GFIL: A Unified Framework for the Importance Analysis of Features, Frequency Bands, and Channels in EEG-Based Emotion Recognition

Yong Peng, *Member, IEEE*, Feiwei Qin, Wanzeng Kong, *Member, IEEE*, Yuan Ge, Feiping Nie, *Senior Member, IEEE*, and Andrzej Cichocki, *Fellow, IEEE*

***Abstract*—Accurately and automatically recognizing the emotional states of human beings is the central task in affective computing. The electroencephalography (EEG) data, generated from the neural activities in brain cortex, provide us with a reliable data source to perform emotion recognition. Besides the recognition accuracy, it is also necessary to explore the importance of different EEG features, frequency bands, and channels in emotion expression. In this article, we propose a unified framework termed graph-regularized least square regression with feature importance learning (GFIL) to simultaneously achieve these goals by incorporating an autoweighting variable into the least square regression. Unlike the widely used trial-and-error manner, GFIL automatically completes the identification once it is trained. Specifically, GFIL can: 1) adaptively discriminate the contributions of different feature dimensions; 2) automatically identify the critical frequency bands and channels; and 3) quantitatively rank and select the features by the learned autoweighting variable. From the experimental results on the SEED_IV data set, we find GFIL obtained improved accuracies based on the feature autoweighting strategy, which are 75.33%, 75.03%, and 79.17% corresponding to the three cross-session recognition tasks (session1->session2, session1->session3, session2->session3), respectively. Additionally, the *Gamma* band is identified as the most important one and the channels locating in the prefrontal and left/right central regions are more important.**

***Index Terms*—Affective brain–computer interface (aBCI), channel, electroencephalography (EEG), emotion recognition, feature importance learning, frequency band.**

Manuscript received 4 December 2020; revised 19 April 2021; accepted 17 May 2021. Date of publication 21 May 2021; date of current version 9 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61971173, Grant 61972121, Grant U1909202, and Grant U20B2074; in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY21F030005; in part by the National Social Science Foundation of China under Grant 19ZDA348; in part by the Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant GK209907299001-008; in part by the China Postdoctoral Science Foundation under Grant 2017M620470; in part by the Key Laboratory of Advanced Perception and Intelligent Control of High-End Equipment of Ministry of Education (Anhui Polytechnic University) under Grant GDSC202015; and in part by the Jiangsu Provincial Key Laboratory for Computer Information Processing Technology (Soochow University) under Grant KJS1841. *(Corresponding author: Yong Peng.)*

Yong Peng and Wanzeng Kong are with the School of Computer Science and Technology and the Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: yongpeng@hdu.edu.cn).

Feiwei Qin is with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China.

Yuan Ge is with the Key Laboratory of Advanced Perception and Intelligent Control of High-End Equipment of Ministry of Education, Anhui Polytechnic University, Wuhu 241000, China.

Feiping Nie is with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China.

Andrzej Cichocki is with the Center for Computational and Data-Intensive Science and Engineering, Skolkov Institute of Science and Technology, 143026 Moscow, Russia, and also with the School of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: a.cichocki@riken.jp).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCDS.2021.3082803.

Digital Object Identifier 10.1109/TCDS.2021.3082803

I. Introduction

ACCURATELY recognizing the emotional states of our human beings is of great significance to achieve harmonic human-to-human interaction. For human–machine interaction, this is also an inevitable step to achieve emotional intelligence, which has been considered as a complementary but more advanced aspect to logical intelligence. Depending on different data sources in use, existing studies in the field of emotion recognition can be roughly categorized into four groups, including vision-based, text-based, speech-based, and the biosignal-based methods.

Vision-based emotion recognition aims to identify the emotional states of facial expressions from images or video frames. It is a typical pattern recognition process, usually including the stages of face detection and localization, feature extraction and tracking on facial landmark points, machine learning model design, and training among which the facial expression related feature extraction and recognition are more important to determine the recognition performance. Text-based emotion recognition methods use the technologies of natural language processing, text mining, and computing linguistics to extract and recognize the subjective emotional representation in the text. Speech emotion recognition refers to the characterization of speech signals or corpora of different tonal emotion states, which are used to train the machine learning model and obtain the corresponding emotional states of different voices. Generally, these three different types of emotion recognition methods can be collectively referred to as nonphysiological signal-based emotion recognition. Their common advantage is that they are relatively easy to implement; however, they cannot guarantee recognition reliability. For example, people can hide their true emotional states by disguising their facial expressions, voices and intonations. What is more, for people

with physical disabilities, it is difficult to perform emotion recognition from their nonphysiological signals.

Compared to the nonphysiological signals, physiological signals have significant advantages, such as being difficult to disguise, being more objective, and providing richer information. Earlier research in this field focused on the peripheral physiological signals, such as skin impedance, respiration, and blood volume pulsation; however, the change rate of these signals is low and, therefore, the time resolution is limited. With recent advances in signal acquisition and wearable computing devices, electroencephalography (EEG) data have become an important data source for emotion recognition because it is generated by the central nervous system activities and directly correlated to the production of emotion. Therefore, EEG has been considered to be the "gold standard" for emotion recognition.

EEG is typically multichannel and multirhythm. In most of the existing studies, features extracted from different EEG-frequency bands were directly concatenated to form the sample vector. These methods essentially assume that each dimension of the feature vector has an equal contribution to emotion recognition; however, it is not the case from the machine learning perspective. Therefore, it is necessary to differentiate the weights of different dimensions of the feature vector (e.g., sample).

As a consequence, in this article, we propose a new computational model termed graph-regularized least square regression with feature importance learning (GFIL) for joint feature importance learning and emotion recognition. Compared to the existing studies, our article has the following contributions.

1) By introducing an autoweighting variable, the proposed GFIL model can adaptively learn the contributions of different EEG feature dimensions in emotion recognition, which serves as a unified framework for joint feature importance learning and emotion recognition.
2) Due to the coupling relationship between the different feature dimensions and the EEG-frequency bands (channels), GFIL can automatically and quantitatively measure the importance of different EEG-frequency bands and channels for emotion recognition.
3) Once the GFIL model is optimized, the obtained feature autoweighting variable provides us with an effective tool to rank the EEG features and then we can select more important features for emotion recognition.
4) Extensive experiments are conducted on the SEED_IV data set and the results well support the above declarations from three perspectives: a) recognition performance is improved by introducing the feature weighting strategy; b) the critical EEG-frequency bands and channels for emotion recognition are identified; and c) feature ranking and selection are achieved when the model training is finished.

The remainder of this article is organized as follows. Section II briefly reviews some recent advances in EEG-based emotion recognition. In Section III, we introduce the model formulation and optimization of GFIL in detail. The EEG data preparation, experimental settings, and results are presented in Section IV. Besides, how GFIL works in critical EEG-frequency bands and channel identification is provided in this section. Section V concludes the whole article.

Notations: We summarize the notations used in this article. Matrices are written in boldface uppercase letters. Vectors are written as boldface lowercase letters. For example, the (i, j)th element of matrix $\mathbf{M}$ is m_{ij}. The squared ℓ_2-norm of matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$ is $\|\mathbf{M}\|_2^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} m_{ij}^2$. By default, we use $\mathbf{m}_i$ to represent the ith column vector of $\mathbf{M}$ and $\mathbf{m}^j$ to represent its jth row vector. Moreover, the five frequency bands widely used in EEG-based BCIs are denoted as *Delta*, *Theta*, *Alpha*, *Beta*, and *Gamma* to avoid confusion with model variables (parameters).

II. Related Works

From the perspective of pattern recognition, the main efforts on EEG-based emotion recognition were made from two aspects: 1) *feature extraction* and 2) *recognition model design*. Below, we present a brief review of existing studies from these two perspectives. Moreover, some related studies on feature weighting and critical EEG-frequency bands identification are included.

A. Feature Extraction

Roughly, we can extract the EEG features for emotion recognition from the time domain, frequency domain, and time–frequency domain. In addition, functional connectivity features are also useful in providing spatial information.

The time-domain EEG features are mainly to identify the characteristics of time series that vary between different emotional states. Some representative ones are the event related potentials, statistical features, nonstationary index, fractal dimension, and high-order crossings (HOCs).

Compared to the time-domain features, the EEG features extracted from the frequency domain are more stable. Power features from different frequency bands are the most popular ones in the field of EEG-based emotion recognition.

Usually, the fast fourier transform is used to implement the signal transformation from the time domain to frequency domain. Commonly used alternatives are the short-time Fourier transform or Welch's method to estimate the power spectra density (PSD). Sometimes, higher order spectra (HOS) features can be built by the group of frequency-domain features. It is worth mentioning that as a variant of the power spectrum, the differential entropy (DE) feature showed excellent performance in diverse EEG-based brain–computer interfaces [1]–[4], which is also employed in this work.

Essentially, EEG data are nonstationary and, therefore, time–frequency domain features are more appropriate in capturing the dynamics of frequencies in terms of time. The discrete-wavelet transform (DWT) is the most widely used method to decompose EEG signal into an approximation and some detail levels corresponding to different frequency ranges. Similarly, the empirical mode decomposition (EMD, also known as the Hilbert–Huang Transform) can decompose the EEG signal into several intrinsic mode functions (IMFs) along with a residue [5], [6].

Apart from the multirhythm property, the EEG signal is also multichannel; therefore, the functional connectivity features can be calculated from combinations of electrodes. For example, the differential asymmetry feature calculates the differences in power bands of corresponding pairs of electrodes; similarly, the rational asymmetry calculates the ratios of features from symmetric electrodes. Phase synchronization of EEG signals is a reflection of the difference in the physiological structure of the white matter in the brain and the phase-locking value (PLV) was widely used to quantify the interaction of EEG signals recorded from different channels [7], [8].

A thorough review of feature extraction and selection methods in EEG-based emotion recognition was conducted by Jenke *et al.* [9]. Recently, with powerful representation learning ability, deep neural networks can directly take the raw EEG data rather than handcrafted features as input [10]–[12] and then generate the recognition results, which is different from traditional machine learning models of performing feature extraction and recognition in separated stages.

B. Recognition Model

In [13], mutual information-based feature selection methods and kernel classifiers were combined to improve the accuracy of EEG-based emotion classification. Zheng [14] proposed a group sparse canonical correlation analysis (GSCCA) method for simultaneous EEG channel selection and emotion recognition, which can effectively model the linear correlation between emotional labels and EEG feature vectors. In [15], a PLV-based emotional brain network was constructed and a multiple feature fusion approach was adopted to combine the complementary activation and connection information for emotion recognition. The nonstationary property of EEG can easily lead to the feature distribution discrepancy among subjects and transfer learning methods were employed to suppress the data misalignment between the source domain and the target domain. In [16] and [17], some domain adaptation techniques for EEG-based emotion recognition were reviewed.

With powerful nonlinear feature learning capability, deep neural networks were widely applied in EEG-based brain-computer interfaces. The deep belief network (DBN) was trained by DE feature to perform binary emotional states classification [18]. Li *et al.* [19] proposed a hybrid deep learning model by combining the convolutional neural network (CNN) and the recurrent neural network (RNN) for extracting the task-related features, mining interchannel correlation and incorporating contextual information. The dynamic graph CNNs (DGCNNs) were proposed for EEG-based emotion recognition [20]. In DGCNN, the graph structure was used to model the multichannel EEG features and the intrinsic relationship between different EEG channels can be dynamically learned in the network training process. To instantaneously detect the video viewers' emotions from EEG signals and facial expressions, a combinational model based on long short-term memory RNNs (LSTM-RNN) and continuous conditional random fields (CCRFs) was proposed to detect emotions automatically and continuously [21]. Two review studies on applying deep learning models into EEG data processing can be found in [22] and [23].

C. Feature Weighting, Channel Selection and Frequency Bands Identification

Generally, there has been little research on EEG feature weighting, critical frequency bands, and channels identification for emotion recognition. Most of the existing studies simply concatenated the EEG features extracted from different frequency bands to form a feature vector (e.g., sample) which would be used as input to machine learning models. For critical frequency bands identification, the "trial-and-error" based methods were widely used; specifically, if the feature vector from a certain frequency band obtained the best recognition accuracy, it can be viewed as the critical frequency band. Since adjacent EEG channels have similar potentials in characterizing brain activities, some rule-based methods were proposed to remove such redundancy to achieve channel selection.

Zhang *et al.* [24] proposed a ReliefF-based channel selection method for EEG-based emotion recognition, whose core idea is to evaluate the quality of features according to their abilities in differentiating samples within a neighborhood. The lagged Poincare's measures of EEG recordings and a source location method (sLORETA) were used to select more important EEG electrodes for emotion recognition [25]. The classification results of the multilayer perceptron neural network (MLPNN) and k-Nearest Neighbor (kNN) were used to determine the top five EEG channels based on which the final feature vectors were obtained to retrain classifiers [26]. Wang *et al.* [27] proposed to select an optimal subset of EEG channels according to the normalized mutual information (NMI) criterion. Paper [28] reviewed the recent methods in the EEG channel selection field along with their applications. In [29], by using the correlation coefficient feature selection method on the DE feature, the top 20 selected features are mostly from the *Beta* and *Gamma* bands and at the lateral temporal and frontal brain areas. To the best of our knowledge, [30] is the only existing study by using learning-based method (the weight distribution of the trained DBNs) to perform important frequency bands and channel identification on EEG-based emotion recognition. From the experimental results, the *Gamma* band is the most important one and the channels in lateral temporal and prefrontal brain areas are more important. However, there are several limitations in the deep learning models, such as their high computational complexities and limited interpretability.

III. Method

In this section, we present the GFIL model, including its formulation and optimization.

A. Model Formulation

Due to the multirhythm and multichannel properties of EEG signals, features extracted from different frequency bands and different channels should contribute differently in emotion recognition. However, such differences did not attract sufficient attention in EEG-based emotion recognition; that is, most existing studies directly concatenated all these features

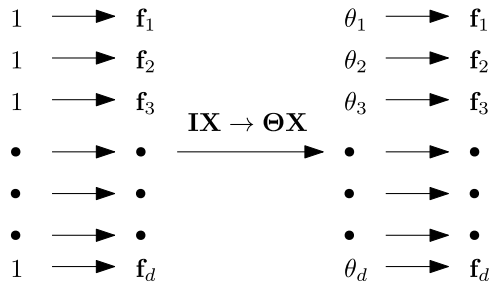


Fig. 1. Graphical show of the feature importance learning.

together to form a complete feature vector that serves as the input of learning models. This potentially assumed that all feature dimensions shared the same importance, as shown in the left part of Fig. 1. In the current research, we introduce an autoweighting variable $\boldsymbol{\theta} \in \mathbb{R}^d$ to perform adaptive feature importance learning. Specifically, the values $\theta_i|_{i=1}^d$ quantitatively characterize the weights of features $\mathbf{f}_i|_{i=1}^d$, as shown in the right part of Fig. 1. Therefore, the general idea is that we can incorporate $\boldsymbol{\theta}$ into learning models to explore the importance of different EEG feature dimensions. Then, $\boldsymbol{\theta}$ becomes an variable of a certain learning model and can be adaptively fitted by EEG data.

In the supervised setting of EEG-based emotion recognition, given n training samples $(\mathbf{x}_i, y_i)|_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the EEG feature vector and y_i is the corresponding emotional label, we adopt the least square regression (LSR) model in this article due to its simplicity and effectiveness. By enforcing $\boldsymbol{\theta}$ to satisfy $\boldsymbol{\theta} \geq \mathbf{0}$ and $\boldsymbol{\theta}^T\mathbf{1} = 1$, we incorporate it into the LSR model and then obtain the objective function of GFIL as

$$\begin{aligned} \min_{\boldsymbol{\theta},\mathbf{W},\mathbf{b}} \ & \left\|\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y}\right\|_F^2 + \alpha\|\mathbf{W}\|_F^2 + \beta\mathcal{R}(\boldsymbol{\theta}, \mathbf{W}) \\ \text{s.t.} \quad & \boldsymbol{\theta} = \text{diag}(\boldsymbol{\Theta}), \boldsymbol{\theta} \geq \mathbf{0}, \boldsymbol{\theta}^T\mathbf{1} = 1. \end{aligned} \tag{1}$$

Here, $\boldsymbol{\Theta} \in \mathbb{R}^{d\times d}$ is a diagonal matrix, whose diagonal elements are corresponding to those in $\boldsymbol{\theta}$. $\mathcal{R}(\boldsymbol{\theta}, \mathbf{W})$ is termed the regularization term to constrain the model variables $\boldsymbol{\theta}$ and $\mathbf{W}$. There are many criteria to define such a term and in this article we follow the label consistency property [31]. Its basic idea can be stated as follows, if two EEG samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are belonging to the same emotional state, they should have similar representations in projected space obtained by $\mathbf{W}$ and $\boldsymbol{\Theta}$. Therefore, a reasonable expression of $\mathcal{R}(\boldsymbol{\theta}, \mathbf{W})$ to satisfy such requirement can be defined as

$$\mathcal{R}(\boldsymbol{\theta}, \mathbf{W}) \triangleq \sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\mathbf{W}^T\boldsymbol{\Theta}\mathbf{x}_i - \mathbf{W}^T\boldsymbol{\Theta}\mathbf{x}_j\right\|_2^2 s_{ij}. \tag{2}$$

In (2), matrix $\mathbf{S} \in \mathbb{B}^{n\times n} = \{0, 1\}^{n\times n}$ is a binary graph affinity matrix in which s_{ij} depicts the connection between sample $\mathbf{x}_i$ and sample $\mathbf{x}_j$. According to the label information of training samples, matrix $\mathbf{S}$ can be constructed as

$$s_{ij} = \begin{cases} 1, & \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ have the same emotional state} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Based on the above-defined $\mathbf{S}$, if EEG samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are belonging to the same emotional state (e.g., $s_{ij} = 1$), $\mathbf{W}^T\boldsymbol{\Theta}\mathbf{x}_i$ will be enforced to be similar to $\mathbf{W}^T\boldsymbol{\Theta}\mathbf{x}_j$ by minimizing (2). In machine learning, such defined $\mathcal{R}(\boldsymbol{\theta}, \mathbf{W})$ is called as a graph regularizer, which is widely used to achieve the local invariance of data. To facilitate the subsequent optimization, we rewrite the graph regularizer as the following matrix form:

$$\mathcal{R}(\boldsymbol{\Theta}, \mathbf{W}) \triangleq \text{Tr}\left(\mathbf{W}^T\boldsymbol{\Theta}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W}\right). \tag{4}$$

As a result, we name the proposed model as GFIL and its objective is represented by (1).

B. Model Optimization

There are three variables, $\boldsymbol{\theta}$, $\mathbf{W}$, and $\mathbf{b}$, in the objective function of GFIL. We propose to optimize (1) by the alternating direction method; that is, we update one variable with the others fixed. Below are the detailed derivations of their updating rules.

1) *Update* $\mathbf{b}$ *With* $\mathbf{W}$ *and* $\boldsymbol{\theta}$ *Fixed:* We can rewrite the objective function (1) as

$$\min_{\mathbf{b}}\left\|\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y}\right\|_2^2. \tag{5}$$

Taking the derivative of (5) to $\mathbf{b}$, and setting it to 0, we have

$$\left(\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y}\right)^T\mathbf{1} = \mathbf{0}. \tag{6}$$

Then, we get the updating rule to $\mathbf{b}$ as

$$\mathbf{b} = \frac{1}{n}\left(\mathbf{Y}^T - \mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}\right)\mathbf{1}. \tag{7}$$

2) *Update* $\mathbf{W}$ *With* $\mathbf{b}$ *and* $\boldsymbol{\theta}$ *Fixed:* By taking the derivative of objective (1) for $\mathbf{W}$ and setting it to 0, we have

$$\boldsymbol{\Theta}\mathbf{X}\left(\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y}\right) + \alpha\mathbf{W} + \beta\boldsymbol{\Theta}\mathbf{X}\mathbf{L}\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W} = \mathbf{0}. \tag{8}$$

Therefore, the updating rule to $\mathbf{W}$ can be expressed as

$$\mathbf{W} = \left(\boldsymbol{\Theta}\mathbf{X}(\mathbf{I} + \beta\mathbf{L})\mathbf{X}^T\boldsymbol{\Theta} + \alpha\mathbf{I}\right)^{-1}\left(\boldsymbol{\Theta}\mathbf{X}\left(\mathbf{Y} - \mathbf{1}\mathbf{b}^T\right)\right). \tag{9}$$

3) *Update* $\boldsymbol{\theta}$ *With* $\mathbf{W}$ *and* $\mathbf{b}$ *Fixed:* The objective function in terms of $\boldsymbol{\theta}$ is

$$\min_{\boldsymbol{\theta}\geq\mathbf{0},\boldsymbol{\theta}^T\mathbf{1}=1}\left\|\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W} + \mathbf{1}\mathbf{b}^T - \mathbf{Y}\right\|_2^2 + \beta\text{Tr}\left(\mathbf{W}^T\boldsymbol{\Theta}\mathbf{X}\mathbf{L}\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W}\right)$$

which is equivalent to

$$\begin{aligned} \min_{\boldsymbol{\theta}\geq\mathbf{0},\boldsymbol{\theta}^T\mathbf{1}=1} \ & \text{Tr}\left(\boldsymbol{\Theta}\mathbf{X}(\mathbf{I} + \beta\mathbf{L})\mathbf{X}^T\boldsymbol{\Theta}\mathbf{W}\mathbf{W}^T\right) \\ & + 2\text{Tr}\left(\mathbf{W}\left(\mathbf{b}\mathbf{1}^T - \mathbf{Y}^T\right)\mathbf{X}^T\boldsymbol{\Theta}\right). \end{aligned} \tag{10}$$

We can see that $\mathbf{W}\mathbf{W}^T$ appears on the right-hand side of $\boldsymbol{\Theta}$ and, therefore, it is not easy to directly calculate the derivative of (10) concerning $\boldsymbol{\Theta}$. Since $\boldsymbol{\Theta}$ is a diagonal matrix, we have the following lemma.

Lemma 1: If $\boldsymbol{\Theta}$ is a diagonal matrix, then $\text{Tr}(\boldsymbol{\Theta}\mathbf{P}\boldsymbol{\Theta}\mathbf{Q}) = \boldsymbol{\theta}^T(\mathbf{P} \circ \mathbf{Q}^T)\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a column vector filled with the diagonal elements of $\boldsymbol{\Theta}$ and $\circ$ denotes the dot product.

Proof: By using $\text{diag}(\mathbf{P}\boldsymbol{\Theta}\mathbf{Q})$ to represent the column vector formed by the diagonal elements of $\mathbf{P}\boldsymbol{\Theta}\mathbf{Q}$, we have

$$\text{Tr}(\boldsymbol{\Theta}\mathbf{P}\boldsymbol{\Theta}\mathbf{Q}) = \boldsymbol{\theta}^T\text{diag}(\mathbf{P}\boldsymbol{\Theta}\mathbf{Q}). \tag{11}$$

Algorithm 1 GFIL

Input: EEG feature matrix $\mathbf{X} \in \mathbb{R}^{d\times n}$ and the corresponding label matrix $\mathbf{Y} \in \mathbb{R}^{n\times c}$;

Output: The label information on unlabeled samples $\mathbf{Y}_U$ and the output weight matrix $\boldsymbol{\beta}$.

1: Initialize $\boldsymbol{\theta} = [1, 1, \ldots, 1]./d \in \mathbb{R}^d$ and $\mathbf{W}$ by least square regression (regularization parameter 0.1);
2: **while** not converged **do**
3: Update $\mathbf{b}$ according to (7);
4: Update $\mathbf{W}$ according to (9);
5: Update $\boldsymbol{\theta}$ by optimizing (14) according to Algorithm 3;
6: **end while**

If $\mathbf{p}_i$ and $\mathbf{q}_i$ $(1 \le i \le d)$, respectively, represent the ith row and the ith column of $\mathbf{P}$ and $\mathbf{Q}$, we have

$$\boldsymbol{\theta}^T \mathrm{diag}(\mathbf{P}\boldsymbol{\Theta}\mathbf{Q}) = \boldsymbol{\theta}^T \begin{bmatrix} \mathbf{p}^1\boldsymbol{\Theta}\mathbf{q}_1 \\ \mathbf{p}^2\boldsymbol{\Theta}\mathbf{q}_2 \\ \cdots \\ \mathbf{p}^d\boldsymbol{\Theta}\mathbf{q}_d \end{bmatrix} = \boldsymbol{\theta}^T \begin{bmatrix} (\mathbf{p}^1 \circ \mathbf{q}_1)^T\boldsymbol{\theta} \\ (\mathbf{p}^2 \circ \mathbf{q}_2)^T\boldsymbol{\theta} \\ \cdots \\ (\mathbf{p}^d \circ \mathbf{q}_d)^T\boldsymbol{\theta} \end{bmatrix} = \boldsymbol{\theta}^T(\mathbf{P} \circ \mathbf{Q}^T)\boldsymbol{\theta} \quad (12)$$

where we assume that $\mathbf{p}^i \circ \mathbf{q}_i|_{i=1}^d$ is a column vector formed by the elementwise product of vectors $\mathbf{p}^i$ and $\mathbf{q}_i$. ■

Based on Lemma 1, by denoting that $\mathbf{P} \triangleq \mathbf{X}(\mathbf{I} + \beta\mathbf{L})\mathbf{X}^T$, $\mathbf{Q} \triangleq \mathbf{W}\mathbf{W}^T$, and $\mathbf{R} \triangleq \mathbf{W}(\mathbf{b}\mathbf{1}^T - \mathbf{Y}^T)\mathbf{X}^T$, we can rewrite the objective (10) as

$$\min_{\boldsymbol{\theta}\ge\mathbf{0},\boldsymbol{\theta}^T\mathbf{1}=1} \boldsymbol{\theta}^T(\mathbf{P} \circ \mathbf{Q}^T)\boldsymbol{\theta} + \boldsymbol{\theta}^T\mathrm{diag}(\mathbf{R}) \quad (13)$$

where diag($\mathbf{R}$) a column vector corresponding to the diagonal of matrix $\mathbf{R}$. Problem (13) is equivalent to the following problem, which essentially is a quadratic programming with simplex constraint:

$$\min_{\boldsymbol{\theta}} \boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{c}, \ \text{s.t.}\ \boldsymbol{\theta}^T\mathbf{1} = 1, \boldsymbol{\theta} \ge \mathbf{0}. \quad (14)$$

The optimization to (14) can be found in the Appendix.

Based on the above analysis, the whole procedure of GFIL is summarized in Algorithm 1. Once it is optimized, we can obtain the quantitative feature importance measure $\boldsymbol{\theta}$ and then the feature ranking and selection can be completed. In the next section, we will explain how GFIL works on critical frequency bands and channel identification in EEG-based emotion recognition by spectra features.

IV. Experimental Studies

In this section, we first describe the EEG data set used in our experiments based on which we explain how to apply GFIL into identifying the critical EEG-frequency bands and channels. Then, we present the results ana analysis on the emotion recognition accuracies, the identification of critical frequency bands and channels, feature ranking, and selection.

A. Data Preparation

In the following experiments, we use the well-known "SEED_IV" emotional data set provided by the center for brain-like computing and machine intelligence, Shanghai Jiao Tong University (SJTU) [32]. This is a video-evoked EEG data set and total 72 video clips were carefully chosen from four different types corresponding to the four emotional states: 1) *sad*; 2) *fear*; 3) *happy*; and 4) *neutral*. Sample video clips to these four emotional states are shown in Fig. 2(a), which are collected from the "SEED" Web site.[1] Fig. 2(b) is a general system flow of video-evoked affective brain–computer interface (aBCI).

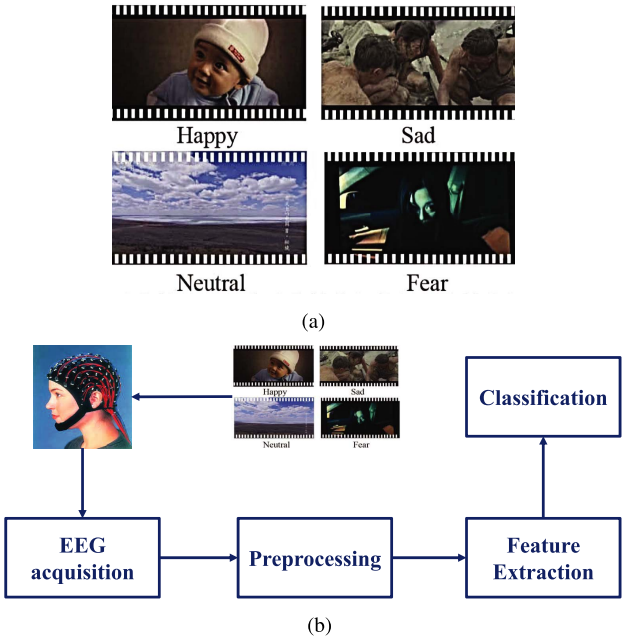


Fig. 2. Sample video clips and a general aBCI system. (a) Sample video clips. (b) General aBCI system.

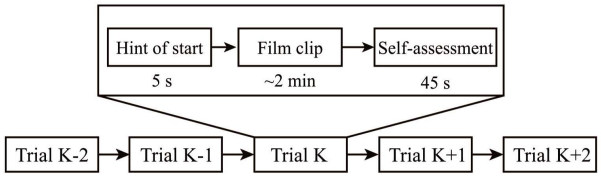


Fig. 3. Schedule diagram of trials taken from the "SEED_IV" data set Web site.

There are 15 healthy subjects who participated in the EEG data collection experiment. The experiment on each subject was divided into three sessions conducted on different days and therefore the EEG data set consists of a total of 45 sessions, three sessions for each subject. In each session, there are 24 video clips serving as the stimuli, among which exact six video clips correspond to each of the four emotional states. During watching the video clips, EEG data were recorded from subjects by the ESI NeuroScan system with a 62-channel electrode cap according to the international 10-20 system placement. The sampling rate is 1000 Hz. Fig. 3 shows the schedule diagram of trials and each trial has three stages, 5-s hint of start, video playing (about 2 min), and 45-s self-assessment.

[1] http://bcmi.sjtu.edu.cn/home/seed/seed-iv.html

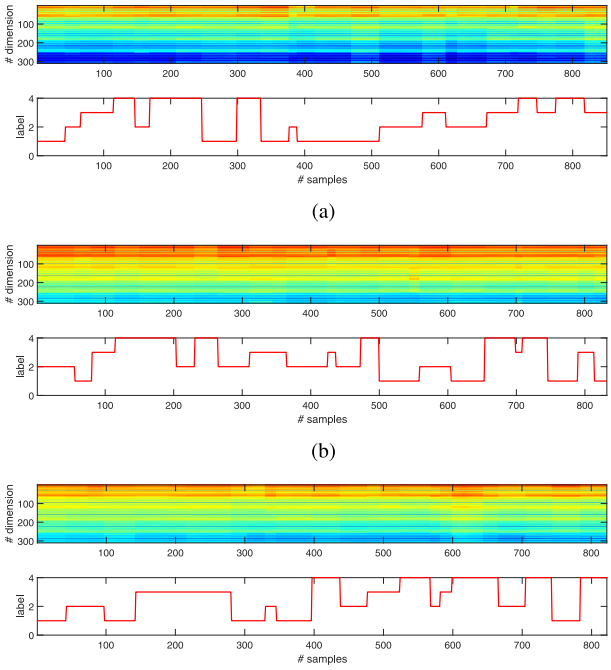


Fig. 4. EEG data collected in different sessions of one subject. The labels 1, 2, 3, and 4, respectively, represent the four emotional states, "sad," "fear," "happy," and "neutral." (a) EEG data (1st row) and label (2nd row) of subject 1 in session1. (b) EEG data (1st row) and label (2nd row) of subject 1 in session 2. (c) EEG data (1st row) and label (2nd row) of subject 1 in session 3.

B. Data Preprocessing and Feature Extraction

After downsampling the EEG data to 200 Hz, the DE feature was extracted from the five frequency bands: 1) *Delta* (1–4 Hz); 2) *Theta* (4–8 Hz); 3) *Alpha* (8–14 Hz); 4) *Beta* (14–31 Hz); and 5) *Gamma* (31–50 Hz), using the short-term Fourier transforms with a 4-s time window without overlapping. The DE is defined as

$$h(X) = -\int p(x)\ln p(x)\mathrm{d}x \tag{15}$$

where X is a random variable with probability density function (p.d.f.) $p(x)$ [1], [2]. For EEG time series, assuming that the p.d.f. $p(x)$ follows the Gaussian distribution $\mathcal{N}(x;\mu,\sigma^2)$, its DE is

$$\begin{aligned} h(X) &= -\int p(x)\left(-\frac{1}{2}\ln\left(2\pi\sigma^2\right)-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathrm{d}x \\ &= \frac{1}{2}\ln\left(2\pi\sigma^2\right)+\frac{\mathrm{Var}(X)}{2\sigma^2} = \frac{1}{2}\ln\left(2\pi e\sigma^2\right). \end{aligned} \tag{16}$$

In the above equation, we used the facts that $\int p(x)\mathrm{d}x = 1$ and $\mathrm{Var}(x) = \int p(x)(x-\mu)^2\mathrm{d}x = \sigma^2$. Taking subject 1 as an example, we show its DE feature-based samples in the first row of each subfigure of Fig. 4. Each EEG sample has 310 feature dimensions formed by concatenating the 62 points of each of the five frequency bands. Since the time durations of different video clips in each session are slightly different, we have approximate 830 samples in each session.

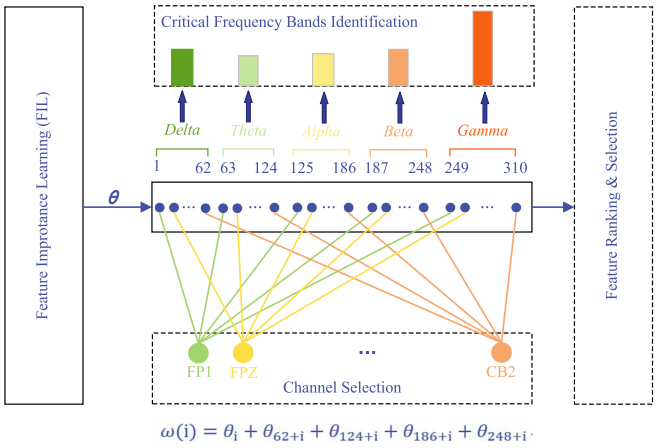


Fig. 5. Graphical show of GFIL-based frequency bands identification and channel selection.

C. GFIL-Based Critical Frequency Bands and Channels Identification

Since EEG data are multirhythm and multichannel, the feature vector can be formed by concatenating the spectra features (e.g., PSD, DE) extracted from different frequency bands (channels). Taking the "SEED_IV" data set for example, we can easily establish the correspondence between the learned feature importance measure $\boldsymbol{\theta}$ and different frequency bands (channels), as shown in Fig. 5.

Obviously, we can quantitatively calculate the importance of the ith frequency band as

$$\omega(i) = \theta_{(i-1)*62+1} + \theta_{(i-1)*62+2} + \cdots + \theta_{(i-1)*62+62} \tag{17}$$

where $i = 1, 2, 3, 4, 5$, respectively, represents the five frequency bands of *Delta*, *Theta*, *Alpha*, *Beta* and *Gamma*. Similarly, the weight of the jth channel can be obtained by

$$\psi(j) = \theta_j + \theta_{j+62} + \theta_{j+124} + \theta_{j+186} + \theta_{j+248} \tag{18}$$

where $j = 1, 2, \ldots, 62$, respectively, represents the 62 EEG channels of *FP1*, *FPZ*, ..., *CB2*. The order of the 62 channels can be found in the description file of the "SEED_IV" data set.

It is worth mentioning that GFIL is a general framework, which is flexible for any spectra features. Moreover, if there exists correspondence between the feature vector and the frequency bands (channels), GFIL is appropriate for identifying critical frequency bands and channels. In this article, we only consider the publicly available "SEED_IV" data set as an example due to the limited space.

D. Experimental Settings and Results

Due to the nonstationary nature, even if the EEG data were collected from the same subject, there might be distribution discrepancies for data collected from different sessions. To investigate whether there exist stable neural patterns in cross-session emotion recognition, we arranged the samples from one session as training samples and the ones from the other session as test samples. Specifically, for each subject, we conducted experiments by setting samples from session 1 as

TABLE I
CROSS SESSION (SESSION 1 -> SESSION 2) EMOTION RECOGNITION RESULTS (%) OF THE FIVE MODELS ON "SEED_IV" DATA SET

	session1->session2				
	1NN	LSR	SVM	FIL	GFIL
subject 1	34.13	42.55	37.50	45.31	**66.59**
subject 2	72.36	87.74	83.89	85.46	**85.70**
subject 3	38.94	61.54	51.44	63.10	**67.79**
subject 4	32.45	72.60	54.33	77.04	**84.98**
subject 5	32.45	61.78	52.88	65.99	**75.12**
subject 6	21.75	66.59	44.23	64.54	**73.32**
subject 7	40.63	75.36	71.63	75.60	**78.97**
subject 8	51.92	67.67	65.38	72.72	**73.20**
subject 9	50.24	74.28	77.28	75.60	**79.33**
subject 10	33.89	56.73	41.59	55.65	**59.25**
subject 11	47.12	59.01	48.80	66.71	**69.59**
subject 12	26.08	66.47	43.63	66.59	**70.07**
subject 13	47.48	68.27	57.21	68.03	**70.67**
subject 14	59.38	72.72	78.61	81.49	**82.21**
subject 15	77.64	91.11	88.82	91.47	**93.15**
average	44.43	68.29	59.81	70.35	**75.33**

TABLE II
CROSS SESSION (SESSION 1 -> SESSION 3) EMOTION RECOGNITION RESULTS (%) OF THE FIVE MODELS ON "SEED_IV" DATA SET

	session1->session3				
	1NN	LSR	SVM	FIL	GFIL
subject 1	37.23	68.86	50.85	72.14	**74.09**
subject 2	56.33	66.67	72.02	72.75	**80.54**
subject 3	41.36	51.70	41.73	70.07	**75.06**
subject 4	49.03	89.54	54.01	85.77	**86.74**
subject 5	34.06	**67.27**	50.24	64.84	66.18
subject 6	53.04	77.49	**84.67**	81.14	82.12
subject 7	40.27	79.93	65.45	**81.27**	**79.20**
subject 8	69.59	76.64	**84.55**	72.99	81.02
subject 9	47.93	46.96	59.49	62.90	**70.19**
subject 10	27.74	59.61	34.79	60.58	**68.61**
subject 11	43.19	**73.72**	62.53	65.45	71.53
subject 12	35.16	52.55	30.17	53.04	**56.57**
subject 13	47.20	56.45	54.62	49.15	**57.91**
subject 14	41.48	74.70	63.38	84.06	**85.28**
subject 15	75.79	80.54	84.06	85.28	**90.27**
average	46.63	68.18	59.50	70.76	**75.02**

TABLE III
CROSS SESSION (SESSION 2 -> SESSION 3) EMOTION RECOGNITION RESULTS (%) OF THE FIVE MODELS ON "SEED_IV" DATA SET

	session2->session3				
	1NN	LSR	SVM	FIL	GFIL
subject 1	44.04	59.61	66.79	**69.59**	**69.59**
subject 2	55.11	68.37	82.12	78.71	**82.60**
subject 3	56.81	65.21	61.07	70.44	**70.92**
subject 4	42.94	85.40	62.04	87.83	**90.15**
subject 5	48.78	68.25	62.04	**72.26**	**72.26**
subject 6	65.09	86.13	67.40	90.27	**92.34**
subject 7	81.75	82.97	81.14	**89.78**	**89.78**
subject 8	49.03	**78.22**	73.97	72.14	74.70
subject 9	31.87	74.94	56.20	**75.67**	**75.67**
subject 10	67.15	76.76	67.03	71.78	**81.14**
subject 11	51.46	56.69	53.65	**64.72**	**64.72**
subject 12	42.46	79.44	64.84	75.18	**79.81**
subject 13	49.64	51.82	50.85	58.64	**61.68**
subject 14	63.99	86.62	82.12	**89.17**	**89.17**
subject 15	87.71	86.01	85.28	86.37	**93.07**
average	55.86	73.76	67.77	76.84	**79.17**

training while the ones from session 2 as test samples, denoted as "session 1->session 2"; similarly, we had "session 1-> session 3" and "session 2->session 3."

We compared the proposed GFIL model with some popular classifiers, including the nearest neighbor (1NN), LSR, and support vector machine (SVM). To show the effectiveness of the graph regularization term in (2), we additionally included the comparison with FIL. Obviously, FIL is the degenerated version of GFIL when the parameter β in objective (1) is set as 0. In SVM, the linear kernel is used. Parameters involved in respective models (C in SVM and the regularization parameter in LSR) were searched from candidates $\{2^{-20}, 2^{-19}, \ldots, 2^{20}\}$. The Euclidean distance was used in 1NN. The best recognition results obtained by different models were reported. Tables I–III, respectively, show the emotion recognition accuracies to the three cross-session settings, where the best results are highlighted in boldface.

From the above results, we have the following findings.

1) Though there exist distribution discrepancies for EEG data collected from different sessions, we still obtained promising recognition accuracies in cross-session emotion recognition. The best average accuracies of these 15 subjects corresponding to these three cross-session recognition tasks are 75.33%, 75.02%, and 79.17%, respectively. This indicates that each subject actually has stable emotional EEG patterns over time. To be specific, we can consider the emotional EEG data as the superposition of inner emotion-related component and outer information. The latter corresponds to the interferences caused by data collection processes in different sessions while the former is stable despite the time variations.
2) It is obvious that different EEG feature dimensions have different discriminative abilities in emotion recognition. For example, FIL, respectively, achieved 2.06%, 3.08%, and 2.58% improvements on the three cross-session emotion recognition tasks in comparison with LSR. This means that treating the contributions of different feature dimensions equally is not beneficial for EEG-based emotion recognition. Instead, adaptively adjusting their weights can effectively improve the emotion recognition accuracy. Moreover, in most cases, FIL and GFIL are better than the other compared methods, which directly take the concatenated EEG features as inputs.
3) From the machine learning perspective, the label consistency property is effective in regulating the related variables in the learning model, e.g., the transform matrix $\mathbf{W}$ and the autoweighting variable $\mathbf{\Theta}$ in GFIL. Compared to FIL, the performance of GFIL is, respectively, improved by 4.98%, 2.33% and 4.26% in the three cross-session emotion recognition tasks. We can view the label consistency as the supervised generalization of the local invariance idea, which has been proven beneficial in enhancing the learning performance by many existing studies.

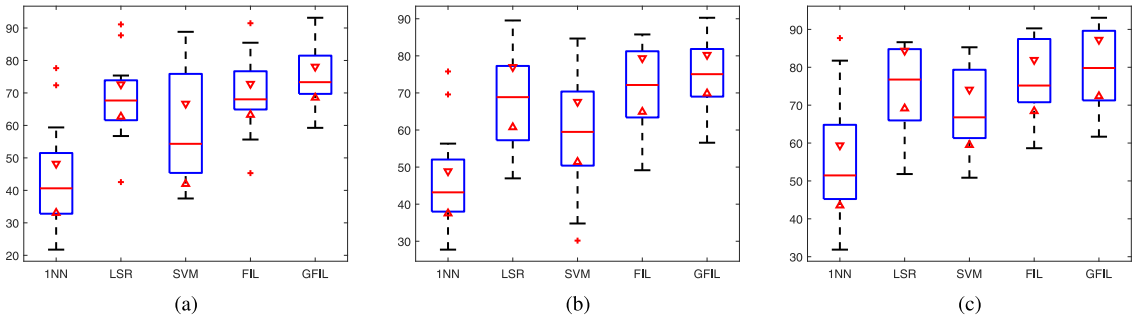


Fig. 6. Emotion recognition results (%) represented by statistical box diagrams. (a) Session 1-> session 2. (b) Session 1-> session 3. (c) Session 2-> session 3.

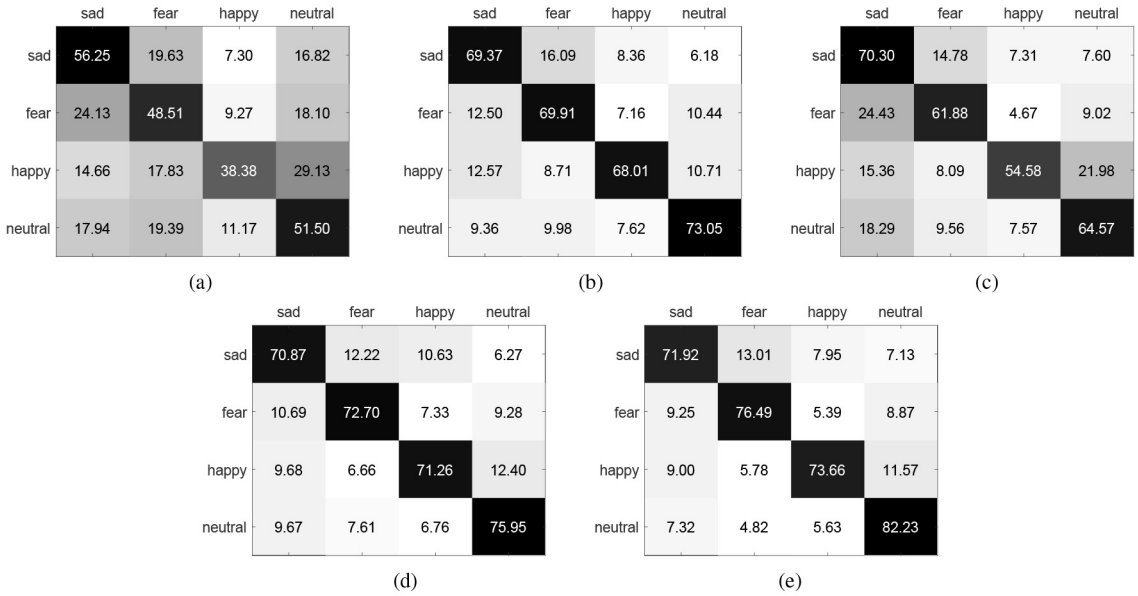


Fig. 7. Confusion matrices of different learning models. (a) 1NN. (b) LSR. (c) SVM. (d) FIL. (e) GFIL.

4) Even for the same task (e.g., session 1 -> session 2), the recognition accuracies vary greatly over these 15 subjects. For example, in Table I, the recognition accuracy of GFIL is 66.59% for subject 1 while it is 93.15% for subject 14. This means that the cross-subject emotion recognition from EEG data might be more challenging due to individual differences. As shown in Fig. 6, we convert the numerical emotion recognition results in Tables I–III into the statistical box diagram representations. Besides the obvious conclusion that GFIL is the best model in terms of average recognition accuracy, we can observe that the results of all these five models have large variances caused by the aforementioned individual differences. Such individual differences act as different outer information mixed with the inner emotional components to form the collected emotional EEG data. This presents us with a more challenging task to build a reliable predictor for simultaneously handling the cross-session and cross-subject differences in EEG-based emotion recognition.

Based on the recognition results, we show the confusion matrices of the five models in Fig. 7 from which we can: 1) obtain the average recognition rate to each of the four emotional states; 2) explore more details about how many samples belonging to a certain emotional state are misclassified to the other states; and 3) find the performance improvement of GFIL on each of the four emotional states in comparison with the other four models. For example, the average recognition accuracy for the *fear* state by FIL is 72.70% while it is 76.49% by GFIL. By checking the confusion matrix of GFIL, we know that 82.23% of the EEG samples belonging to the *neutral* emotional state were correctly recognized while 7.32%, 4.82% and 5.63% of them were misidentified as *sad*, *fear*, and *happy*, respectively. Based on our experimental results, we can generally conclude that the *sad* state is the most difficult emotional state and the *neutral* state is the easiest one to recognize.

E. Importance Evaluation of Frequency Bands

Based on the learned feature importance measure $\boldsymbol{\theta}$, we can automatically judge the critical EEG-frequency bands to emotion recognition according to (17). Taking subject 1 as an example, we, respectively, show the learned $\boldsymbol{\theta}$ s corresponding to the three cross-session recognition tasks in the first three subfigures of Fig. 8, whose average is shown in Fig. 8(d). We can easily find that the features from the *Gamma* frequency

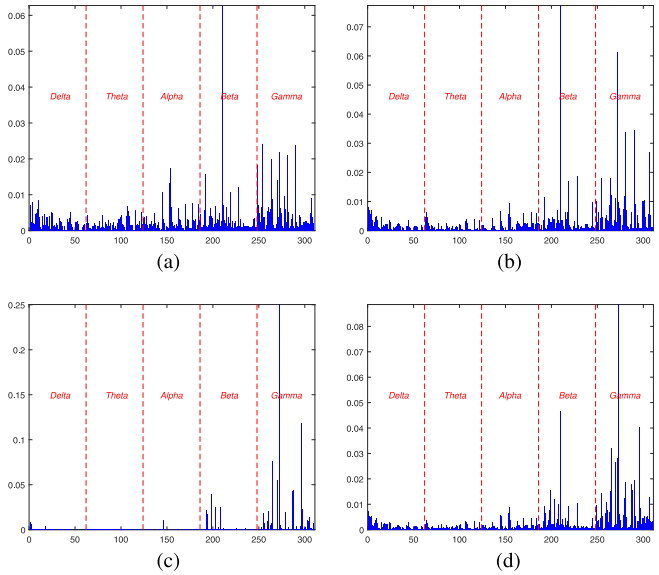


Fig. 8. Learned $\boldsymbol{\theta}$s corresponding to the three cross-session recognition tasks and their average. (a) Session 1-> session 2. (b) Session 1-> session 3. (c) Session 1-> session 3. (d) Subject1.

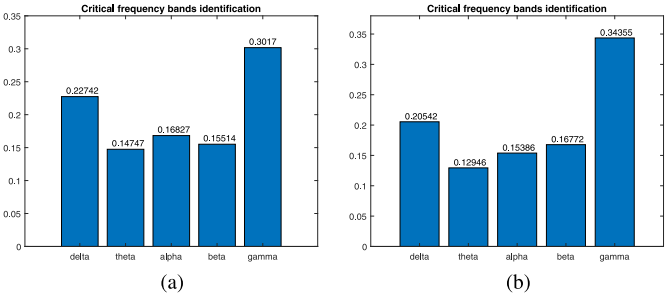


Fig. 9. Frequency bands identification based on the learned $\boldsymbol{\theta}$. (a) FIL. (b) GFIL.

band is the most important for recognizing the emotional states of subject 1.

To suppress the individual differences, we compute the quantitative measure of each frequency band over all the 15 subjects in Fig. 9 by both FIL and GFIL, which also indicates that the *Gamma* band is the most important one in EEG-based emotion recognition. This finding coincides with some existing studies [30], [33]. Compared to the "try-and-error" methods, our proposed feature importance learning-based methods (e.g., FIL and GFIL) are more flexible and adaptive to complete the identification process. What is more, the results obtained by FIL and GFIL are more interpretable than those achieved by deep learning models [30].

F. Channel Selection

Similar to the analysis on critical frequency bands identification, different EEG channels are assumed to contribute differently in emotion recognition, corresponding to different brain areas. According to rule (18), we can make adaptively channel selection and further identify the key brain areas in EEG-based emotion recognition.

Fig. 10 lists the top ten channels, respectively, selected by FIL and GFIL, where the heights of bars quantitatively

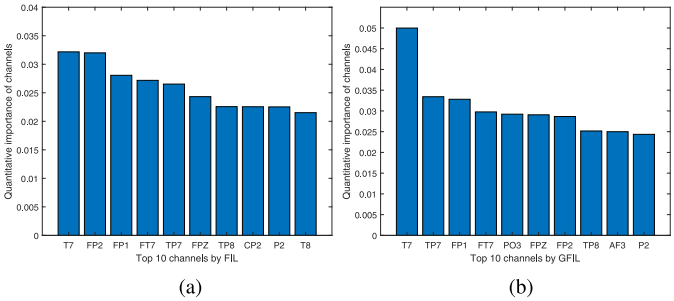


Fig. 10. Top 10 channels, respectively, selected by (a) FIL and (b) GFIL.

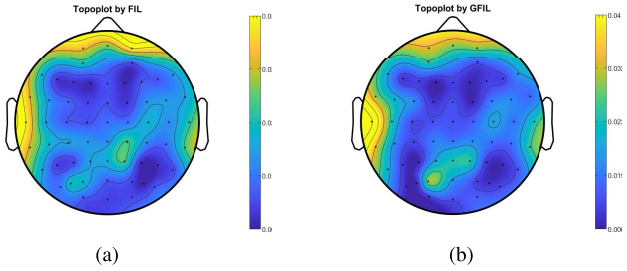


Fig. 11. Topgraphical show of the importance of all the 62 channels, respectively, obtained by (a) FIL and (b) GFIL.

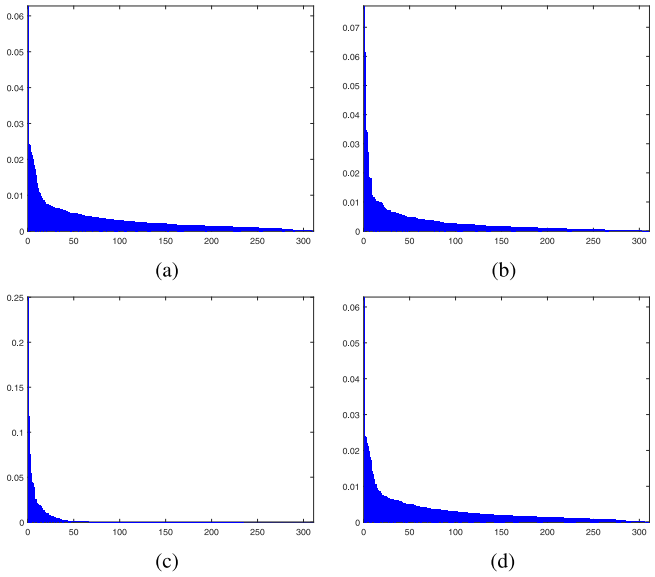


Fig. 12. Ranked feature importance representations of subject 1. (a) Session 1-> session 2. (b) Session 1-> session 3. (c) Session 2-> session 3. (d) Subject 1.

provide us with the importance of different channels. By calculating the importance of all channels according to the learned autoweighting variable $\boldsymbol{\theta}$, the topographical show of the importance of all the 62 channels is provided in Fig. 11. These results demonstrate that the channels locating in the prefrontal and the left/right central regions are more important for emotion recognition. This finding is generally consistent with existing studies [29], which may provide some insights for emotion research from the cognitive neural science perspective and also for engineers to customize the devices for emotional EEG data acquisition.

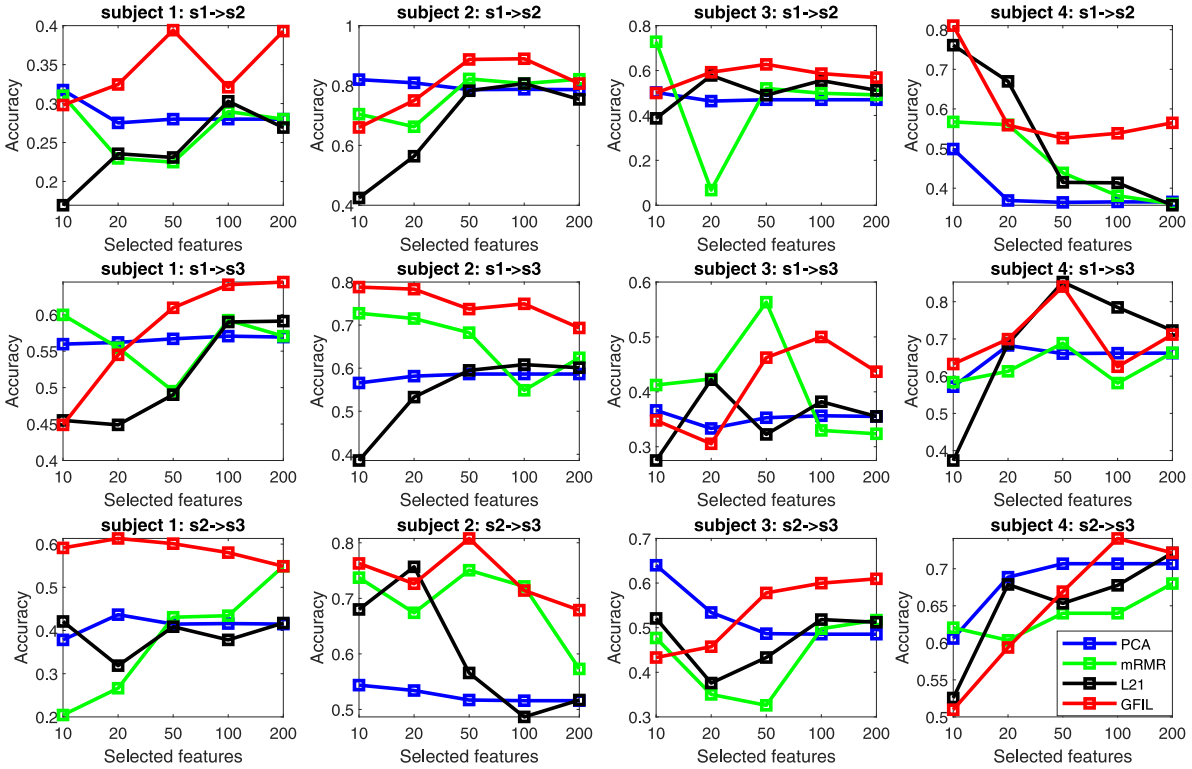


Fig. 13. Feature selection results on subjects 1, 2, 3, and 4.

G. Feature Ranking and Selection

In Fig. 12, we show the ranked feature importance representations of subject 1. We know that the length of the DE-based feature vector in the "SEED_IV" data set is 310; however, only a few of them contribute significantly in emotion recognition from this figure. There are a lot of redundant or less important features in the EEG feature vector of 310-D representation. Similar phenomena can be found in the other subjects. Therefore, as an effective way to improve the recognition performance as well as decreasing the computing and storage resources, feature selection is of great necessity in EEG-based emotion recognition.

Since the learned $\boldsymbol{\theta}$ offers us the rank of features and can be deservedly utilized to perform feature selection, we evaluated the effectiveness of GFIL in feature selection by comparing it with some popular feature selection models, including the minimum redundancy maximum relevance (mRMR) [34] and the $\ell_{2,1}$-norm [35]. Moreover, we included the principal component analysis (PCA) as a baseline, which is essentially an unsupervised dimensionality reduction model. The SVM classifier with linear kernel was employed to perform classification on the EEG data with selected features. The regularization parameter C was set to 1.

Fig. 13 shows the feature selection results of the four models on four subjects in which the number of selected features were, respectively, set as 10, 20, 50, 100, and 200. From this figure, it is obvious that GFIL achieved very competitive performance in comparison with the other models. Moreover, we find that the best results were not always obtained when the number of selected features is 200. Therefore, it is necessary to explore more discriminative features as much as possible rather than handle all the feature dimensions equally in EEG-based emotion recognition.

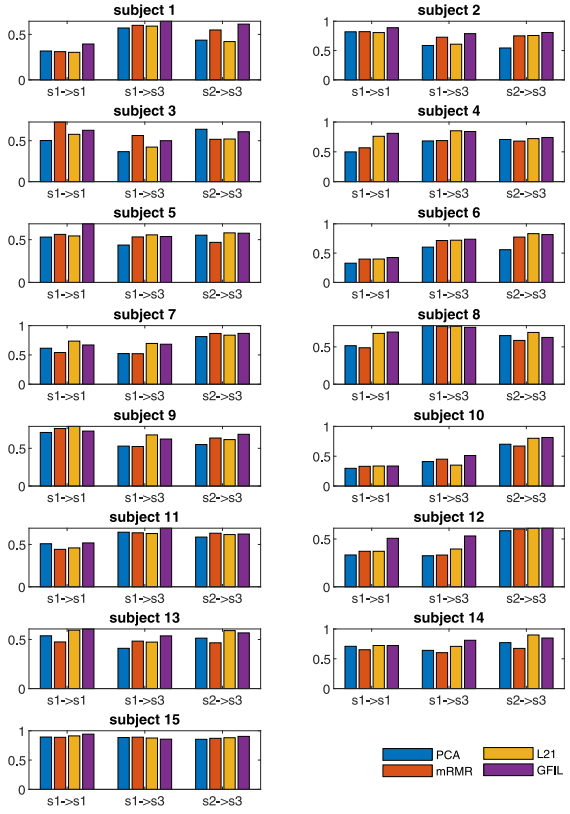


Fig. 14. Best results in term of all the number of selected features, respectively, achieved by the four different models.

The average performance of different models in terms of all the number of selected features on these 15 subjects is provided in Fig. 14. Based on these results, we can conclude that the autoweighting variable in GFIL is effective in selecting more important EEG features for emotion recognition.

V. Conclusion

In this article, we proposed a GFIL for EEG-based emotion recognition. Besides providing the recognition results, the proposed GFIL model can simultaneously complete three tasks of critical frequency band identification, channel selection, and feature selection, which is a pure data-driven model for knowledge mining from emotional EEG data. Extensive experiments were conducted on the SEED_IV emotional EEG data set and the results demonstrated its excellent performance in emotion recognition. We expect that the identification of important frequency bands and brain areas by GFIL might provide some insights into the cognitive research on EEG-based emotion recognition.

In this work, we use the supervised learning paradigm to investigate the cross-session EEG-based emotion recognition tasks. As our future work, we will extend GFIL into semisupervised learning paradigm [36], [37] in which the samples from one session are labeled while those from the other session(s) are unlabeled.

Appendix
Optimization to (10)

We introduce an auxiliary variable $\mathbf{v}$ with respect to $\boldsymbol{\theta}$ to make the objective function separable. Then, problem (13) becomes

$$\min_{\boldsymbol{\theta},\mathbf{v}} \mathbf{v}^T\mathbf{A}\boldsymbol{\theta} - \mathbf{v}^T\mathbf{c}, \ \text{s.t.}\ \boldsymbol{\theta}^T\mathbf{1} = 1, \boldsymbol{\theta} \geq \mathbf{0}, \boldsymbol{\theta} = \mathbf{v}. \tag{19}$$

The corresponding augmented Lagrangian function is

$$\begin{aligned} \min_{\boldsymbol{\theta},\mathbf{v}} \ & \boldsymbol{\theta}^T\mathbf{A}\mathbf{v} - \mathbf{v}^T\mathbf{c} + \frac{\mu}{2}\left\|\boldsymbol{\theta} - \mathbf{v} + \frac{\boldsymbol{\lambda}}{\mu}\right\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{\theta}^T\mathbf{1} = 1, \boldsymbol{\theta} \geq \mathbf{0} \end{aligned} \tag{20}$$

where $\boldsymbol{\lambda}$ is the Lagrangian multiplier and μ is an increment parameter.

Below, we give the derivation of the updating rule to each variable in (20).

1) Update $\mathbf{v}$ with $\boldsymbol{\theta}$ fixed. By taking the derivative of problem (20) with respect to $\mathbf{v}$ and setting it to 0, we have

$$\mathbf{v} = \boldsymbol{\theta} + \frac{1}{\mu}(\mathbf{c} - \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\lambda}). \tag{21}$$

2) Update $\boldsymbol{\theta}$ with $\mathbf{v}$ fixed. When $\mathbf{v}$ fixed, problem (20) becomes

$$\min_{\boldsymbol{\theta}^T\mathbf{1}=1,\boldsymbol{\theta}\geq\mathbf{0}} \mathbf{v}^T\mathbf{A}\boldsymbol{\theta} + \frac{\mu}{2}\left\|\boldsymbol{\theta} - \mathbf{v} + \frac{\boldsymbol{\lambda}}{\mu}\right\|_2^2 \tag{22}$$

which can be further reformulated into

$$\min_{\boldsymbol{\theta}^T\mathbf{1}=1,\boldsymbol{\theta}\geq\mathbf{0}} \frac{\mu}{2}\left\|\boldsymbol{\theta} - \mathbf{v} + \frac{1}{\mu}\left(\boldsymbol{\lambda} + \mathbf{A}^T\mathbf{v}\right)\right\|_2^2. \tag{23}$$

Mathematically, subproblem (23) defines an Euclidean projection on the simplex [38], [39]. Denote $\mathbf{m} = \mathbf{v} - (1/\mu)(\boldsymbol{\lambda} + \mathbf{A}^T\mathbf{v})\|_2^2$ and then problem (23) becomes

$$\min_{\boldsymbol{\theta}^T\mathbf{1}=1,\boldsymbol{\theta}\geq\mathbf{0},\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{m}\|_2^2. \tag{24}$$

The Lagrangian function of problem (24) is

$$\mathcal{L}(\boldsymbol{\theta}, \gamma, \boldsymbol{\eta}) = \frac{1}{2}\|\boldsymbol{\theta} - \mathbf{m}\|_2^2 - \gamma\left(\boldsymbol{\theta}^T\mathbf{1} - 1\right) - \boldsymbol{\eta}^T\boldsymbol{\theta} \tag{25}$$

where γ and $\boldsymbol{\eta}$ are the Lagrangian multipliers to be determined. Assume that $\boldsymbol{\theta}^*$ is the optimal solution and γ^* and $\boldsymbol{\eta}^*$ are the Lagrangian multipliers. According to the KKT condition, we have

$$\begin{cases} \forall t \in [1, d], \ \ \theta_t^* - m_t - \gamma^* - \eta_t^* = 0 \\ \forall t \in [1, d], \ \ \theta_t^* \geq 0 \\ \forall t \in [1, d], \ \ \eta_t^* \geq 0 \\ \forall t \in [1, d], \ \ \theta_t^*\eta_t^* = 0. \end{cases} \tag{26}$$

The first equation in (26) can be rewritten in a vector form as

$$\boldsymbol{\theta}^* - \mathbf{m} - \gamma^*\mathbf{1} - \boldsymbol{\eta}^* = \mathbf{0}. \tag{27}$$

By considering the constraint $\boldsymbol{\theta}^T\mathbf{1} = 1$, (27) can be reformulated into

$$\gamma^* = \frac{1 - \mathbf{1}^T\mathbf{m} - \mathbf{1}^T\boldsymbol{\eta}^*}{d}. \tag{28}$$

By substituting (28) into (27), we have

$$\boldsymbol{\theta}^* = \mathbf{m} - \frac{\mathbf{1}\mathbf{1}^T}{d}\mathbf{m} + \frac{1}{d}\mathbf{1} - \frac{\mathbf{1}^T\boldsymbol{\eta}^*}{d}\mathbf{1} + \boldsymbol{\eta}^*. \tag{29}$$

Suppose that $\bar{\eta}^* = ([\mathbf{1}^T\boldsymbol{\eta}^*]/d)$ and $\mathbf{g} = \mathbf{m} - (\mathbf{1}\mathbf{1}^T/d)\mathbf{m} + (1/d)\mathbf{1}$, the above equation can be rewritten as

$$\boldsymbol{\theta}^* = \mathbf{g} + \boldsymbol{\eta}^* - \bar{\eta}^*\mathbf{1}. \tag{30}$$

Therefore, for any $t \in [1, d]$, we have

$$\theta_t^* = g_t + \eta_t^* - \bar{\eta}^*. \tag{31}$$

According to (26) and (31), we have $g_t + \eta_t^* - \bar{\eta}^* = (g_t - \bar{\eta}^*)_+$ where $(f(\cdot))_+ = \max(f(\cdot), 0)$. Therefore, we have

$$\theta_t^* = \left(g_t - \bar{\eta}^*\right)_+. \tag{32}$$

Now, if the optimal $\bar{\eta}^*$ can be determined, the optimal solution $\boldsymbol{\theta}^*$ can be obtained from (32). Similarly, (31) can also be rewritten as $\eta_t^* = \theta_t^* + \bar{\eta}^* - g_t$ such that $\eta_t^* = (\bar{\eta}^* - g_t)_+$. Therefore, $\bar{\eta}^*$ can be calculated as

$$\bar{\eta}^* = \frac{1}{d}\sum_{t=1}^{d}\left(\bar{\eta}^* - g_t\right)_+. \tag{33}$$

According to the constraint $\boldsymbol{\theta}^T\mathbf{1} = 1$ and (32), we can define the following function:

$$f(\bar{\eta}) = \sum_{t=1}^{d}(g_t - \bar{\eta})_+ - 1 \tag{34}$$

Algorithm 2 Algorithm to Solve Subproblem (24)

Input: vector $\mathbf{m} \in \mathbb{R}^d$
Output: variable $\boldsymbol{\theta} \in \mathbb{R}^d$
1: Compute $\mathbf{g} = \mathbf{m} - \frac{\mathbf{1}\mathbf{1}^T}{d}\mathbf{m} + \frac{1}{d}\mathbf{1}$;
2: Use Newton's method to obtain the root $\bar{\eta}^*$ of (33);
3: Obtain the optimal solution $\theta_t^* = (g_t - \bar{\eta}^*)_+$ for each $t \in [1, d]$;

Algorithm 3 Algorithm to Solve Problem (14)

Input: matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and vector $\mathbf{c} \in \mathbb{R}^d$;
1: Set $\rho > 1$, and initialize $\boldsymbol{\theta} = \frac{\mathbf{1}}{d}$, μ and $\boldsymbol{\lambda}$;
2: **while** not converged **do**
3: Update $\mathbf{v}$ by equation (21);
4: Update $\boldsymbol{\theta}$ by solving problem (23) via Algorithm 2;
5: Update $\boldsymbol{\lambda} = \boldsymbol{\lambda} + \mu(\boldsymbol{\theta} - \mathbf{v})$;
6: Update $\mu = \rho\mu$;
7: **end while**

and the optimal $\bar{\eta}^*$ should satisfy $f(\bar{\eta}^*) = 0$. When (34) equals 0, the optimal $\bar{\eta}^*$ can be obtained by the Newton method [40], namely

$$\bar{\eta}_{k+1} = \bar{\eta}_k - \frac{f(\bar{\eta}_k)}{f'(\bar{\eta}_k)}. \tag{35}$$

We know that $f(\bar{\eta})$ is a piecewise linear and monotonically increasing function. When $g_t \geq \bar{\eta}$, $f(\bar{\eta}) = \sum_{t=1}^{d} g_t - \bar{\eta} - 1$ and we have $f'(\bar{\eta}) = -1$. When $g_t < \bar{\eta}$, $f(\bar{\eta}) = -1$ and its derivative $f'(\bar{\eta}) = 0$. Therefore, we can obtain $f'(\bar{\eta})$ by counting the number of positive values in $(g_t - \bar{\eta})|_{t=1}^{d}$.

Consequently, the optimization procedure to subproblem (24) is listed in Algorithm 2 and the algorithm to solve problem (14) is summarized in Algorithm 3.

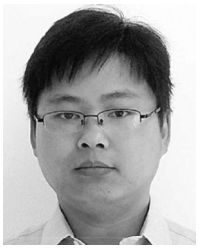
ACKNOWLEDGMENT

The authors would like to thank the editors and reviewers for their valuable comments on this article, and the researchers from Shanghai Jiao Tong University for their kindly providing the "SEED_IV" emotional EEG data set.

REFERENCES

[1] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu, "Differential entropy feature for EEG-based vigilance estimation," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2013, pp. 6627–6630.
[2] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. Int. IEEE/EMBS Conf. Neural Eng.*, 2013, pp. 81–84.
[3] W.-L. Zheng and B.-L. Lu, "A multimodal approach to estimating vigilance using EEG and forehead EOG," *J. Neural Eng.*, vol. 14, no. 2, 2017, Art. no. 026017.
[4] Y. Peng, Q. Li, W. Kong, J. Zhang, B.-L. Lu, and A. Cichocki, "Joint semi-supervised feature auto-weighting and classification model for EEG-based cross-subject sleep quality evaluation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 946–950.
[5] N. E. Huang *et al.*, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London A Math. Phys. Eng. Sci.*, vol. 454, no. 1971, pp. 903–995, 1998.
[6] S. K. Hadjidimitriou and L. J. Hadjileontiadis, "Toward an EEG-based recognition of music liking using time-frequency analysis," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3498–3510, Dec. 2012.
[7] W. Kong, L. Wang, S. Xu, F. Babiloni, and H. Chen, "EEG fingerprints: Phase synchronization of EEG signals as biomarker for subject identification," *IEEE Access*, vol. 7, pp. 121165–121173, 2019.
[8] Z. Ghanbari and M. H. Moradi, "FSIFT-PLV: An emerging phase synchrony index," *Biomed. Signal Process. Control*, vol. 57, pp. 1–12, Mar. 2020.
[9] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Jul.–Sep. 2014.
[10] Y. Gao, H. J. Lee, and R. M. Mehmood, "Deep learninig of EEG signals for emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2015, pp. 1–5.
[11] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Map.*, vol. 38, no. 11, pp. 5391–5420, 2017.
[12] M. S. Hossain, S. U. Amin, M. Alsulaiman, and G. Muhammad, "Applying deep learning for epilepsy seizure detection and brain mapping visualization," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 1s, pp. 1–17, 2019.
[13] J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Exp. Syst. Appl.*, vol. 47, pp. 35–41, Apr. 2016.
[14] W. Zheng, "Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis," *IEEE Trans. Cogn. Develop. Syst.*, vol. 9, no. 3, pp. 281–290, Sep. 2017.
[15] P. Li *et al.*, "EEG based emotion recognition by combining functional connectivity network and local activations," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 10, pp. 2869–2881, Oct. 2019.
[16] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, "Domain adaptation techniques for EEG-based emotion recognition: A comparative study on two public datasets," *IEEE Trans. Cogn. Develop. Syst.*, vol. 11, no. 1, pp. 85–94, Mar. 2019.
[17] D. Wu, Y. Xu, and B.-L. Lu, "Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016," *IEEE Trans. Cogn. Develop. Syst.*, early access, Jul. 2, 2020, doi: 10.1109/TCDS.2020.3007453.
[18] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, "EEG-based emotion classification using deep belief networks," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2014, pp. 1–6.
[19] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, "Emotion recognition from multi-channel EEG data through convolutional recurrent neural network," in *Proc. IEEE Int. Conf. Bioinformat. Biomed.*, 2016, pp. 352–359.
[20] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 532–541, Jul.–Sep. 2020.
[21] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, Jan.–Mar. 2016.
[22] F. Movahedi, J. L. Coyle, and E. Sejdić, "Deep belief networks for electroencephalography: A review of recent contributions and future outlooks," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 3, pp. 642–652, May 2018.
[23] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *J. Neural Eng.*, vol. 16, no. 3, 2019, Art. no. 031001.
[24] J. Zhang, M. Chen, S. Zhao, S. Hu, Z. Shi, and Y. Cao, "ReliefF-based EEG sensor selection methods for emotion recognition," *Sensors*, vol. 16, no. 10, p. 1558, 2016.
[25] A. Goshvarpour and A. Goshvarpour, "A novel approach for EEG electrode selection in automated emotion recognition based on Lagged Poincare's indices and sLORETA," *Cogn. Comput.*, vol. 12, no. 3, pp. 602–618, 2020.
[26] M. S. Özerdem and H. Polat, "Emotion recognition based on EEG features in movie clips with channel selection," *Brain Informat.*, vol. 4, no. 4, pp. 241–252, 2017.
[27] Z.-M. Wang, S.-Y. Hu, and H. Song, "Channel selection method for EEG emotion recognition using normalized mutual information," *IEEE Access*, vol. 7, pp. 143303–143311, 2019.
[28] T. Alotaiby, F. E. A. El-Samie, S. A. Alshebeili, and I. Ahmad, "A review of channel selection algorithms for EEG signal processing," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 8, pp. 1–21, Aug. 2015. [Online]. Available: https://dblp.uni-trier.de/db/journals/ejasp/ejasp2015.html
[29] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 417–429, Jul.–Sep. 2019.

[30] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.
[31] Y. Peng, S. Wang, X. Long, and B.-L. Lu, "Discriminative graph regularized extreme learning machine and its application to face recognition," *Neurocomputing*, vol. 149, pp. 340–353, Feb. 2015.
[32] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.
[33] Y. Peng and B.-L. Lu, "Discriminative manifold extreme learning machine and applications to image and EEG signal classification," *Neurocomputing*, vol. 174, pp. 265–277, Jan. 2016.
[34] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[35] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint ℓ2,1-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
[36] X. Chen, G. Yuan, F. Nie, and Z. Ming, "Semi-supervised feature selection via sparse rescaled linear square regression," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 1, pp. 165–176, Jan. 2020.
[37] X. Chen, F. Nie, G. Yuan, and J. Z. Huang, "Semi-supervised feature selection via rescaled linear regression," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 1525–1531.
[38] A. Kyrillidis, S. Becker, V. Cevher, and C. Koch, "Sparse projections onto the simplex," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 235–243.
[39] Y. Peng, Q. Li, W. Kong, F. Qin, J. Zhang, and A. Cichocki, "A joint optimization framework to semi-supervised RVFL and ELM networks for efficient data classification," *Appl. Soft Comput.*, vol. 97, pp. 1–15, Dec. 2020.
[40] A. H. Sherman, "On Newton-iterative methods for the solution of systems of nonlinear equations," *SIAM J. Numer. Anal.*, vol. 15, no. 4, pp. 755–771, 1978.



Yong Peng (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2015.

He is currently an Associate Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. His main research interests are machine learning, pattern recognition, and EEG-based brain–computer interfaces.

Dr. Peng was awarded by the President Prize of Chinese Academy Sciences in 2009 and the Third Prize of Chinese Institute of Electronics in 2018.

Feiwei Qin received the Ph.D. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2014.

He is currently an Associate Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. His research interests include artificial intelligence, computer vision, computer graphics, computer-aided design, and medical image analysis.

Wanzeng Kong (Member, IEEE) received the Ph.D. degree from the Department of Electrical Engineering, Zhejiang University, Hangzhou, China, in 2008.

He was a Visiting Research Associate with the Department of Biomedical Engineering, University of Minnesota Twin Cities, Minneapolis, MN, USA, from 2012 to 2013. He is currently a Full Professor and the Director of the Cognitive Computing and BCI Laboratory, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. His current research interests include biomedical signal process, brain–computer interface, cognitive computing, and pattern recognition.

Yuan Ge received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2011.

He is currently a Professor with the School of Electrical Engineering, Anhui Polytechnic University, Wuhu, China. His main research interests include artificial intelligence, energy Internet, and network-based system analysis and control.

Feiping Nie (Senior Member, IEEE) received the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2009.

He is currently a Full Professor with Northwestern Polytechnical University, Xi'an, China. He has published more than 100 papers in the following top journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *International Journal of Computer Vision*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *Bioinformatics*, ICML, NIPS, KDD, IJCAI, AAAI, ICCV, CVPR, and ACM MM. His papers have been cited more than 10 000 times. His research interests include machine learning and its applications, such as pattern recognition, data mining, computer vision, image processing, and information retrieval.

Prof. Nie is currently an Associate Editor or a PC member for several prestigious journals and conferences in the related fields.

Andrzej Cichocki (Fellow, IEEE) received the M.Sc. (Hons.), Ph.D., and Dr.Sc. (Habilitation) degrees in electrical engineering from Warsaw University of Technology, Warszawa, Poland.

He spent several years with the University of Erlangen–Nuremberg, Erlangen, Germany, as an Alexandervon-Humboldt Research Fellow and a Guest Professor. He was a Senior Team Leader and the Head of the Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Saitama, Japan. He is currently a Professor with the Skolkovo Institute of Science and Technology, Moscow, Russia, and also an Adjunct Professor with Hangzhou Dianzi University, Hangzhou, China, since April 2018. He has authored more than 500 technical journal papers and six monographs in English (two of them translated to Chinese). His current research focus on multiway blind source separation, tensor decompositions, tensor networks, deep learning, human–robot interactions, and brain–computer interface.

Dr. Cichocki served as an Associated Editor for IEEE TRANSACTION ON SIGNALS PROCESSING, IEEE TRANSACTION ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANS ON CYBERNETICS, and *Journal of Neuroscience Methods*. He is the Founding Editor-in-Chief of *Journal Computational Intelligence and Neuroscience*.