. Shenoy, P. L. Rrashmi, K. R. Venugopal, and L. M. Patnaik, “Classification of alzheimer’s disease and Parkinson’s disease by using machine learning and neural network methods,” in Proceedings of the Second International Conference on Machine Learning and Computing, Bangalore, India, February 2010.

[22] X. C. Pang, Z. Wang, J. S. Fang, W. W. Lian, and G. H. Du, “Network pharmacology study of effective constituents of traditional Chinese medicine for Alzheimer’s disease treatment,” Yao xue xue bao = Acta pharmacutica Sinica, vol. 51, no. 5, pp. 725–731, 2016.

[23] H.-Y. Chen, J.-Q. Chen, J.-Y. Li, H.-J. Huang, X. Chen, and H.-Y. Zhang, “Deep learning and random forest approach for finding optimal TCM formula for treatment of Alzheimer’s disease,” Journal of Chemical Information and Modeling, vol. 4, 2019.

[24] J. Lu, P. Tang, W. Qiu, H. Wang, and J. Guo, Effectiveness Analysis of Traditional Chinese Medicine for Anti-alzheimer’s Disease Based on Machine Learning, Springer, Berlin, Germany, 2020.

[25] Chinese Pharmacopoeia Commission, Pharmacopoeia of the People’s Republic of China, Vol. 1, China Medical Science Press, Beijing, China, 2015.