IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 10, 2021, accepted January 27, 2021, date of publication February 5, 2021, date of current version February 23, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3057382*

# Hybrid Feature Embedded Sparse Stacked Autoencoder and Manifold Dimensionality Reduction Ensemble for Mental Health Speech Recognition

**HONG CHEN[1], YUAN LIN[2], YONGMING LI[2], (Member, IEEE), WEI WANG[1], PIN WANG[2], AND YAN LEI[2]**
[1]Chongqing University Cancer Hospital, Chongqing 400030, China
[2]School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China

Corresponding authors: Yongming Li (yongmingli@cqu.edu.cn) and Wei Wang (abbystina98@163.com)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61771080; in part by the Fundamental Research Funds for the Central Universities under Grant 2019CDQYTX019, Grant 2019CDCGTX306, and Grant 2019CDQYTX017; in part by the Basic and Advanced Research Project in Chongqing under Grant cstc2018jcyjAX0779, Grant cstc2020jcyj-msxmX0523, Grant cstc2020jscx-msxm0369, and Grant 2020ZY023659; and in part by the Research on Tumor Metastasis and Individualized Diagnosis and Treatment in Cancer Hospital Affiliated to Chongqing University Chongqing Key Laboratory Open Project Fund General Project.

**ABSTRACT** Speech feature learning is the key to speech mental health recognition. Deep feature learning can automatically extract the speech features but suffers from the small sample problem. The traditional feature extract method is effective, but cannot find the inter-feature structure to generate the new high-quality features. This paper proposes an embedded hybrid feature deep sparse stacked autoencoder ensemble method to solve this problem. Firstly, the speech features are extracted based on prior knowledge and called original features. Secondly, the original features are embedded into the deep network (Sparse Stacked Autoencoder) to filter the output of the hidden layer, to enhance the complementarity between the deep features and the original features. Thirdly, the L1 regularized feature selection mechanism is designed to reduce the hybrid feature set formed by the combination of deep features and original features. Finally, a manifold projection classifier ensemble is designed to enhance the stability of classification. Besides, this paper for the first time proposes a speech collection scheme for mental health recognition. We construct a large-scale Chinese mental health speech database for verification of the proposed algorithm of mental health. In the experimental section, the proposed algorithm is verified and compared with the representative related algorithms. The experimental results show that the proposed algorithm has better classification accuracy than the other representative algorithms. The proposed method combines the advantages of deep feature learning and traditional feature extraction methods more efficiently to solve the small sample problem.

**INDEX TERMS** Embedded hybrid feature sparse stacked autoencoder, ensemble learning, feature fusion, L1 regularization, speech mental health recognition.

## I. INTRODUCTION

Mental disease refers to the disorder of brain function and leads to different degrees of mental disorders in cognitive, emotional, volitional, and behavioral activities [1]. Most patients will have cognitive impairment during the disease. The course of mental disease patients is generally intermittent, with repeated attacks and deterioration. Some patients eventually suffer from mental decline and mental disability.

Studies have shown that speech disorder is one of the early symptoms of mental patients [2]. With the deepening of the disease, the problems of clarity and fluency of speech will gradually appear. Now, with the development of computer technology and acoustic analysis technology,

The associate editor coordinating the review of this manuscript and approving it for publication was Filbert Juwono.

the pronunciation characteristics of mental patients are gradually paid attention to. Based on the pathological features of speech, the painless and noninvasive objective auxiliary diagnosis of mental health using machine learning technology has become a major research hotspot [3]. Liu Z *et al.* [4] have proved that there is a significant difference between depression and healthy people in pronunciation time and pause time. Stasak *et al.* [5] further analyzed how pronunciation is affected by depression. Compared with several common mental disease diagnosis methods such as medical images and electroencephalogram signals, the operation process of mental disease diagnosis method based on speech is more simple and convenient, and the price of diagnosis is cheaper and it will not bring any side effects to the subjects [6], [7]. This diagnostic method has gradually attracted people's attention.

Feature learning is the key to the mental health recognition method based on speech data analysis. In recent years, although there have been relevant studies, the accuracy is still not satisfactory. According to the relevant papers [8], clinical studies showed that the Mel frequency cepstrum coefficient (MFCC) score of psychiatric patients was significantly lower than that of the control group. Alghowinem *et al.* [9] have proved that the MFCC has important statistical significance for the detection of depression. Joshi *et al.* [10] further summarized and compared the statistical characteristics of several low-order descriptors in the classification of depression. Mitra *et al.* [11] extracted four low-level descriptor features including damped oscillator cepstrum coefficients to identify depression, and achieved preliminary results. Wei *et al.* [12] extracted 26 features such as MFCC, and established a prediction model with logistic regression, which achieved good results in the recognition of depression. In the depression recognition experiment of Asgari *et al.* [13], the extracted speech features mainly include loudness-related features (loudness, etc.), vocalization-related features (jitter, frequency, etc.) and speech-related features (MFCC, etc.). The classification accuracy of depression can reach 74%. On this basis, Kaya *et al.* [14] adopted feature selection methods including Maximum Collective Relevance Canonical Correlation Analysis (MCR-CCA), Redundancy Maximum Relevance Canonical Correlation Analysis (mRMR-CCA), and Based Feature Selection (BFS) to further improve the classification accuracy. Jiang *et al.* [15] used Principal Component Analysis (PCA) as the dimension reduction algorithm, and used K-Nearest Neighbor (KNN), Gaussian Mixed Model (GMM), and Support Vector Machine (SVM) classifiers in the classification model, and achieved good psychiatric recognition results.

However, the traditional feature processing algorithms are based on the shallow feature learning of empirical knowledge, which cannot effectively mine the internal nonlinear complex relationship between data, so there are some limitations [16]. Deep learning is an important method in machine learning research, and people gradually began to use deep learning in the detection of mental illness. Lang *et al.* [17] used deep convolutional neural network (DCNN) to learn deep features from spectrum images. The results are combined with the classification results of shallow level features for the final depression prediction. Majtner *et al.* [18] and Sun *et al.* [19] proposed to combine deep features and shallow features in the decision-making level. Firstly, the deep features were extracted by the neural network, and then the classifiers were trained with deep features and shallow features respectively. Wang *et al.* [20] used deep feature and artificial feature fusion to reduce false-positive results. Compared with the traditional method, the feature fusion method has a better classification effect.

These methods have achieved good results in the diagnosis of mental health, but there are still some problems. First of all, in most studies above, the deep feature learning method and the traditional feature learning method are involved respectively. As we said before, the former suffers from a small sample problem, and the latter fails to obtain high-quality new features automatically. Second, even if the two kinds of features are considered, the nonlinear complex relationship between data is ignored. Only the decision-making level fusion of the original features and deep features is carried out in the studies. The features are not well fused, so they cannot well represent the class features of speech.

Relevant research [21]–[23] shows that high-level features (deep features) and low-level features (shallow features) reflect different side information of the target recognition, and they have good complementarity. Therefore, it is important to consider how to fuse the two types of features.

Autoencoder (AE) is a typical neural network, which has attracted more and more attention in recent years [24]–[26]. Stacked autoencoder (SAE) can realize stacking easily by taking the output of the last hidden layer of AE as the input of the next AE. On this basis, sparse stacked autoencoder (SSAE) can learn more representative features by introducing sparse constraint [27]–[29]. Therefore, the SSAE is used for deep feature learning here. Although SSAE has achieved some success in subsequent applications [30]–[33], feature fusion of SSAE and original features is still a challenging problem. First of all, deep features need to complement the original features. However, the existing deep autoencoder does not consider the original function in the training process and the hidden layer. Secondly, due to a large number of parameters, the deep autoencoder is easy to suffer from the overfitting problem, which limits the generalization ability of deep autoencoder in learning effective features. Therefore, it is necessary to introduce feature reduction and data enhancement in SSAE to further improve its ability to resist overfitting.

Based on this idea above, this paper proposes an embedded hybrid feature sparse stacked autoencoder (EHFSSAE) for feature fusion. The basic idea of EHFSSAE is to embed the original features into the coded output of each AE, and then fuse these mixed features into the more abstract feature representation of higher hidden layer, and retain some useful information for classification tasks. At the same time, the robustness of the network is improved.

At the same time, in order to solve the high-dimensional problem caused by the combination of two kinds of features, a new feature level fusion strategy called hybrid feature fusion model is proposed. In other words, the feature selection algorithm based on L1 regularization is used to select more discriminative and robust features among hybrid features.

Besides, in order to further eliminate the redundancy and improve the generalization ability of the proposed algorithm, the weighted local discriminant preserving projection (w_LPPD) and SVM are combined to construct an ensemble model (w_LPPD-SVM ensemble). W_LPPD is a new feature extraction method, which considers the outliers in the samples and effectively removes some samples far away from the class center. The idea of ensemble learning and dimension reduction is to improve accuracy. Through feature extraction of each base classifier, the diversity of the base classifiers is improved.

The main structure of this paper is as follows: section 1 introduced the background of this manuscript. Section 2 described the data and the proposed method. Section 3 elaborated groups of experiments and analyzed the results. Section 4 discussed the contributions of this paper and future work.

## II. THE PROPOSED METHOD

### A. DATA COLLECTION SCHEME

A total of 299 subjects were included in this data collection study. Among them, 130 patients are with schizophrenia, 67 patients are with depression, and both of them were collected from Chongqing mental health center. 102 healthy people were composed of graduate students from Chongqing University and employees of the Chengdu LanTu company. All subjects were diagnosed and screened by experienced psychological experts and psychiatrists according to the MINI International Neuropsychiatric Interview (MINI) [34] and Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [35].

Among all the subjects, the age of schizophrenia patients was 18-63 years old (mean ± std: 31.9 ± 10.6); the age of depression patients was 15-71 years old (mean ± std: 36.2 ± 14.2); the age of healthy people was 20-36 years old (mean ± std: 28 ± 4.5). As to gender, there are 58 men and 72 women with schizophrenia; 21 men and 46 women with depression; 62 men and 40 women with health status. All the subjects had no other mental diseases such as substance abuse, substance dependence, personality disorder, and no serious physical disease or suicidal behavior. The subjects were all above the level of primary school education. Please see TABLE 1 for details.

The experiment of data collection was conducted in a well quiet room. The linguistic data for different subjects were the same. The subjects sat about 1 m in front of a 21-inch computer screen. A piece of Chinese text would be displayed on the screen. The subjects needed to read the text carefully. The Chinese text contains 13 tasks, including continuous vowels, numbers, words, and short sentences. These sound segment types were selected by psychiatrists from a group of oral exercises corpus. Each speech sample was obtained by making some subjects finish one speech task. Please See TABLE 2 for details.

**TABLE 1.** Basic information of datasets.

| Class | Instances | Features | Age | Male | Female |
|---|---|---|---|---|---|
| health | 102 | 26 | 28±4.5 | 62 | 40 |
| schizophrenia | 130 | 26 | 31.9±10.6 | 58 | 72 |
| depression | 67 | 26 | 36.2±14.2 | 21 | 46 |

**TABLE 2.** Specific information of speech task.

| Task number | Speech task |
|---|---|
| 1 | a, e, i, o, u |
| 2 | miao, yuan, guang, qiao, suan |
| 3 | Chair |
| 4 | Mother |
| 5 | Radish |
| 6 | Tofu |
| 7 | Stone |
| 8 | What |
| 9 | Flower |
| 10 | Under the bridge in front of gate…… |
| 11 | This is a dog…… |
| 12 | Goose, goose, googse….. |
| 13 | Ah, spring is coming…… |

The recording is done by a Sony recorder in the frequency range of 50 Hz to 13 kHz. The recorder is set at 96 kHz, 30 dB, placed 10 cm away from the subject's mouth, and then the subject is asked to read the specified text. The schematic diagram is shown in FIGURE 1.

The data are saved in the form of. Wav (lossless compression format). On this basis, the Praat acoustic analysis software [36] is used for feature extraction, thereby generating the original features. Based on the previous related work of the authors, a set of 26 speech features including time-frequency are extracted from each speech sample. Please see TABLE 3 for detail.

### B. THE PROPOSED METHOD

As the analysis above, it is necessary to combine the original features and deep features. Since the number of speech samples is small, deep learning should be conducted on the original features rather than the original signal. In order to improve the complementarity between the deep features extracted and the original features, a stacked autoencoder fusion model with embedded deep and shallow features is designed in this paper for speech psychiatric recognition. The model consists of three main parts: the first part is to design an EHFSSAE, which embeds the original features into the hidden layer of the stack encoder so that the improved stacked autoencoder can learn high-quality deep features from the original features. The second part is the hybrid feature fusion

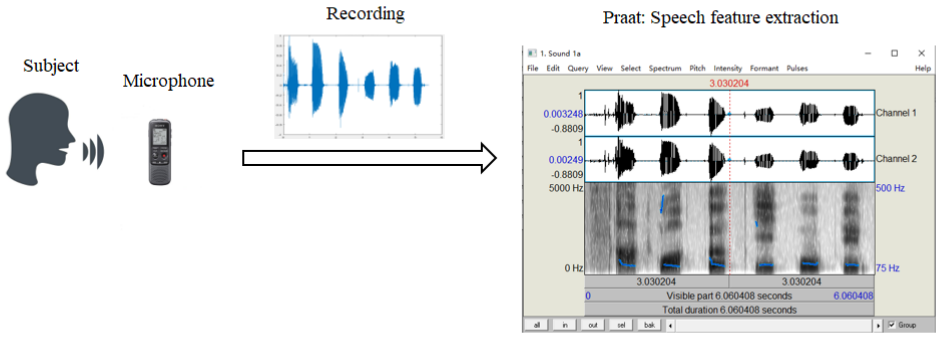
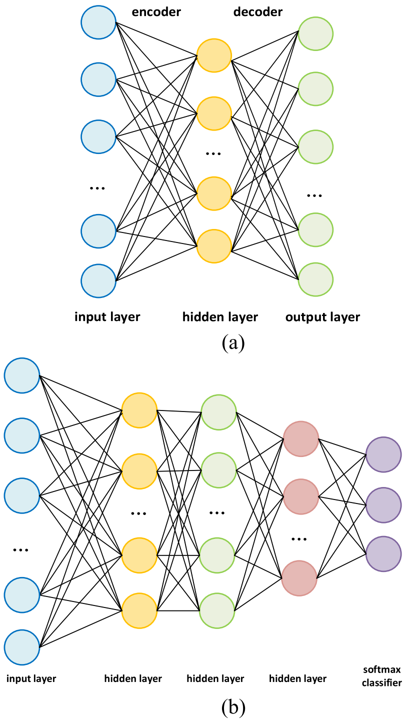

Recording
Praat: Speech feature extraction
Subject
Microphone
1. Sound 1a
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses Help
3.030204
0.003248
-0.8809
Channel 1
0.00249
-0.8809
Channel 2
5000 Hz
500 Hz
0 Hz
75 Hz
3.030204
3.030204
0
Visible part 6.060408 seconds
6.060408
Total duration 6.060408 seconds
all in out sel bak
Group

**FIGURE 1.** Schematic diagram of speech task acquisition.

**TABLE 3.** Audio characteristics of speech features.

| Features | Group |
|---|---|
| Jitter (local)<br>Jitter (local, absolute)<br>Jitter (rap)<br>Jitter (ppq5)<br>Jitter (ddp) | Frequency parameters |
| Shimmer (local)<br>Shimmer (local, dB)<br>Shimmer (apq3)<br>Shimmer (apq5)<br>Shimmer (apq11)<br>Shimmer (dda) | Amplitude parameters |
| Autocorrelation<br>Noise-to-Harmonic<br>Harmonic-to-Noise | Harmonicity parameters |
| Median pitch<br>Mean pitch<br>Minimum pitch<br>Maximum pitch<br>Standard deviation | Pitch parameters |
| Number of pulses<br>Number of periods<br>Mean period<br>Standard dev. of period | Pulse parameters |
| Fraction of locally unvoiced frames<br>Number of voice breaks<br>Degree of voice breaks | Voicing parameters |



(a)

(b)

**FIGURE 2.** (a) Autoencoder(AE); (b) stacked autoencoder(SAE).

mechanism based on L1 regularization. The third part is the dimensionality reduction ensemble model based on w_LPPD and SVM. The three-step processing method can effectively remove the feature redundancy, enhance the feature's discrimination capability, improve the reliability of the classification results, and increase the generalization ability and stability.

### 1) EMBEDDED HYBRID FEATURE SPARSE STACKED AUTOENCODER(EHFSSAE)

Autoencoder is a well-known and popular artificial neural network, which includes three main layers: input layer, hidden layer and output layer. It is composed of an encoder and decoder, as shown in FIGURE 2 (a). Due to the simple structure, the representative power of a single autoencoder is limited. Inspired by the biological model of the human visual cortex [37], the construction of a deep network helps to discover highly nonlinear and complex patterns in data [38]. The traditional SAE is composed of taking the outputs from the hidden units of the lower layer as the input to the upper layer's input units, as shown in FIGURE 2 (b).

The training of the stacked autoencoder is based on a greedy hierarchical unsupervised learning algorithm [39]. The key idea of this algorithm is to train one layer once by minimizing the reconstruction error in this layer. The representation of the $i$-th hidden layer is used as input for the $(i + 1)$-th hidden layer. However, such a structure will lead to unsatisfactory recognition ability of the coding features due to the small sample problem. We realize that the original

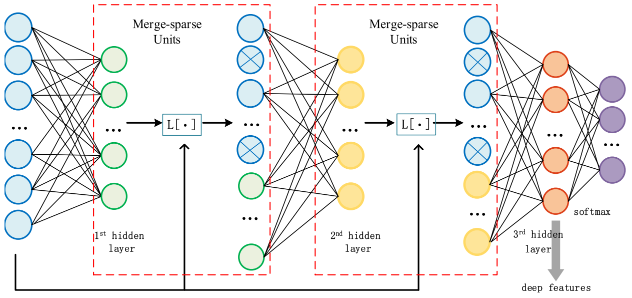


**FIGURE 3. Hybrid feature embedded sparse autoencoder.**

features contain useful information due to prior knowledge, which can be introduced into the deep network to maintain the initial information when the network goes deep. Therefore, a merged sparse unit (MSU) is designed between the encoded feature and the original feature. It can construct a EHFSSAE to filter the bad feature representation in the hidden layer, as shown in FIGURE 3.

MSU is an important part of the proposed EHFSSAE. Given the original feature samples $X \in R^{N\times n}$ and the encoded feature samples of autoencoder $H \in R^{N\times d}$, the purpose of MSE is to obtain the optimal subset of the X and H, thereby constructing the hybrid feature set. It can be defined as follows:

$$L(X \oplus H) = G^T(X \oplus H) \tag{1}$$

where $\oplus$ represents concatenating original features and feature representations in hidden layer, $L$ represents the sparse operation, and $G$ is the corresponding sparse matrix composed of 0 and 1. Objective function of sparse operation can be defined as:

$$\begin{aligned} &\max_{G} \quad tr(G^T(X \oplus H)(X \oplus H)^T G) \\ &s.t. \quad tr(G) = d \end{aligned} \tag{2}$$

where $d$ is the number of hidden units. The diagonal elements of the covariance matrix in formula (2) are sorted, and the $d$-th maximum value is selected as the threshold T. The elements of G can be defined as:

$$G_{ij} = \begin{cases} 1, & i = j, D_{ij} >= t \\ 0, & others \end{cases} \tag{3}$$

where $D_{ij}$ is the diagonal element of the covariance matrix. Through the sparse matrix, the features with low divergence will be zero, so these features will not be sent to the subsequent layer for further coding.

After introducing the MSU between the encoders, the encoder part of the k layer (k > 1) AE in EHFSSAE can be defined as:

$$H^{(k)} = f(W_{k1}L(X \oplus H^{(k-1)}) + b_{k1}) \tag{4}$$

where $H^{(k)} \in R^{N\times d^{(k)}}$ is the output of the $k$-th AE hidden layer, $W_{k1}$ and $b_{k1}$ are the weight matrix and deviation vector of the $k$-th AE, respectively. Decoder function is:

$$L'(X \oplus H^{(k)}) = g(W_{k2}H^{(k)} + b_{k2}) \tag{5}$$

where $W_{k2}$ and $b_{k2}$ are the weight matrix and bias vector, $L'(X \oplus H^{(k)})$ is the reconstruction function of $L(X \oplus H^{(k-1)})$.

In addition, the sparse criterion is applied to the hidden layer to discover latent structures in input data. Generally, Kullback-Leibler (KL) divergence as a tractable unsupervised objective is introduced in order to make representation sparse. It uses relative entropy to measure the difference between two Bernoulli random variables: $\hat{\rho}_j$ of the $j$-th hidden unit and average activation $\rho$ of the target. It is expressed as follows:

$$\begin{aligned} \sum_{j=1}^{d} KL(\rho||\hat{\rho}_j) &= \sum_{j=1}^{d} \rho \log(\frac{\rho}{\hat{\rho}_j}) + (1-\rho)\log(\frac{1-\rho}{1-\hat{\rho}_j}) \\ \hat{\rho}_j &= \frac{1}{N}\sum_{i=1}^{N} f^j(x^{(i)}) \end{aligned} \tag{6}$$

where $f^j(x^{(i)})$ is the activation value of the $i$-th input vector to the $j$-th unit of hidden layer.

The value increases monotonously along with the increasing difference value between $\rho$ and $\hat{\rho}_j$, and when $\hat{\rho}_j = \rho$, $KL(\rho||\hat{\rho}_j) = 0$. Therefore, most of the average outputs of hidden units are zeros by setting a small sparse parameter $\rho$. Thus sparse representation can be achieved. According to eq. (4-6), the training object function of $k$-th AE can be defined as follows:

$$\begin{aligned} \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} &||L(x_i \oplus h_i^{(k)}) - L'(x_i \oplus h_i^{(k)})||^2 \\ &+\lambda(||W_1||_2 + ||W_2||_2) \\ &+\beta(\sum_{j=1}^{d^{(k)}} KL(\rho||\rho_j)) \end{aligned} \tag{7}$$

where $\beta$ denotes regularization parameter of the sparsity constraint, $d^{(k)}$ is the number of $k$-th hidden layer units.

Training process with eq. (7) is called pre-training, after which the hidden layer of pre-trained autoencoders are cascaded one by one to form a stacked autoencoder, and its initial parameters are determined by pre-training. Focusing on the ultimate goal is to obtain features with better class representation ability, so we further optimize the whole network in a supervised manner. In order to achieve that, another classification layer is stacked on top of the stacked autoencoder as the output layer. The fine-tuning process of the stacked network is based on back-propagation with gradient descent. Because of the characteristics of pre-tune, the fine-tune flowing can reduce the risk of falling into the local optimum. The proposed EHFSSAE algorithm is summarized in Algorithm 1.

The learned nonlinear transformation by EHFSSAE can be regarded as a good feature learning. It can not only learn the potential relationships among data as a normal deep network does, but also improve the complementarity of the deep features and original features. After the training of the whole network for every input original feature vector $x_i = (x_{i1}, x_{i2}, \cdots, x_{in})$, a new feature vector can be obtained in each hidden layer and a different layer represents different level

**Algorithm 1** EHFSSAE

**Input**: training sample $X$

```
Setting parameters, including: $\lambda$, $\beta$, $\rho$, and the number of hidden
        units in every AE, the number of iteration.
Training the first AE with object function, then extract the output
        $H^{(1)}$ of hidden layer
for k =2: K (K is the total layer numbers of EHFSSAE )
    Calculate transformation matrix G by Eq. (2-3)
    Embed the original features into the features $H^{(k-1)}$ by Eq. (1)
    Train the $k$th AE with $L(X \oplus H^{(k-1)})$ as input according to eq. (7)
    extract the representation $H^{(k)}$ in hidden layer.
end
Stack the hidden layer and adding a softmax layer, of which the input is $H^{(k)}$
Fine-tune the whole network based on back-propagation with gradient decent, and the objective function is to minimize classification loss.
Extracting the output $H^{(k)}$ of final hidden layer in the
    fine-tuned HESSAE as deep feature $X_d$
```

**Output**: Deep features

information. Generally, the higher the layer in the network, the more complicated or abstract the patterns inherent in the input data. According to that, we take the outputs of the last hidden layer, which are the input of the classification layer, as the deep feature vector, recorded as $x_i' = (x_{i1}', x_{i2}', \cdots, x_{iq}')$. Then we construct an augmented feature vector $\hat{x}_i$ as hybrid features by concatenating $x_i$ and $x_i'$:

$$\hat{x}_i = (x_{i1}, \mathrm{x}_{i2}, \cdots, x_{id}, x_{i1}', x_{i2}', \cdots, x_{iq}') \in R^{(n+q)} \tag{8}$$

### 2) HYBRID FEATURE SELECTION ALGORITHM BASED ON L1 REGULARIZATION

Although our hybrid feature sets have more abundant category information, they will lead to high-dimensional problems. On the other hand, considering that deep features are learned from original features, we believe that these two sets of features are not independent of each other. In other words, there is some redundant information between the two groups of features. It is necessary to develop a new algorithm for effective feature reduction. Therefore, we design a feature reduction algorithm based on $L1$ regularization to optimizing the hybrid features.

Specifically, $L1$ regularization uses a penalty term to control the sum of the absolute values of the parameters to be small, thereby giving a sparse feature vector. For the new dataset $D = \{(\hat{x}_i, y_i)\}_{i=1}^N$, where $\hat{x}_i \in R^{n+q}$ denotes $i$-th sample with hybrid feature and $y_i$ is corresponding label. Considering the simplest regression model with the squared error as loss function, the optimization objective function can be defined as:

$$\arg\min_w \sum_{i=1}^{N} (y_i - \sum_{p=1}^{n+q} w_p \hat{x}_{ip})^2 \tag{9}$$

To prevent overfitting, $L1$ regularization is introduced to solve this problem:

$$\arg\min_w \sum_{i=1}^{N} (y_i - \sum_{p=1}^{n+q} w_p \hat{x}_{ip})^2 + \kappa \sum_{p=1}^{n+q} |w_p| \tag{10}$$

where $N$ is sample number, $\hat{x}_{ip}$ is $p$-th feature of $i$-th sample and $w_p$ is $p$-th feature's regression coefficients. $k$ is a sparsity control parameter. The larger it is, the sparser the model. Proximal Gradient Descent is used to solve Eq. (10), and the iteration of each step should be:

$$w^{(k+1)} = \arg\min_w \frac{L}{2} ||w - (w^{(k)} - \frac{1}{L} \frac{\partial(\sum_{i=1}^{N} (y_i - (w^{(k)})^T \hat{x}_i)^2)}{\partial w^{(k)}})||_2^2 + \kappa ||w||_1 \tag{11}$$

where $w = (w_1, w_2, \cdots w_{n+q})$, $L$ is a constant greater than zero.

Assuming $u = w^{(k)} - \frac{1}{L} \frac{\partial(\sum_{i=1}^{N} (y_i - (w^{(k)})^T \hat{x}_i)^2)}{\partial w^{(k)}}$, the closed-form solution of Eq. (11) can be calculated by:

$$w_p^{(k+1)} = \begin{cases} u_p - \kappa/L, & \kappa/L < u_p; \\ 0, & |u_p| \leq \kappa/L; \\ u_p + \kappa/L, & u_p < -\kappa/L; \end{cases} \tag{12}$$

In Eq. (12), $w_p^{(k+1)}$ and $u_p$ are the $p$-th component of $w^{(k+1)}$ and $u$ respectively. The result of solving the $L1$ regularization shows that only the features corresponding to the non-zero component of $w_p$ can be selected into the final feature subset.

### 3) FEATURE REDUCTION ENSEMBLE MODEL BASED ON W_LPPD AND SVM (W_LPPD-SVM ENSEMBLE)

W_LPPD is a novel effective feature reduction method proposed by the authors before. It considers outliers in the samples and removes some samples away from the center of the class. Firstly, it introduces random subspace sampling. Secondly, locality preserving discriminant projection is established based on the proposed objective function. Finally, multi-space mapping matrices are synthesized to construct the final mapping matrix. Assuming $k_{mc}$ denotes the number of samples sampling for $c$-th times, the total number of the samples after sampling is $k_m = \sum_{c=1}^{C} k_{mc}$. The local between-class scatter matrix $S_{LB}$ with the $k_m$ nearest neighbors of the center $\mu_{lb}$, and the local within-class scatter matrix $S_{LW}$ with the $k_{mc}$ nearest neighbors of the class center $\mu_{lwc}$ can be defined as follows:

$$S_{LB} = \sum_{c=1}^{C} N_{lc}(\mu_{lbc} - \mu_{lb})(\mu_{lbc} - \mu_{lb}) \tag{13}$$

$$S_{LW} = \sum_{c=1}^{C}\sum_{i=1}^{k_{mc}} (x_i^{(c)} - \mu_{lwc})(x_i^{(c)} - \mu_{lwc})^T \tag{14}$$

where local numbers $k_m = \lfloor r_b \cdot N \rfloor$ and $k_{mc} = \lfloor r_w \cdot N_c \rfloor$, $r_b$ and $r_w$ are sampling ratio coefficients, $N$ and $N_c$ are the number of total sample and $c$-th sample respectively. $\mu_{lb} = \frac{1}{k_m}\sum_{i=1,x\in N_{km}(m)}^{k_m} x_i$ is the center of local part for $S_{LB}$ computation, $\mu_{lb} = \frac{1}{k_m}\sum_{i=1,x\in N_{km}(m)}^{k_m} x_i$ is the center of the $c$-th local part for $S_{LB}$ computation, $N_{lc}$ is the number of the $c$-th class in local part, and $\mu_{lwc} = \frac{1}{k_{mc}}\sum_{i=1,x^{(c)}\in N_{k_{mc}}(m_c)}^{N_{lc}} x_i^{(c)}$ is the center of the $c$-th local class for $S_{LW}$ computation.

Furthermore, the locality preservation regularization term is shown as follows:

$$\begin{aligned}
&\min_W \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}\left\| W^T x_i - W^T x_j \right\|^2 \\
&= 2Tr(W^T \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}(x_i x_i^T - x_i x_j^T)W) \\
&\Leftrightarrow Tr(W^T \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}(x_i x_i^T - x_i x_j^T)W) \\
&= Tr(W^T X(D-A)X^T W) \\
&= Tr(W^T XLX^T W)
\end{aligned} \tag{15}$$

where $L$ is the Laplacian matrix, $D_{ii} = \sum_j A_{ij}$ is the diagonal matrix, and $A$ is the affinity matrix, which can be calculated in following manners:

$$A_{ij} = \begin{cases} 1, & if\ x_i \in N_k(x_j) || x_j \in N_k(x_i) \\ 0, & others \end{cases} \tag{16}$$

With Eq. (13-15) and the proposed w_LPPD can be formulated as:

$$\begin{aligned}
&\min_W \ Tr(W^T S_{LW} W) \\
&s.t. \ \ Tr(W^T S_{LB} W) - \gamma Tr(W^T XLX^T W) = \alpha I
\end{aligned} \tag{17}$$

where $\gamma$ represent the regularization coefficient, $\alpha$ is constant. It can be seen from the objective function that the w_LPPD aims to minimize the trace of local within-class scatter matrix and maximize the between-class scatter matrix while preserving the locality of the sample.

By introducing Lagrange multiplier $\gamma$, the objection function Eq. (17) finally can be written as:

$$\begin{aligned}
L(W,\lambda) = &\ Tr(W^T S_{LW} W) \\
&-\lambda(W^T S_{LB} W - \gamma W^T XLX^T W - \alpha I)
\end{aligned} \tag{18}$$

The optimal solutions are obtained by taking the partial derivation of $W$.

$$\begin{aligned}
&\frac{\partial L(W,\lambda)}{\partial W} \\
&= 0 \\
&\Rightarrow \frac{\partial \{Tr(W^T S_{LW} W) - \lambda(W^T S_{LB} W - \gamma W^T XLX^T W - \alpha I)\}}{\partial W} \\
&= 0 \\
&\Rightarrow Tr(2S_{LW} W - 2\lambda(S_{LB} W - \gamma XLX^T W)) = 0 \\
&\Leftrightarrow S_{LW} W = \lambda(S_{LB} - \gamma XLX^T)W \\
&\Rightarrow (S_{LB} - \gamma XLX^T)^{-1} S_{LW} W = \lambda W
\end{aligned} \tag{19}$$

Apparently, through Eq. (19), the projection matrix $W$ can be easily obtained by generalized eigenvalue decomposition. The vector $W_k = (w_1, w_2, \ldots, w_k)$ is comprised of the top $k$ eigenvectors of $W$. Then, the original data can be projected into a low dimension space spanned by the columns of $W_k$ to achieve dimensionality reduction. As mentioned before, we exploit w_LPPD on random subspace, so we can get $p$ projection matrixes $W_k^1, W_k^2, \cdot\cdot\cdot, W_k^p$. The final mapping matrix $W_k^F$ is obtained by weighting $W_k^1, W_k^2, \cdot\cdot\cdot, W_k^p$. Its mathematical expression as follows:

$$W_k^F = \alpha_1 W_k^1 + \alpha_2 W_k^2 + \cdots + \alpha_p W_k^p = \sum_{i=1}^{p} \alpha_i W_k^p \tag{20}$$

where $\alpha_i$ is the weight coefficient, it can be determined by grid search. Note that $\sum_{i=1}^{p} \alpha_i = 1$.

Through w_LPPD, we can further map the hybrid feature subset selected by $L1$ regularization to another low-dimension feature space, where the distance of samples from different classes will be farther, and the distance of samples from the same class will be closer. According to that, the features obtained by this way own more effective class representation ability.

In order to improve the generalization and reliability of the classification model, this paper uses ensemble learning to construct a fusion mechanism. Specifically, assuming the sampling ratio of samples and features are $\delta_1$ and $\delta_2$ respectively, and sampling $k$ times to form $k$ subsets. Then w_LPPD is applied on each subsets. The final $k$ training subsets obtained by w_LPPD will be fed into $k$ classifiers to train sub classifiers respectively. In this paper, SVM is used as base classifier. The classification result of validation samples will be decided by weighting voting mechanism. The weight of each classifier can be calculated according to the following formula:

$$w_k = \frac{\sum_{i=1}^{N_{train}} \phi(C_{ik}, y_i)}{N_{train}} \tag{21}$$

where $\phi(C_{ik}, y_i) = \begin{cases} 1, if\ C_{ik} = y_i \\ 0, others \end{cases}$, $N_{train}$ means the number of training set. Assuming the class number of dataset is $C$, for $i$-th sample $x_i$ with label $y_i$, $C_{ik}$ is the result of $k$-th classifier.

The probability of sample $x_i$ belongs to $c$-th class can be expressed as:

$$P(x_i \in x^c)|_{c=1}^{C} = \sum_{k=1}^{K} w_{kc}\phi(C_{ik}, c) \tag{22}$$

Then the final class label predicted by our ensemble model can be decided by following formula:

$$y_i' = \max_c\{P(x_i \in x^1), P(x_i \in x^2), \cdots, P(x_i \in x^C)\} \tag{23}$$

The proposed algorithm in this paper is summarized in Algorithm 2.

**Algorithm 2** The proposed algorithm

```
Input: training sample $X$
1: Learning deep features using EHFSSAE(refer to Algorithm 1),
        recorded as $X'$
2: Concatenating $X$ and $X'$ to construct hybrid features by Eq. (8)
3: Hybrid feature selection for $\hat{X}$ base on Eq. (9-11)
4: w_LPPD-SVM ensemble
5:   Set $\delta_1$, $\delta_2$.
6:   Sampling T times to form T subsets.
7:   Training t-th SVM:
8:   Training data: $t$-th subset, sampling number ns, number of subspace $p$, the regularization coefficients $\lambda$ and $\gamma$,
        the local numbers $km$ and $kmc$,the new subspace's dimension $k$.
9:   For i = 1:p
10:    Choosing ns training samples randomly.
11:    Calculating the $\mu$lb, $\mu$lbc and $\mu$lwc
12:    Calculating the scatter matrix $SLB$ and $SL$W based on
          formula (13) and (14)
13:    Constructing the affinity matrix $\mathbf{A}$ based on formula (16).
14:    Calculating the diagonal matrix $\mathbf{D}$ and the Laplacian matrix $\mathbf{L}$.
15:    Solving the mapping matrix $\mathbf{W}$ by Eq. (19)
16:  End for
17:    Searching for the optimal weight of Eq. (20) to obtain final mapping matrix $W_K^F$.
18:    Mapping $t$-th subset to training SVM
   19: Calculating weight of $t$-th classifier by Eq. (21)
20:  Getting the ultimate class label by Eq. (22-23)
   Output: Predicted label
```

## III. EXPERIMENTAL RESULTS AND ANALYSIS

### A. EXPERIMENTAL CONDITIONS

In the experimental part, there are three classes of our own datasets. In order to make full use of the data we collected, and better verify our algorithm, our data was divided into four datasets and several sets of experiments were performed to verify the proposed method. The four datasets are healthy people and depression patients (HD), healthy people and schizophrenia patients (HS), depression and schizophrenia patients (DS), and healthy people, depression and schizophrenia patients (HDS). The datasets are about two-class classification and three-class classification. Brief information about the dataset is shown in TABLE 4.

**TABLE 4.** The data set partition used in this experiment.

| Datasets | Instance | Original features | Class |
|---|---|---|---|
| Health and Depression(HD) | 16 | 26 | 2 |
| Health and Schizophrenia(HS) | 232 | 26 | 2 |
| Depression and Schizophrenia(DS) | 197 | 26 | 2 |
| Health, Depression and Schizophrenia(HDS) | 299 | 26 | 3 |

All experiments are carried out in a unified experimental environment: the experimental operating system is Windows 10, 64-bit, and the memory size is 128GB. The programming tool is MATLAB, version R2018b.

There are many model evaluation indicators used in speech diagnosis of mental health. In this paper, classification accuracy (Acc) is used as the model evaluation indicators. It is constructed from the confusion matrix, which stores the number of correctly and incorrectly classified examples in each class. The form of the confusion matrix is shown in Table 5.

Form the confusing matrix, the classification accuracy can be defined as:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \tag{24}$$

In our experiment, the structure and parameter setting of EHFSSAE model are shown in TABLE 6. The main parameters of the EHFSSAE ensemble model include regularization parameter $\lambda$, $\beta$ and sparse parameter $\rho$. Because our data set is too small, and the overfitting risk is high, so we set a small iteration value of 500 to stop fine tuning.

For the ensemble model, local ratio parameter in w_LPPD are set as $r_b = 0.9, r_w = 0.9$, the sampling ration $\delta_1$=0.7, $\delta_2$=0.5, and the number of base classifier $K$=5. In experiment, we use the hold-out cross-validation method for verification. That means for all four datasets, the labeled samples were split into two subsets, one accounted for one-third of all samples as test data, and the rest as train data. In order to eliminate the influence of accidental factors, each experiment was repeated five times to calculate the average accuracy and standard deviation as the final performance.

### B. EXPERIMENTAL RESULTS AND ANALYSIS

#### 1) EFFECTIVENESS OF PROPOSED ALGORITHM

In order to verify the effectiveness of the proposed algorithm, we compare the proposed hybrid feature selection based

**TABLE 5.** Confusion matrix.

| Prediction label<br>Real label | positive | negative |
|---|---|---|
| Positive | True Positive(TP) | False Negative(FN) |
| Negative | False Positive(FP) | True Negative(TN) |

**TABLE 6.** Parameter setting of EHFSSAE.

| Algorithm Symbol | Meaning | Parameter settings |
|---|---|---|
| $\lambda$ | regularization parameter | $10^{-5}$,$10^{-4}$,$10^{-3}$,$10^{-2}$,$10^{-1}$,1 |
| $\beta$ | regularization parameter | 1,2,3,4,5,6 |
| $\rho$ | sparsity parameter | 0.02,0.04,0.06,0.08,0.10,0.12 |
| Hidden units | hidden layer units in EHFSSAE algorithm | 90 - 40 - 20 |

**TABLE 7.** Classification accuracy of different algorithms in mental health speech dataset.

| Datasets | HD | HS | DS | HDS |
|---|---|---|---|---|
| PCA(%) | 87.5±3.3 | 81.4±4.2 | 67.7±3.8 | 74.3±2.9 |
| LDA(%) | 87.1±2.8 | 81.5±2.4 | 66.9±3.2 | 74.1±3.9 |
| Relief(%) | 86.7±3.1 | 81.0±2.5 | 66.3±2.9 | 73.2±3.2 |
| P_value(%) | 86.1±3.5 | 80.5±4.1 | 66.7±4.5 | 72.5±5.3 |
| HF&L1 (%) | **88.7±4.1** | **82.6±3.9** | **68.9±3.5** | **75.8±4.7** |

**Notes** HF&L1: Hybrid feature selection with L1 regularization.

on L1 regularization with the representative feature learning methods. These methods include feature selection and extraction algorithms: Relief [40], P_value [41], PCA [42], and LDA [43]. Considering that the basic classifier in our method is SVM, so we also use SVM as a classifier to evaluate the methods above for fairness. The average accuracy of the experiment is shown in TABLE 7.

The results in TABLE 7 show that the proposed L1 regularization based feature selection algorithm is superior to the traditional methods. Regardless of different datasets, this method has the best accuracy, and the improvement can achieve 3.3%. The results show that the method can effectively reduce the redundancy of the hybrid features.

In order to verify the feature extraction ability of the EHFSSAE, we compare it with the basic SAE and SSAE. To ensure fairness, the three autoencoders are composed of three hidden layers and one softmax layer, and the regularization and sparse parameters are set to the same value. The classification accuracy of the three autoencoders is shown in TABLE 8.

As seen in TABLE 8, the classification effect of the proposed EHFSSAE algorithm is better than SAE and SSAE. Besides, the standard deviation of the proposed autoencoder is the best, which means it is the most stable. The main reason may be that the original features are embedded in the network structure and training, which improves the complementarity of the two groups of features.

**TABLE 8.** Classification accuracy of different deep autoencoders.

| Datasets | HD | HS | DS | HDS |
|---|---|---|---|---|
| SAE(%) | 83.4±6.4 | 79.5±7.1 | 62.1±5.1 | 72.1±3.9 |
| SSAE(%) | 86.5±5.3 | 81.4±5.8 | 64.6±5.2 | 73.5±3.3 |
| EHFSSAE(%) | **89.0±3.6** | **82.6±3.1** | **69.2±4.3** | **75.8±2.5** |

In order to verify that the hybrid features learned by EHFSSAE can be regarded as the potential representation of discriminative information hidden in the data, we designed experiments using only shallow features or deep features for comparison. In order to verify the effectiveness of the EHFSSAE and L1 regularization methods, we consider hybrid feature (HF), HF with L1 algorithm (HF&L1), and HF&L1 with ensemble model (HF&L1&Ensemble, that is, the complete proposed method) for comparison by ablation method. The results are shown in TABLE 9.

**TABLE 9.** Classification accuracy of the major parts in the proposed method.

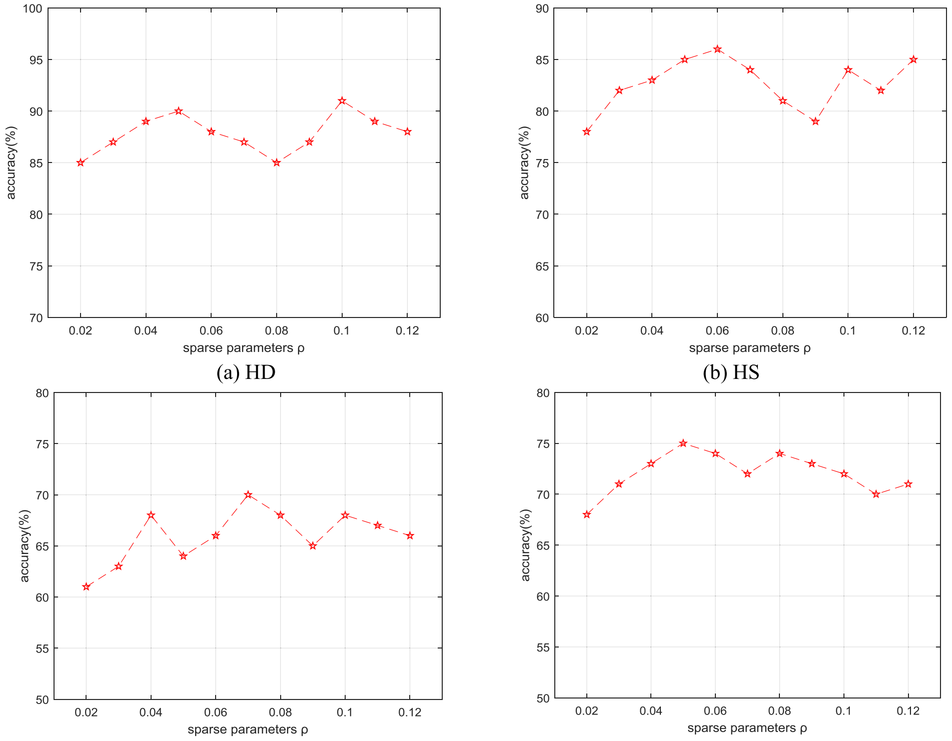
| Datasets | HD | HS | DS | HDS |
|---|---|---|---|---|
| OF (%) | 86.7±1.4 | 78.2±1.2 | 65.3±2.3 | 72.6±2.7 |
| DF (%) | 79.2±2.1 | 71.0±1.9 | 62.1±2.5 | 66.2±3.7 |
| HF (%) | 82.0±1.8 | 75.6±2.3 | 63.2±2.9 | 71.5±3.3 |
| HF&L1 (%) | 86.7±3.2 | 80.2±2.3 | 67.4±2.8 | 74.1±3.5 |
| HF&L1& Ensemble (%) | **89.0±4.1** | **82.6±3.9** | **69.2±3.5** | **75.8±4.7** |

**Notes** OF: Original feature; DF: Deep features learned by EHFSSAE; HF: Hybrid feature (Cascade of original features and deep features); HF&L1: Hybrid feature selection with L1 regularization; HF&L1&Ensemble : HF&L1 followed by w_LPPD-SVM ensemble.

As shown in TABLE 9, the results show that the effect of only using deep features is not good. However, if the original data and depth features are simply combined, the performance is not as good as using only the original features. This may be due to the high dimension and redundancy brought by simple combination. At the same time, experimental results show that the feature fusion based on L1 regularization and w_LPPD-SVM ensemble model can effectively remove redundancy, retain effective classification information, and improve the accuracy and stability of classification.

In order to verify the performance improvement of our ensemble model, we compare it with the random forest (RF) and extreme learning machine (ELM). The results are shown in TABLE 10.

As shown in TABLE 10, the proposed ensemble model maximizes the classification accuracy. The classification performance is improved by at least 1.5% compared with the popular classifier. Besides, the standard deviation of the ensemble model is the best, and it means that the proposed ensemble model is more stable. One of the possible reasons is that w_LPPD can conduct high-quality feature reduction. The second possible reason is that the ensemble model based on bagging has good complementarity of base classifiers.



(a) HD

(b) HS

(c) DS

(d) HDS

**FIGURE 4. Classification accuracy on datasets with different sparse parameter.**

**TABLE 10. Classification accuracy under different classifiers.**

| Datasets | HD | HS | DS | HDS |
|---|---|---|---|---|
| RF (%) | 86.7±7.2 | 78.2±6.4 | 66.7±5.7 | 72.4±6.5 |
| ELM (%) | 87.5±5.6 | 79.0±5.3 | 67.2±4.6 | 73.5±4.2 |
| Ensemble model (%) | **89.0±4.1** | **82.6±3.9** | **69.2±4.7** | **75.8±3.5** |

**Notes** Ensemble model: w_LPPD-SVM ensemble model

### 2) PARAMETER ANALYSIS OF EHFSSAE

The design of EHFSSAE model is the key to this proposed method. Therefore, we evaluated the impact of the parameters on the classification accuracy of EHFSSAE. Firstly, we studied the influence of the sparse parameters on classification accuracy. Generally, the sparse parameter is a small value close to zero. In our experiment, it is selected in the range of 0.02 to 0.12. The classification results of different sparse parameters on our mental health speech dataset are shown in Figure 4.

It can be seen from the results that the sparse constraint has a significant impact on accuracy. As seen in FIGURE 4, we can see that with the increase of sparse parameter $\rho$, the classification accuracy has obvious fluctuation. However, they are generally showing a trend of first rising and then decreasing. Besides, when the $\rho$ is around 0.5, it tends to be the best. It means that the parameter has an apparent effect on accuracy.

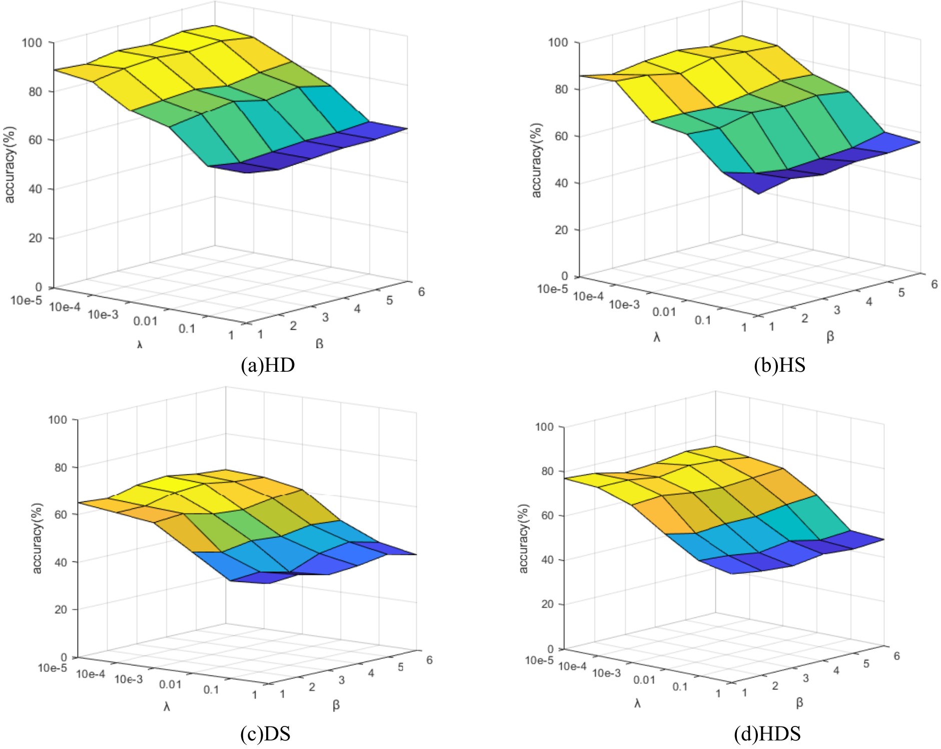
Penalty parameter $\lambda$ and $\beta$ are the relevant penalty parameters of the EHFSSAE. In order to study the effect of parameters on the performance of proposed EHFSSAE, we combine them for joint analysis. According to the prior knowledge, the range of $\lambda$ is set $10^{-5}$ to 10, and the range of $\beta$ is 1-6. The relationship between the parameters and classification accuracy is shown in FIGURE 5.

We can see that the regularization parameter $\lambda$ of weight attenuation is very important in our EHFSSAE. When the range of $\lambda$ is from $10^{-5}$ to $10^{-3}$, the classification effect is stable. When $\lambda$ is greater than 0.01, the classification accuracy will decrease sharply. The possible reason for this phenomenon is that too much penalty will lead to too many connections with zero weight. At the same time, we can see that for a fixed $\lambda$, $\beta$ has a relatively small impact on the classification results, which is almost negligible.

### 3) COMPARISON WITH RELEVANT ALGORITHMS

In order to verify the effectiveness of the proposed method, the methods in reference [12], reference [15], and reference [17] are compared with the proposed method. The compared algorithms are representative speech recognition algorithms of mental health. The results are shown in TABLE 11.

Reference [12] uses the classification algorithm of logical regression. It can be seen that the classification results for HD and HS datasets are acceptable, but the classification result for the DS datasets is very poor. The possible reason is that logistic regression is a kind of linear regression and cannot



(a)HD

(b)HS

(c)DS

(d)HDS

**FIGURE 5.** **Effect of parameter and on classification performance.**

**TABLE 11.** **Comparison of different classification methods based on mental health speech dataset.**

| Datasets | HD | HS | DS | HDS |
|---|---|---|---|---|
| Reference [12] | 81.9±5.4 | 75.9±4.6 | 58.6±4.8 | —— |
| Reference [15] | 86.4±4.3 | 78.5±3.5 | 65.6±4.2 | 71.2±2.8 |
| Reference [17] | 84.5±5.7 | 78.2±5.3 | 63.7±6.4 | 70.4±3.5 |
| Proposed method | **89.0±4.1** | **82.6±3.9** | **69.2±4.7** | **75.8±3.5** |

effectively distinguish similar classes. In reference [15], PCA is used as feature dimension reduction, and KNN, GMM, and SVM are used as classifiers. The best results are shown in TABLE 11. The classification effect is acceptable, but compared with the proposed method, it is still worse. The possible reason is that PCA only reduces the dimension of features, and does not obtain high-quality speech features for classification, so it can not represent our mental disease class information well. Reference [17] uses DCNN to extract deep features and original audio features from the spectrogram, but the effect is not as good as the proposed method. The reason may be that only deep features are not suitable for the small sample problem.

From the analysis and results above, it can be seen that the reference [12], [15] only uses the shallow features of speech, and does not effectively mine the deep information of speech data. Only the shallow features cannot fully represent the class labels. In reference [17], although DCNN is used to extract the deep features from the spectrogram, deep features are not effectively fused with shallow features. So the classification accuracy was not further improved due to the small sample problem. The proposed algorithm effectively fuses the deep and shallow features in network training, eliminates the redundancy of features, obtains the most representative features, and improves the classification accuracy.

## IV. CONCLUSION

This paper systematically and completely explained the methods of mental health speech data collection, feature learning and recognition. In order to solve the problem of mental health speech diagnosis, this paper constructs a mental health speech data set and proposes a new mental health speech recognition algorithm: speech recognition algorithm of mental health based on embedded hybrid feature stacked sparse autoencoder ensemble.

This paper has the following contributions and innovations mainly:

1) We collected and constructed a Chinese mental health speech data set by ourselves, which solved the problem of insufficiency related speech data set.

2) A new stacked autoencoder EHFSSAE is designed to extract deep features with more complementarity. EHFSSAE filters some bad features learned by the previous layer in the pre-training by embedding the original features. Compared with the standard stacked autoencoders, the robustness of deep feature is improved.

3) Aiming at the high-dimensional small sample problem caused by the combination of deep features and shallow features, a dimension reduction algorithm based on L1 regularization for feature selection is designed.

4) A dimension reduction ensemble model based on w_LPPD and SVM is designed. The model can effectively reduce the dimension and improve the accuracy, stability and generalization ability of classification.

5) By combining the EHFSSAE, L1 regularization method, and w_LPPD & SVM ensemble model, a new recognition algorithm for mental health speech recognition is proposed.

6) This paper for the first time proposes a speech collection scheme for mental health recognition. The scheme completed data collection and constructed a large-scale Chinese mental health speech database to verify the proposed mental health recognition algorithm.

Although the proposed method is effective, there is still some work to be done in this research. The future work is to optimize the structure or training method of the stacked autoencoder to further enhance the quality of mental health speech features and improve the classification accuracy. Besides, other deep neural networks need to be considered to further verify and improve the proposed method.

## ACKNOWLEDGMENT

*(Hong Chen and Yuan Lin contributed equally to this work.)* The authors would like to thank the editor and reviewers for their valuable comments and suggestions.

They would also like to thank those individuals (or institutions) that have provided data support for this research.

## REFERENCES

[1] R. S. Kern, M. F. Green, B. D. Marshall, W. C. Wirshing, D. Wirshing, S. McGurk, S. R. Marder, and J. Mintz, "Risperidone vs. Haloperidol on reaction time, manual dexterity, and motor learning in treatment-resistant schizophrenia patients," *Biol. Psychiatry*, vol. 44, no. 8, pp. 726–732, Oct. 1998.

[2] A. Wolkin, M. Sanfilipo, A. P. Wolf, and B. Angrist, "Negative symptoms and hypofrontality in chronic schizophrenia," *Arch. Gen. Psychiatry*, vol. 49, no. 2, pp. 959–965, Dec. 1992.

[3] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[4] Z. Liu, H. Kang, L. Feng, and L. Zhang, "Speech pause time: A potential biomarker for depression detection," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 2020–2025.

[5] B. Stasak, J. Epps, and A. Lawson, "Analysis of phonetic markedness and gestural effort measures for acoustic speech-based depression classification," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Oct. 2017, pp. 165–170.

[6] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apaydin, and O. Kursun, "Collection and analysis of a parkinson speech dataset with multiple types of sound recordings," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 4, pp. 828–834, Jul. 2013.

[7] R. J. Holmes, J. M. Oates, and D. J. Phyland, "Voice characteristics in the progression of Parkinson's disease," *Int. J. Lang. Commun. Disorders*, vol. 35, no. 3, pp. 407–418, Sep. 2000.

[8] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski, "Analysis of acoustic space variability in speech affected by depression," *Speech Commun.*, vol. 75, pp. 27–49, Dec. 2015.

[9] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: A comparison between spontaneous and read speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7547–7551.

[10] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *J. Multimodal User Interfaces*, vol. 7, no. 3, pp. 217–228, Sep. 2013.

[11] V. Mitra, E. Shriberg, D. Vergyri, B. Knoth, and R. M. Salomon, "Cross-corpus depression prediction from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4769–4773.

[12] W. Pan, "Depression recognition based on speech," *Sci. Bull.*, vol. 63, no. 20, pp. 2081–2092, Sep. 2018.

[13] M. Asgari, I. Shafran, and L. B. Sheeber, "Inferring clinical depression from speech and spoken utterances," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2014, pp. 21–24.

[14] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller, "CCA based feature selection with application to continuous depression recognition from acoustic speech features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4–9.

[15] H. Jiang, B. Hu, Z. Liu, L. Yan, T. Wang, F. Liu, H. Kang, and X. Li, "Investigation of different speech types and emotions for detecting depression using different classifiers," *Speech Commun.*, vol. 90, pp. 39–46, Jun. 2017.

[16] Y. Liu, S. Zhao, Q. Wang, and Q. Gao, "Learning more distinctive representation by enhanced PCA network," *Neurocomputing*, vol. 275, pp. 924–931, Jan. 2018.

[17] L. He and C. Cao, "Automated depression analysis using convolutional neural networks from speech," *J. Biomed. Informat.*, vol. 83, pp. 103–111, Jul. 2018.

[18] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg, "Combining deep learning and hand-crafted features for skin lesion classification," in *Proc. 6th Int. Conf. Image Process. Theory, Tools Appl. (IPTA)*, Dec. 2016, pp. 1–6.

[19] W. Sun, S. Xie, J. Yu, L. Niu, and W. Sun, "Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 919–923.

[20] C. Wang, A. Elazab, J. Wu, and Q. Hu, "Lung nodule classification using deep feature fusion in chest radiography," *Computerized Med. Imag. Graph.*, vol. 57, pp. 10–18, Apr. 2017.

[21] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, "Query-adaptive late fusion for image search and person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1741–1750.

[22] X. Zhang, H. Zhang, Y. Zhang, Y. Yang, M. Wang, H. Luan, J. Li, and T.-S. Chua, "Deep fusion of multiple semantic cues for complex event recognition," *IEEE Trans. Image Process.*, vol. 25, no. 3, pp. 1033–1046, Mar. 2016.

[23] Z. Liu, S. Wang, L. Zheng, and Q. Tian, "Robust ImageGraph: Rank-level feature fusion for image search," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3128–3141, Jul. 2017.

[24] Y. Qi, Y. Wang, X. Zheng, and Z. Wu, "Robust feature learning by stacked autoencoder with maximum correntropy criterion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 6716–6720.

[25] S. Gao, Y. Zhang, K. Jia, J. Lu, and Y. Zhang, "Single sample face recognition via learning deep supervised autoencoders," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 10, pp. 2108–2118, Oct. 2015.

[26] X. Yuan, B. Huang, Y. Wang, C. Yang, and W. Gui, "Deep learning-based feature representation and its application for soft sensor modeling with variable-wise weighted SAE," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3235–3243, Jul. 2018.

[27] F. Lv, M. Han, and T. Qiu, "Remote sensing image classification based on ensemble extreme learning machine with stacked autoencoder," *IEEE Access*, vol. 5, pp. 9021–9031, May 2017.

[28] C. Tao, H. Pan, Y. Li, and Z. Zou, "Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2438–2442, Dec. 2015.

[29] G. Jiang, H. He, P. Xie, and Y. Tang, "Stacked multilevel-denoising autoencoders: A new representation learning approach for wind turbine gearbox fault diagnosis," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 9, pp. 2391–2402, Sep. 2017.

[30] J. Dai, H. Song, G. Sheng, and X. Jiang, "Cleaning method for status monitoring data of power equipment based on stacked denoising autoencoders," *IEEE Access*, vol. 5, pp. 22863–22870, Aug. 2017.

[31] Y. Lei, W. Yuan, H. Wang, Y. Wenhu, and W. Bo, "A skin segmentation algorithm based on stacked autoencoders," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 740–749, Apr. 2017.

[32] Y. Qi, C. Shen, D. Wang, J. Shi, X. Jiang, and Z. Zhu, "Stacked sparse autoencoder-based deep network for fault diagnosis of rotating machinery," *IEEE Access*, vol. 5, pp. 15066–15079, Jul. 2017.

[33] L. Wang, Z. Zhang, and J. Chen, "Short-term electricity price forecasting with stacked denoising autoencoders," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2673–2681, Jul. 2017.

[34] I. M. Van Vliet and E. De Beurs, "The MINI-international neuropsychiatric interview. A brief structured diagnostic psychiatric interview for DSM-IV en ICD-10 psychiatric disorders," *Tijdschrift Voor Psychiatrie*, vol. 49, no. 6, pp. 393–397, Jul. 2007.

[35] A. M. Vijay "Diagnostic and statistical manual of mental disorders," *Psychiatry Res.*, vol. 189, no. 1, pp. 158–159, Jan. 2011.

[36] C. Gorris, A. R. Maccarini, F. Vanoni, and M. Poggioli, "Acoustic analysis of normal voice patterns in Italian adults by using praat," *J. Voice*, vol. 34, no. 6, pp. 9–18, May 2019.

[37] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[38] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *J. Mach. Learn. Res.*, vol. 1, pp. 1–40, Jan. 2009.

[39] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[40] L. Jianhai, L. U. Kai, and S. Yuxuan, "A novel relief feature selection algorithm based on mean-variance model," *J. Inf. Comput.al Sci.*, vol. 8, no. 16, pp. 3921–3929, Dec. 2011.

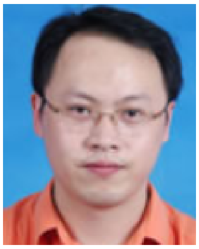
[41] D. Donoho and J. Jin, "Higher criticism thresholding: Optimal feature selection when useful features are rare and weak," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 39, pp. 14790–14795, Oct. 2008.

[42] I. T. Jolliffe, "Principal component analysis," *J. Marketing Res.*, vol. 87, no. 4, p. 513, Aug. 2002.

[43] C.-H. Li, B.-C. Kuo, and C.-T. Lin, "LDA-based clustering algorithm and its application to an unsupervised feature extraction," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 1, pp. 152–163, Feb. 2011.

**HONG CHEN** received the M.M. degree from the Chengdu University of TCM, Chengdu, China, in 2008. She works as a Deputy Chief Physician with the Chongqing University Cancer Hospital. Her current research interests include the prevention and rehabilitation treatment of malignant tumor, TCM tumor intelligent research, and TCM tumor precision treatment.

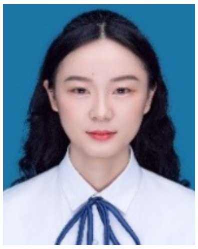
**YUAN LIN** received the bachelor's degree in electronic information engineering from Chongqing University, Chongqing, China, in 2018, where he is currently pursuing the master's degree. His research interests include feature learning, fusion of traditional machine learning, and deep learning.



**YONGMING LI** (Member, IEEE) received the B.S. degree from the University of Electronic Science and Technology of China, in 1999, and the M.E. and Ph.D. degrees from Chongqing University, Chongqing, China, in 2003 and 2007, respectively. He is currently working as an Associate Professor with Chongqing University. His current research interests include medical signal and information processing, artificial intelligence, and data analysis and mining.

**WEI WANG** received the M.M. and Ph.D. degrees in integrated traditional Chinese and Western medicine and internal medicine from Chongqing Medical University, Chongqing, China, in 2008 and 2015, respectively. She currently works as the Director of Cancer Treatment Center of Traditional Chinese Medicine, Chongqing University Cancer Hospital, and the Oncology Department, Traditional Chinese Medicine Hospital. She is also the Master Tutor with the School of Medicine and the Bioengineering College, Chongqing University. Her current research interests include the prevention and rehabilitation treatment of malignant tumor, TCM tumor intelligent research, and TCM tumor precision treatment.

**PIN WANG** received the M.S. degree from Chongqing University, China, in 2003, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2008. She held a postdoctoral position with Nanyang Technological University, from 2008 to 2009. From 2009 to 2010, she was funded by the National Energy Laboratory Postdoctoral Fund to conduct postdoctoral research at the University of Pittsburgh. She is currently an Associate Professor with Chongqing University. Her current research interests include hyperspectral imaging and detection, intelligent information processing, and big data analysis decision.



**YAN LEI** received the bachelor's degree in communication engineering from the Wuhan University of Science and Technology, China, in 2018. She is currently pursuing the master's degree with Chongqing University, Chongqing, China. Her research interests include feature learning, and fusion of traditional machine learning and deep learning.