ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20220124122948
Original scientific paper

Intelligent Diagnosis Approach for Depression Using Vocal Source Features

Yuan GAO, Yinan XIN, Li ZHANG*

Abstract: Depression is the most widely affecting of mental illnesses for public health concern. Although there are many treatments for depression, barriers to diagnosis still exist. The intelligent diagnosis relying on extraction of biomarkers provides reliable indicators of depression. This paper proposed a new method of machine learning diagnosis based on vocal source features. The short-term and long-term features were combined for classification and evaluation. The long-term features contained four important short-term features selected by decision trees, and the random forest algorithm and extreme gradient boosting algorithm were used for classification. The results showed that our method was feasible to classify the degree of depression, F1 scores and sensitivity of non-depression were better than traditional short-term features, long-term features, and deep learning approaches. Our study provides a useful tool for preventing and diagnosing early depression.

Keywords: depression; feature combination; intelligent diagnosis; random forest algorithm

1 INTRODUCTION

Depression is a typical affective disorder which can lead to physical dysfunction. Major depressive disorder would even cause serious disability and increase the risk of mortality. The 2019 report from Healthy China Initiative showed that morbidity rate of depression in China was 2.1% and taking on a growth trend [1]. Although there are many effective treatments for depression, barriers to screening and diagnosis still exist. Thus early diagnosis of depression and improving its diagnostic accuracy is an important problem to be solved. The clinical diagnosis of depression was mainly based on facial expressions, voice, eye movements and emotional information [2, 3]. Accurate diagnosis required special equipment and professional doctors for measurement and evaluation, which made the cost high for patient to obtain medical service. Recently, the researches on finding simple and effective diagnostic approaches have been a new hotspot and have received conspicuous achievements. One of the methods is automatic monitor relying on the extraction of biomarkers to provide reliable indicators of depression [4]. Vocal source biomarkers for depression are easy to obtain and contain rich emotional information, thus were suitable for capturing depression symptom heterogeneity [5-7]. Previous study has suggested that high levels of depression could lead to pitch variations, increase the degree of aspiration [8]. Other acoustic measures such as jitter and shimmer have also been shown to classify the depression [9]. On the Audio/Visual Emotion Challenge Conference 2016, Valstar et al. used short-term features with linear support vector machine algorithm based on stochastic gradient descent, making the recognition rate of depression reach 85% [10]. Li et al. weakened individual differences by multi-scale audio differential normalization and used deep regression model to predict the score of the depression scale with smaller mean square error [11]. However, the depression category classifications of the above experiments were not effective. The main problem is that it is more difficult to see the features of the audio only by rhyming features, the representation capability is not enough, the rhyming features combined with machine learning methods for classification often have the problem of under-fitting. While spectral based correlation features (such as MFCC, etc.) combined with deep learning methods would have the over-fitting problem. Moreover, none of them considered gender characteristics. Recent research pointed out that there was gender difference in audio emotion recognition [12]. Thus, in this study we added gender features to determine whether it is a key factor to improve the recognition rate of non-depression categories while ensuring the recognition rate for depressed categories and ing the overall performance of the classifier.

For the problem of intelligent diagnosis of depression, we also proposed a new depression recognition method based on audio signal. The method used feature combination to construct a new long-term feature vector as classification feature, and used the random forests (RF) algorithm [13] to construct a classification model, applied the eXtreme Gradient Boosting (XGBoost) algorithm [14] as a comparison to verify whether the classification results were specific to a particular model. This study showed that our method improved the F1 scores and sensitivity of non-depression classes was better than traditional short-term features, long-term features and deep learning. The results can be useful for preventing and diagnosing early depression.

2 AUDIO DATASETS AND DATA PREPROCESSING

2.1 Datasets

The audio data and associated depression indicators we used were provided by the Distress Analysis Interview Corpus(DAIC-WOZ) datasets [15], which is also the datasets used by AVEC. The datasets collected audio data by talking to subjects through a virtual interviewer and contained a total of 189 audio data, with the average length of the audio data being 16 minutes. Subjects had completed a Psychiatric Questionnaire (PHQ-8) [16] prior to participating in the interviewer. The maximum score of the PHQ-8 is 30, 15 is a cut-off value, equal to and greater than 15 is rated as depressed patients, and below 15 is non-depressed patients. The datasets contain voice data and corresponding PHQ-8 scores for 107 available subjects.

2.2 Data Preprocessing

First of all, the interviewer's voice and silent segments were removed from the audio data. For the new audio data the pyAudioAnalysis library [17] was used for the feature extraction operation. The sampling frequency was taken at

16 kHz due to the frequency of the recording microphone being 16 kHz in the interviews.

Combinations of short-term features and long-term features will be used for classification and evaluation, with the long-term features being derived from combinations of selected short-term features. Firstly, the short-term features are extracted by adding a Hamming window with a window size of 50 ms. In order to make a smooth transition between each frame and maintain its continuity, the step size is taken to be 25 ms, the overlapping part of each frame is 50%. In total, 68 short-term features were extracted, such as zero-cross rate, short-time energy, energy entropy, and spectral centre.

If all the features are used directly in the classifier, it will not only be time consuming but also affect the classification effect [18]. There exists power relationship between generated features and short-term features. The base number of power function is 3. So 4 short-term features combined could generate 81 features, if more short-term features are selected, it is easy to cause overfitting problems. Therefore, using the decision tree method, the four most important short-term features are selected as the research features. The four most important features are zero-cross rate (ZCR), energy entropy (EE), spectral spread (SS) and spectral entropy (SE).

Then the long-term characteristics were obtained by accumulating the short-term features of 80 Hamming windows, and the length of the long-term feature is exactly equal to 2.025 seconds.

After data pre-processing and feature selection, a total of 27506 feature matrices for 4 × 80 are obtained, with each row representing a feature and each column representing a feature of a frame. The number of non-depression categories was about three times the number of depression categories. The non-depression categories were down sampled to equalize the positive and negative samples, the depression category is labeled as positive category, and the non-depression category is marked as negative category. The final number of feature matrices for both positive and negative samples was 7313.

3 MACHINE LEARNING MODEL

3.1 Feature Discretization

Depression detection based on audio signals usually uses short-term features or long-term features of the audio signal for classification. Short-term features vary widely from frame to frame, and long-term features ignore information such as changes in features within audio segments, neither of which reflect the characteristics of the audio signal very objectively. Therefore, we use feature combination approaches to create new long-term features for classification. The method is that we count the frequencies of the low, middle and high values of the four selected short-term features (first method) and their joint occurrence (second method) in a time interval (80 frames) to reflect the changes of speech.

Feature combination will increase the number of combined features exponentially. In order to reduce the total number of features whereas generating the combined features, the features are first discretized into discrete features. Specifically, the four selected features are first discretized into low, medium and high values based on thresholds, and the thresholds used are the upper and lower third of the feature value.

Table 1 List of discrete features

Features	Threshold		Eigenvalue		
ZCR	0.08	0.20	*ZCR*[0]	*ZCR*[1]	*ZCR*[2]
EE	0.31	0.55	*EE*[0]	*EE*[1]	*EE*[2]
SS	0.06	0.16	*SS*[0]	*SS*[1]	*SS*[2]
SE	0.06	0.16	*SE*[0]	*SE*[1]	*SE*[2]

The two values in the "Threshold" column of Tab. 1 are the lower and upper third loci, and the "[0], [1], [2]" in the feature values correspond to the low, medium and high in the feature size.

3.2 Combination of Features

We use two methods of feature combination to generate long-term features. In the first method, the combined long-term feature vector V consists of a combination of four separate long-term feature vectors, with each separate long-term feature generated as follows.

First generate the short-term eigenvector $W(s)$ by using Eq. (1), as in

$$W(s)=\begin{cases}[1,0,0], a > L(s)\\ [0,1,0], b > L(s) \geq a\\ [0,0,1], b \leq L(s)\end{cases} \tag{1}$$

$W(s)$ represents the vectorized representation of the short-term features at time t, $L(s)$ represents the eigenvalues at time t, a and b represent the threshold points for the discretization of the eigenvalues.

The short-term eigenvector $W(s)$ is accumulated in the time interval Δt to generate a long-term feature vector, as shown in Eq. (2)

$$V=\int_{t}^{t+\Delta t} W(s)\,\mathrm{d}s \tag{2}$$

Δt represents the time interval of 2.025s for long-term features, and also represents the 80 frames of audio segments counted.

Four independent long-term eigenvectors (V_{ZCR}, V_{EE}, V_{SS}, and V_{SE}) are generated from Eq. (1) and Eq. (2), which represent the long-term features of the *ZCR*, *EE*, *SS* and *SE*.

In the second method, the discrete features of all short-term features are combined in a co-occurring manner to form new short-term features. Each feature corresponds to three discrete features with low, medium and high values. The combination of two features will produce a 9-dimensional (3 × 3 = 9) short-term combined feature vector, and the four features will have a dimension of 81 when combined. The resulting 9-dimensional feature vector, using the combination of the zero-crossing rate and energy entropy as an example, is shown in Tab. 2.

Table 2 Feature combination

	ZCR[0]	*ZCR*[1]	*ZCR*[2]
EE[0]	Dimension 1	Dimension 2	Dimension 3
EE[1]	Dimension 4	Dimension 5	Dimension 6
EE[2]	Dimension 7	Dimension 8	Dimension 9

The feature consists of nine dimensions, in Tab. 2, the *ZCR* is a low value, denoted by *ZCR*[0], *EE* is a low value denoted by *EE*[0]. If these two low values are co-occurring in a frame, the dimension 1 takes the value of 1, the values of the other dimensions are zero, in a similar fashion, a combined short-term feature vector is produced from Eq. (3), as in

$$WZCR*EE(S)= \begin{cases} [1,0,0,0,0,0,0,0,0], ZCR(s)<a_1, EE(s)<a_2 \\ [0,1,0,0,0,0,0,0,0], ZCR(s)<a_1, b_2>EE(s)\ge a_2 \\ [0,0,1,0,0,0,0,0,0], ZCR(s)<a_1, EE(s)\ge b_2 \\ [0,0,0,1,0,0,0,0,0], b_1>ZCR(s)\ge a_1, EE(s)<a_2 \\ [0,0,0,0,1,0,0,0,0], b_1>ZCR(s)\ge a_1, b_2>EE(s)\ge a_2 \\ [0,0,0,0,0,1,0,0,0], b_1>ZCR(s)\ge a_1, EE(s)\ge b_2 \\ [0,0,0,0,0,0,1,0,0], ZCR(s)\ge b_1, EE(s)<a_2 \\ [0,0,0,0,0,0,0,1,0], ZCR(s)\ge b_1, b_2>EE(s)\ge a_2 \\ [0,0,0,0,0,0,0,0,1], ZCR(s)\ge b_1, EE(s)<b_2 \end{cases} \quad (3)$$

ZCR(*s*) and *EE*(*s*) denote the eigenvalues of the *ZCR* and *EE* at time *t*, a_1, b_1, a_2 and b_2 are the upper and lower thresholds of the *ZCR* and *EE*, respectively. W_{ZCR*EE} is the combined eigenvectors at time *t*.

Finally, the long-term characteristics are generated by Eq. (4), as in

$$VZCR*EE=\int_t^{t+\Delta t} W_{ZCR*EE}(s)\,\mathrm{d}s \quad (4)$$

The 9-dimensional long-term feature vector combined by *ZCR* and *EE* is obtained from Eq. (4). Furthermore, the four optimal features are combined, Eq. (3) becomes 81-dimensional from 9-dimensional. The long-term feature vector $V_{ZCR*EE*SS*SE}$ of 81-dimensional combined by the four optimal features is finally obtained from Eq. (4).

3.3 Data Normalization

There are large differences in the scales of the long-term features generated by the above methods, for example, the eigenvalues of the combined features ($V_{ZCR*EE*SS*SE}$) are all integers, but there are significant differences in the order of magnitude. In order to compare the differences between different audio segments, different combined features due to the degree of depression, and to create a more stable model, the combined feature obtained is normalized without changing the original distribution of the data, but only scaling the data.

3.4 Creating Datasets

Each long-term feature has the same label and corresponding feature matrix. The choice of Δt = 2.025 s as the time window (according to the size of Hamming window and step in section 2.2, the 2.025 s audio segment exactly contains 80 frames of feature values), and the feature vectors is obtained through Eqs. (1), (2), (3) and (4). The corresponding dataset is produced, which has the same data distribution as the feature matrix in section 2.2. In order to verify whether gender is an important feature, one dimension is added to the long-term feature vector to represent gender using a sparse representation when creating the datasets. The gender dimension was set to 1 if the subject was male and 0 for female. The combining features generate a long-term feature vector with a total of 81 dimensions, and after adding the gender feature, the total dimensions number of the feature vector (V = [*gender*, $V_{ZCR*EE*SS*SE}$]) is 82.

3.5 Evaluation Criteria

The depression category is labeled as 1 and positive category, and the non-depression category is marked as 0 and negative category.Sensitivity, precision and F_1 are usually used to measure the goodness of the classification results. Sensitivity, also known as true positive rate (*TPR*), or as recall (*R*), is how sensitive the diagnostic method is to the disease (ability to recognize it). It is given by

$$Sens = R = \frac{TP}{TP+FN} \quad (5)$$

where *TP* is number of true positives, *FN* is number of false negatives. The higher the sensitivity, the lower the probability of missing the diagnosis. Precison is the ratio of the true positive cases to the positive cases obtained by classification, which can also be called the accuracy rate. It is calculated as shown in Eq. (6)

$$P = \frac{TP}{TP+FP} \quad (6)$$

Recall and precison are influenced by each other, in general, high precison and low recall, low recall and high precison. In the process of evaluating a model, it is not possible to fully evaluate a model by only using precision or recall, so F_1 score is uusally adopted as the actual scoring criterion for the model. F_1 considers both precision and recall; it is the harmonic mean of precision and recall, given by

$$F_1 = 2\frac{P\cdot R}{P+R} \quad (7)$$

When both accuracy and recall are high, the F_1 value will also be high.

4 RESULTS AND ANALYSIS

In this section, the results for long-term features, short-term features and long-term features constructed in this paper as classification features are compared with each other, and the results of different machine learning algorithms are compared, as well as with the results of studies by Baseline [10] and Huang [19]. The machine learning algorithm (RF or XGBoost) in this paper is implemented using the sklearn library (version 0.23.2) of python version 3.7. And the results of the studies by Baseline and Huang are obtained from their papers.

Tab. 3 shows the F_1 scores and the sensitivities of the combined long-term feature, the purely long-term feature and short-term feature. Because of the large variation among subjects, in order to validate the performance of the model in the new datasets, the datasets are divided into training and validation sets in a ratio of 4:1. To ensure that each subject's data can be distributed in both the training and validation sets and to increase the generalization ability of the model, a randomization method is used to divide the training and validation sets. The model was tuned for parameters in the training set using a five-fold crossover [20], and the F_1 scores and sensitivities were obtained in the validation set using the built model.

To compare the classification performance of feature combinations and individual features as well as long and short term features xperiment 1 used long-term features as classification features and Experiment 2 used short-term features as classification features. 68 extracted features were selected for classification in Experiments 1 and 2, and the classification features used in Experiments 3 is long-term feature generated from feature combination. The results of the experiments are shown in Tab. 3. (In brackets indicate negative categories, F_1 scores and sensitivities for non-depression categories).

Table 3 Comparison of different features and algorithm results

Experiment	RF Algorithm		XGBoost algorithm	
	F_1 score	Sensitivity	F_1 score	Sensitivity
1. Long-term features	0.45 (0.58)	0.39 (0.66)	0.34 (0.64)	0.24 (0.82)
2. Short-term features	0.52 (0.68)	0.46 (0.61)	0.51 (0.52)	0.50 (0.53)
3. Feature combination	0.66 (0.70)	0.75 (0.61)	0.64 (0.60)	0.67 (0.58)

From Tab. 3, it can be seen that among the two machine learning methods, Experiment 3 has a greater improvement in F1 score and sensitivity compared to Experiment 1 and Experiment 2, indicating that the feature combination method has a significant improvement in the results of depression classification compared to simple long-term and short-term features. Due to the excessive number of features selected, Experiment 1 and Experiment 2 showed severe overfitting, resulting in poorer results in the training set.

Table 4 Comparison with other studies

	F_1 score	Sensitivity
This paper	0.66 (0.70)	0.75 (0.61)
Huang	0.52 (0.70)	1.00 (0.54)
Baseline	0.46 (0.68)	0.85 (0.54)

Tab. 4 shows the comparison of this experimental method with other studies from the same datasets, and the F_1 score and sensitivities of Experiment 3 and the Random Forest algorithm were chosen to make comparisons with the other methods. In terms of sensitivity, Huang [19] achieved a recognition rate of 100 percent for the depression category, with a leakage rate of 0. This is where the deep learning method has the advantage of being able to better handle the changing situation of audio signals and extract deeper features for sequence-type data like audio, but the recognition rate for the non-depression category was only 0.54, which was 0.07 less than in this paper, and also resulted in the depression of method category F_1 score decreasing, 0.14 less than in this paper; in the baseline presented at the 2016 AVEC, the sensitivity was at 0.85, 0.10 higher than in this paper, but the depression category F_1 score was only 0.46, 0.20 less than in this paper. In the study of Huang [19], the rhythmic features of audio were not considered, and with the help of neural network models to extract advanced features with poor interpretability, the feature extraction function was based on convolution kernel and the processing function of long and short-term memory network for contextual connection of sequence data achieved surprising results in the judgement of depression category; in the baseline of 2016 AVEC, short-term features, and parameters such as mean and variance of different low-level features were used as feature values, and linear support vector algorithm with stochastic gradient descent was used as the classifier, also achieved good results. The method chosen in this paper performs feature combination for a small number of features, and the long-term features created by feature combination method, which combines high, medium and low levels of individual features in a co-occurring manner, with different levels of individual features considered in each dimension, has a shortfall in the recognition rate of the depression category, but overall has less false positives for the depression category and a significant improvement in F_1 scores.

5 CONCLUSIONS

In this paper, we used audio signals to achieve intelligent diagnosis of depression. Firstly, four short-term audio features were selected abiding by the order of importance, then two separate approaches were used to generate a long-term feature judgment vector, and a gender feature was added to the long-term judgment feature vector, followed by classification using the Random Forest algorithm, and the XGBoost classifier was applied as an algorithmic comparison. The results were compared with single long-term or short-term judgment features and other studies. Our method ensured the recognition rate of the depressed category and improved the recognition rate of the F_1 score and the non-depression category. In addition, we have ranked the importance of the different feature values of the feature vector by the way of the *RF* algorithm, all tree nodes for a feature in the *RF* were found and the average of their Gini coefficient reductions was calculated as a measure of feature importance. It was found that low value of *SE* plays a key role in the classification of depression, and the gender feature is not an important factor in the identification of depression.

However, the method in this paper also has certain flaws. For example, when performing the feature combination, only four optimal short-term features were selected, dropping the more important energy features, which represent the level of sound and also have a more important role in the classification of depression. The reason why the fifth short-term feature was not selected is that the long-term feature vector generated by the combination of four features is 81 dimensions, and if five features were selected, a 243-dimensional feature vector would be generated, which would make the model more complex and even produce over-fitting. This does not mean that only four features can be used. Secondly, the combination of features is still another way of representing

the long-term features of audio segments, which ignores the changing information of audio signals and cannot fully exploit the rich emotional information contained in audio signals. Finally, the datasets used in this paper are not in Chinese, and there is a huge difference between the pronunciation of English and Chinese, so the model is limited in identifying depressed patients in China. Future re-search will seek to model and classify the Chinese language datasets so that the model can be useful for early diagnosis of depressed patients in China.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities (Grants No. CZZ21007 and CZY19021).

6 REFERENCES

[1] Health China Action Promotion Committee. Health China Action (2019-2030). See http://www.gov.cn/xinwen/2019-07/15/ content_5409694.htm

[2] Le, Y., Jiang, D., & Hichem, S. (2018). Integrating deep and shallow models for multi-modal depression analysis - hybrid architectures. *IEEE Transactions on Affective Computing*, *12*(1), 239-253. https://doi.org/10.1109/TAFFC.2018.2870398

[3] Wang, Z., Chen, L., Wang, L., & Diao, G. (2020). Recognition of audio de-pression based on convolutional neural network and generative antagonism network model. *IEEE Access*, *8*, 101181-101191. https://doi.org/10.1109/ACCESS.2020.2998532

[4] Safa, R., Bayat, P., & Moghtader, L. (2021). Automatic detection of depression symptoms in twitter using multimodal analysis. *The Journal of Supercomputing*, *4*. https://doi.org/10.1007/s11227-021-04040-8

[5] Arevian, A. C., Bone, D., Malandrakis, N., Martinez, V. R., Wells, K. B., Miklowitz, D. J., & Narayanan, S. (2020). Clinical state tracking in serious mental illness through computational analysis of speech. *PloS One*, *15*(1), e0225695. https://doi.org/10.1371/journal.pone.0225695

[6] Schumann, I., Schneider, A., Kantert, C., Löwe, B., & Linde, K. (2012). Physicians' attitudes, diagnostic process and barriers regarding depression diagnosis in primary care: a systematic review of qualitative studies. *Family Practice*, *29*(3), 255-263. https://doi.org/10.1093/fampra/cmr092

[7] Horwitz, R., Quatieri, T., Helfer, B., Yu, B., Williamson, J. R., &Mundt, J. (2013). On the relative importance of vocal source, system, and prosody in human depression. *IEEE International Conference on Body Sensor Networks*, *Cambridge, MA, USA*, 1-6. https://doi.org/10.1109/BSN. 2013. 6575522

[8] Takaya, T., Hirokazu, T., Kiyotaka, N., Masayuki, S., Toru, N., Ryuki, T., Masafumi, N., & Tetsuaki, A. (2018). Major depressive disorder discrimination using vocal acoustic features. *Journal of Affective Disorders*, *225*, 214-220. https://doi.org/10.1016/j.jad.2017.08.038

[9] Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investig Otolaryngol*, *5*(1), 96-116. https://doi.org/10.31219/osf.io/5pwze

[10] Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres, M. T., Scherer, S., Stratou, G., Cowie, R., & Pantic, M. (2016). AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 3-10. https://doi.org/10.1145/2988257.2988258

[11] Li, J. & Fu, X. (2019). Audio Depression Recognition Based on Deep Learning. *Computer Applications and Software*, *36*(9), 161-167. https://doi.org/10.3969/j.issn.1000-386x.2019.09.029

[12] Cao, X., Li, H., & Wang, W. (2019). A study of gender differences in corpus-based audio emotion recognition. *Nanjing University*, *55*(5), 758-764. https://doi.org/10.13232/j.cnki.jnju.2019.05.007

[13] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1),5-32. https://doi.org/10.1023/A:1010933404324

[14] See https://github.com/dmlc/xgboost

[15] Gratch, J., Arstein, R., Lucas, G., & Stratou, G. (2014). The distress analysis interview corpus of human and computer interviews. *The Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 3123-3128.

[16] Dhingra, S. S., Kroenke, K., Zack, M. M., Strine, T. W., & Balluz, L. S. (2011). PHQ-8 Days: a measurement option for DSM-5 Major Depressive Disorder (MDD) severity. *Population Health Metrics*, *9*, 11. https://doi.org/10.1186/1478-7954-9-11

[17] Giannakopoulos, T. (2015), pyAudioAnalysis: An open-source python library for audio signal analysis. *PloS one*, *10*(12),1-17. https://doi.org/10.1371/journal.pone.0144610

[18] Gu, X., Ni, T., & Wang, H. (2014). New Fuzzy Support Vector Machine for the Class Imbalance Problem in Medical Datasets Classification. *The Scientific World Journal*, 536434. https://doi.org/10.1155/2014/536434

[19] Ma, X., Yang, H., Chen Q., Huang, D., & Wang, Y. (2016) DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, *Amsterdam, The Netherlands*, 35-42. https://doi.org/10.1145/2988257.2988267

[20] Cawley, G. C. & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079-2107. https://doi.org/10.1007/s10846-010-9395-x

Contact information:

Yuan GAO, Assistant Professor
South-Central Minzu University, School of Biomedical Engineering,
Wuhan 430074, China
E-mail: gaoyuan@scuec.edu.cn

Yinan XIN, MSc
South-Central Minzu University, School of Biomedical Engineering,
Wuhan 430074, China
E-mail: 2019110468@scuec.edu.cn

Li ZHANG, Professor
(Corresponding author)
South-Central Minzu University, School of Biomedical Engineering,
Wuhan 430074, China
E-mail: zhangli@scuec.edu.cn