Computer Methods and Programs in Biomedicine 140 (2017) 93–110

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Recognition of emotions using multimodal physiological signals and an ensemble deep learning model

Zhong Yin[a,*], Mengyuan Zhao[b], Yongxiong Wang[a,*], Jingdong Yang[a], Jianhua Zhang[c]

[a] Engineering Research Center of Optical Instrument and System, Ministry of Education, Shanghai Key Lab of Modern Optical System, University of Shanghai for Science and Technology, Shanghai, 200093, PR China
[b] School of Social Sciences, University of Shanghai for Science and Technology, Shanghai, 200093, PR China
[c] Department of Automation, East China University of Science and Technology, Shanghai 200237, PR China

ARTICLE INFO

Article history:
Received 20 May 2016
Revised 31 October 2016
Accepted 12 December 2016

Keywords:
Emotion recognition
Affective computing
Physiological signals
Deep learning
Ensemble learning

ABSTRACT

Background and Objective: Using deep-learning methodologies to analyze multimodal physiological signals becomes increasingly attractive for recognizing human emotions. However, the conventional deep emotion classifiers may suffer from the drawback of the lack of the expertise for determining model structure and the oversimplification of combining multimodal feature abstractions.
Methods: In this study, a multiple-fusion-layer based ensemble classifier of stacked autoencoder (MESAE) is proposed for recognizing emotions, in which the deep structure is identified based on a physiological-data-driven approach. Each SAE consists of three hidden layers to filter the unwanted noise in the physiological features and derives the stable feature representations. An additional deep model is used to achieve the SAE ensembles. The physiological features are split into several subsets according to different feature extraction approaches with each subset separately encoded by a SAE. The derived SAE abstractions are combined according to the physiological modality to create six sets of encodings, which are then fed to a three-layer, adjacent-graph-based network for feature fusion. The fused features are used to recognize binary arousal or valence states.
Results: DEAP multimodal database was employed to validate the performance of the MESAE. By comparing with the best existing emotion classifier, the mean of classification rate and F-score improves by 5.26%.
Conclusions: The superiority of the MESAE against the state-of-the-art shallow and deep emotion classifiers has been demonstrated under different sizes of the available physiological instances.

1. Introduction

1.1. Overview

Since collaborations between human and machines (or computers) exist in various working or living environments, researchers in the area of ergonomics and intelligent systems attempt to improve efficiency and flexibility of human-computer interaction (HCI) with high satisfaction levels of human agent [1]. Such intelligent HCI systems require the capability of self-adaptation of computers [2], in which the accurate comprehension of human communications is necessary for machine agent to trigger proper feedback [3]. The human intentions can be expressed in a verbal or a non-verbal manner that carries different *emotions*. A key point of approaching computer adaptability is to develop its functionality of understanding human affective behaviors [4]. This emerging research area is known as affective computing [5–7] regarding the fact that most of the contemporary HCI systems suffer from the lack of intelligence for recognizing emotional cues related to human psychophysiological states [8–10].

Emotions are known as a group of affective states of human being arising as responses to some stimuli from external environments or interpersonal events [11]. Different emotions possess critical influences on self-motivation generation and preferences of decision-making [12]. Representations of emotions include discrete scales in terms of angry, nervous, pleased, bored and so forth or using arousal-valence plane [13–15]. For the latter, 2-dimensional coordinates describe the nature of emotional experience via the core of the affections [16]. The arousal dimension is used to quantify different degrees from calm to excitement levels while the valence dimension indicates whether human feelings are positive (happy) or negative (sad) [17–20]. Fig. 1 shows a typical layout

* Corresponding author.
E-mail addresses: yinzhong@usst.edu.cn, seesawxe@126.com (Z. Yin), wyxiong@usst.edu.cn (Y. Wang).

http://dx.doi.org/10.1016/j.cmpb.2016.12.005
0169-2607/

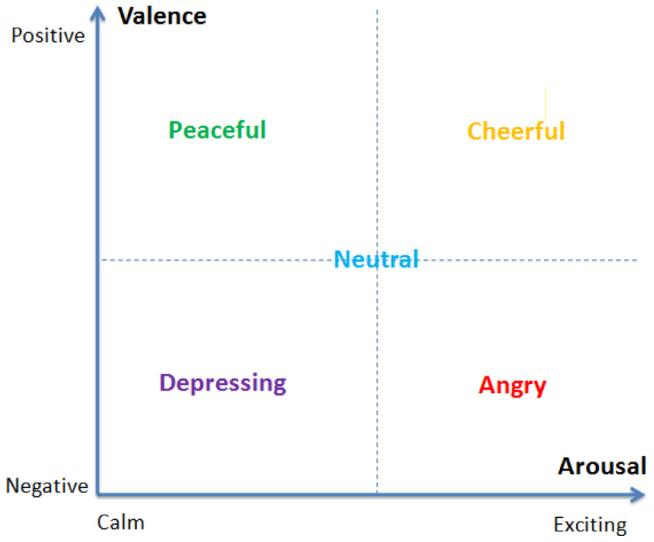

Fig. 1. Arousal-valence plane.

of the arousal-valence plane, where multiple discrete emotional states, e.g., neutral, cheerful, peaceful, depressing and angry, can be defined with different combinations of arousal and valence levels.

1.2. Emotion recognition using physiological signals and pattern classifiers

The function of an intelligent emotion estimator or classifier is to detect emotional clues from human reactions, integrate emotional responses and finally give the prediction of the transient emotional state. The corresponding approaches are mainly classified as two categories, i.e., facial/vocal expressions [21] and physiological signals [22]. Due to specific users who are conditioned to be expressionless, the generalization capability of the behavior data may be limited [23,24]. On the other hand, physiological measures that record electrophysiological information in real time from central nervous system (CNS) or peripheral nervous system (PNS) become attractive because of their the repeatability and objectivity to infer human cognitive or affective state as well as ease of use with a portable implementation via wireless data transmission devices [25].

In well-documented works, the accessibility of various physiological signals for evaluating emotions has been investigated [26–34]. More specifically, emotion variation could be identified via electroencephalogram (EEG) from several frontal and parietal cortical areas. Verma and Tiwary indicated the EEG power spectral density (PSD) in alpha (8–13 Hz) band significantly varies with different valence levels [26]. Frantzidis et al. showed that an EEG feature subset of delta (1–4 Hz) and theta (4–7 Hz) PSD extracted from three central channels (Cz, Fz, and Pz) are quite useful for indicating both arousal and valence levels [28]. The phase synchronization and coherence between EEG channel pairs in the brain areas with functional connectivity were found as effective emotion indicators [29]. The usability of event-correlated potential (ERP) was also examined. Konstantinidis et al. extracted the ERP components of N100 and N200 to classify emotions in arousal-valence plane [30]. Frantzidis et al. calculated P100 and P300 for emotion recognition [29]. In addition, the multimodal PNS physiological signals, e.g., galvanic skin response (GSR) [31], electrooculargram (EOG) [32], electromyogram (EMG) [33], and electrocardiogram (ECG) [34], were extensively explored.

Considering high spatial and temporal resolutions of sophisticated CNS and PNS signal acquisition devices, machine learning approaches facilitate analyzing the massive volume of neurophysiological data [35–39]. In particular, pattern classifiers could fuse physiological features of different modality. Recently, Iacoviello et al. have combined discrete wavelet transformation, principal component analysis (PCA) and support vector machine (SVM) to build a hybrid classification framework [38]. Khezri et al. employed three-channel forehead EEG combined with GSR to recognize six basic emotions via *K*-nearest neighbors (KNN) classifiers [31]. Verma et al. [26] developed an ensemble classification approach fusing EEG, EMG, ECG, GSR, and EOG. Mehmood and Lee used independent component analysis to extract emotional indicators from EEG, EMG, GSR, ECG, and ERP [39].

Due to the superiority of abstracting high-dimensional physiological features, a number of deep learning approaches were investigated for emotion classification and elicit promising results. The popular deep learning primitives include deep belief networks (DBN), stacked autoencoders (SAE) and convolutional neural nets (CNN). In particular, Wand and Shang adopted the standard DBN to extract features from raw physiological data based on unsupervised pre-training and build three deep classifiers to estimate the levels of arousal, valance, and liking [40]. The classification accuracies of DEAP database are 60.9%, 51.2%, and 68.4%, respectively. Similarly, Li et al. adopted a two-layer DBN ensemble to fuse multichannel EEG data in DEAP and the binary emotion classification accuracies of arousal and valence scales are 0.5840, and 0.6420, respectively [41]. Li et al. employed the supervised restrict Boltzmann machine (RBM) to modify the standard DBN and proposed the supervised DBN based affective state recognition (SDA) model [42]. By using the EEG data of DEAP as the deep model inputs, the average AUC (i.e., the area under the receiver operating characteristic curve) score is 0.75. Jia et al. proposed the semi-supervised deep learning model (semi-DLM) based on DBNs for binary emotion classification [43]. The essential of the semi-DLM classifier is to utilize the label information for EEG channel selections instead of the pre-training procedure of the DBNs. The average AUC score of the liking scale in DEAP is 0.7890. Jirayucharoensak et al. combined the dimensionality reduction technique, i.e., PCA, with the standard SAE network to build emotion classifiers [44]. The designed SAE network possessed two hidden layers with 100 hidden neurons in each layer. Based on three levels of arousal and valence scales targeted in DEAP, the average classification accuracies are 0.4952 and 0.4603, respectively. Besides the physiological signals, Acar et al. combined the CNN and SVM to identify four affective categories from the audio and visual modality of videos [45].

1.3. Motivation of the present study

The brief literature review suggests the machine-learning-based methodologies are promising to reveal the latent patterns of certain emotional states hidden in the physiological signals. In particular, the deep classifiers are able to abstract the intermediate representations of physiological features in multiple modalities via hierarchical architectures. However, the deep network structure of emotion classifiers is usually selected based on the prior knowledge from other domains. Considering the nature of the high dimensionality and limited training instances of the physiological data, transferring the empirical expertise from massive data problems may not be always reliable. More specifically, too deep network with too many hidden neurons in each layer may lead to the severe model overfitting. The oversimplification and insufficient abstraction of physiological features could arise when employing too simple model structure. Hence, it is necessary to develop a physiological-data-driven approach to identify the optimal topology of the deep emotion classifier.

On the other hand, the classifier ensemble has the capability to tackle the multimodality in physiological signals since it improves

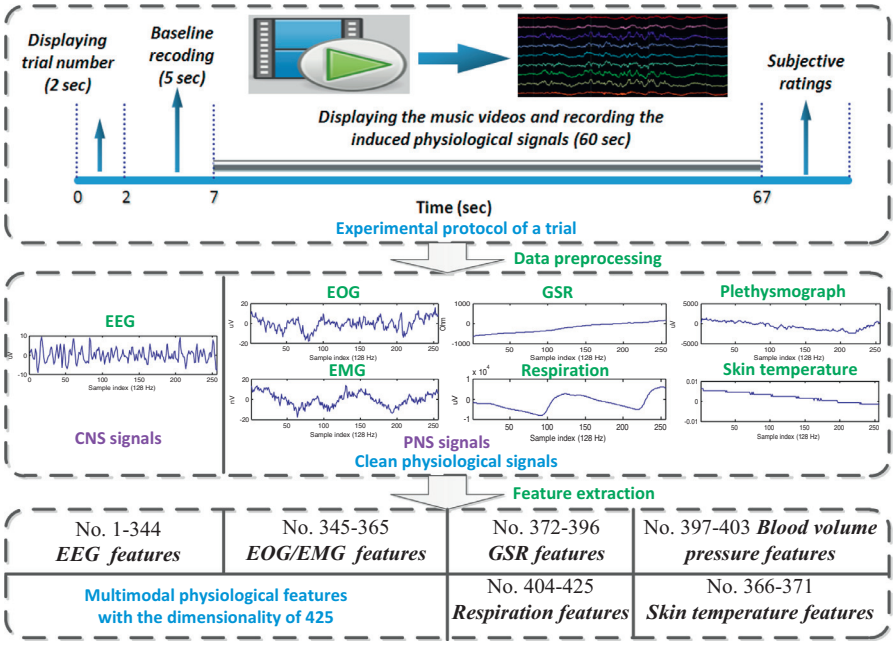

Fig. 2. Illustration of physiological data acquisition of DEAP database and the employed modality for feature extraction.

the classification robustness against the insignificance of specific modality caused by individual variance [46]. The conventional ensembles of shallow classifiers were mainly constructed via the decision fusion scheme. That is, multiple member classifiers process different modalities separately and an ensemble classification committee is constructed to yield the final emotion predictions. However, the discrimination capability in higher abstraction levels of physiological features has been ignored and the ensemble classifier may inherit the defects of specific member shallow models. Hence, a classifier ensemble fusing physiological abstraction via deep architectures is promising to further improve the generalizability.

To address the above two issues, we develop an abstraction-fusion-based deep-learning emotion classifier with the network structure identified by physiological data. The intermediate representations of physiological features in multiple modalities are separately extracted via deep architectures using stacked autoencoders (SAEs) [47]. For each member deep model, a structural loss function is proposed and validated to identify the optimal number of hidden layers and neurons. Then, an adjacent graph based feature fusion network is constructed to merge those high-level abstractions to further enhance the discrimination capability between different emotional states. A Bayesian-model based decision layer is used to obtain the final emotion predictions. The proposed multiple-fusion-layer based ensemble classifier of SAE (MESAE) is validated via the DEAP database of 32 participants and compared against several shallow classifiers, different fusion schemes of multimodal data, popular deep learning approaches and the existing studies on the same database.

The contributions of the present study can be summarized as three aspects: (1) A new physiological-data-driven approach for structure identification of deep models, (2) a new abstraction fusion method that is specifically designed for merging multimodal physiological representations in high levels, and (3) a new hybrid emotion classification framework using deep-ensemble learning paradigm that incorporates the above two approaches.

The rest of the paper is organized as follows. A short description of the DEAP database, procedures of extracting physiological features and the details of MESAE emotion classification framework are presented in Section 2. In section 3, the main classification results on arousal and valence dimensions are shown. Both of the classification performance comparison and the statistical analysis are performed. Some useful discussions on MESAE properties are given in Section 4. A short conclusion is drawn in Section 5.

2. Materials and methods

2.1. Descriptions of DEAP database

The database for emotion analysis using physiological signals (DEAP) was employed to validate the effectiveness of MESAE classifier. Koelstra et al. built the DEAP database aiming at examining spontaneous human affective states that are specifically induced by music videos [48]. The dataset contains 32 healthy participants (19–37 years, *mean* = 26.9, 50% females). For each participant, 40 videos were separately presented in 40 trials with the EEG and peripheral physiological signals simultaneously recorded. In each of them, the index of the current trial was first displayed for 2-sec; and a consecutive 5-sec recording proceeded as the baseline condition; then the music video was shown for 1 minute; finally, the subjective-ratings on arousal, valence, liking and dominance scales were collected (as shown in Fig. 2).

2.2. Data preprocessing and feature extraction

In this study, the 60-sec-length physiological signals with a sampling rate of 128 Hz and the corresponding subjective ratings for 32 participants are used for generating inputs and target emotions for classifiers, respectively. EEG data of 32 channels based on 10–20 system and the peripheral physiological data of 13 channels are available. We adopted independent component analysis to remove EOG artifacts in EEG. A band-pass filter with cutoff frequencies of 4.0 and 45.0 Hz was applied to correct the low frequency respiration and high frequency scalp EMG disturbance for EEG and EOG signals. Then, we extracted 425 salient physiological features from EEG, EOG, EMG, skin temperature, GSR, blood volume pressure, and respiration signals according to [31–35, 48] (see Table 1). In order to investigate the performance of emotion classifiers under different sizes of available instances, all physiological

Table 1
Index and corresponding notations of physiological features.

Feature index	*Notations of the extracted features*
No. 1–160 ***EEG power features***	Average PSD in theta (4–8 Hz), slow-alpha (8–10 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–45 Hz) bands for all EEG channels (5 power × 32 channels): Fp1, AF3, F3, F7, FC5, FC1, C3, T7, CP5, CP1, P3, P7, PO3, O1, Oz, Pz, Fp2, AF4, Fz, F4, F8, FC6, FC2, Cz, C4, T8, CP6, CP2, P4, P8, PO4, and O2.
No. 161–216 ***EEG power differences***	Difference of average PSD in theta, alpha, beta, and gamma bands for 14 EEG channel pairs between right and left scalp (4 power differences × 14 channel pairs): Fp2- Fp1, AF4- AF3, F4-F3, F8-F7, FC6- FC5, FC2-FC1, C4- C3, T8- T7, CP6- CP5, CP2- CP1, P4- P3, P8-P7, PO4- PO3, and O2- O1.
No. 217–344 ***EEG time-domain features***	Mean, variance, zero-crossing rate, and approximate entropy of 32 EEG channels (4 features × 32 channels).
No. 345–349 ***EOG and EMG frequency-domain features***	Eye blink rate, average PSD of vertical, horizontal EOG, Trapezius EMG, and Zygomaticus EMG (1 feature + 1 feature × 4 channels).
No. 350–365 ***EOG and EMG time-domain features***	Mean, variance, zero-crossing rate, and the approximate entropy of 4 EOG and EMG channels (4 features × 4 channels).
No. 366–371 ***Skin temperature features***	Average PSD in the frequency bands (0–0.1 Hz) and (0.1–0.2 Hz), mean, variance, approximate entropy, and mean of derivative (6 features).
No. 372–396 ***GSR features***	Mean, mean of derivative, mean of negative derivative values, proportion of negative values in all derivative values, number of local minima, mean of rising time, 15 PSD values in frequency band (0–2.4 Hz), zero-crossing rates and means for the bands of (0–0.2 Hz) and (0–0.8 Hz) (25 features).
No. 397–403 ***Blood volume pressure features***	Power ratio between the frequency bands of (0.04–0.15 Hz) and (0.15–0.5 Hz), average PSD in the frequency bands of (0.1–0.2 Hz), (0.2–0.3 Hz), (0.3–0.4 Hz), (0.01–0.08 Hz), (0.08–0.15 Hz), and (0.15–0.5 Hz) (7 features).
No. 404–425 ***Respiration features***	Power ratio between frequency bands of (0.05–0.25 Hz) and (0.25–0.5 Hz), mean, mean of derivative, centroid of PSD, respiration rate, 15 values of PSD in frequency band (0–2.4 Hz), mean and media of peak-to-peak time (22 features).

features are computed in two cases. For *Case 1*, one feature vector is computed from a trial and the data matrix of each participant is with the size of 40×425 (trials × features). For all participants, $40 \times 32 = 1280$ instances are available. For *Case 2*, 60-sec physiological signals are evenly divided into 10 consecutive, non-overlapped segments. The 6-sec length of each segment is employed because we found the rising-time of GSR and the respiration rate cannot be stably computed under a shorter length. The segments are non-overlapped aiming at avoiding the intra-trial redundant information for training classifiers. Accordingly, 10 feature vectors are elicited from 10 segments in a trial. The data matrix of 40 trials of each participant is with the size of 400×425. For all participants, $400 \times 32 = 12{,}800$ instances are available. To remove the difference in feature scales, the data of each feature for both of the two cases was standardized for each participant. That is, each column of a data matrix is mapped to mean = 0 and s.d. = 1.

The self-assessment markers on arousal, valence, liking, and dominance of each trial are from the range of 1 to 9 and the levels of the emotion scales from low (1) to high (9) were indicated. In this study, we focused on arousal and valence scales for the following two reasons. (1) According to Koelstra et al. [48], Wilcoxon signed-rank tests results indicate there is a significant difference between low and high valence ratings given either low or high arousal conditions ($p < 0.0001$). The significant difference between low and high arousal ratings was also found when the low or high valence stimuli are given ($p < 0.001$). That is, the arousal and valence conditions are considered as two complementary aspects for stimulating emotions in DEAP experiments and the independent analysis of the two dimensions are facilitated. This observation was supported by their insignificant correlation of 0.18. (2) The high correlations between liking and valence (0.62) as well as between dominance and valence scales are found (0.51). The correlation coefficient between liking and dominance scales are also marginally large (0.31). Koelstra et al. [48] suggests it may be caused by the fact that people liked music and may give them a positive feeling as well as a feeling of empowerment. To this end, besides the risk for misclassification between low and high emotion levels, building a liking or dominance classifier may increase the additional risk of misclassification across different emotional dimensions. To investigate the algorithm performance in a simpler environment, the dominance and liking scales are excluded in the current study.

To find the proper thresholds for labeling different valence and arousal classes, *k*-means clustering is performed to elicit optimal clusters of the self-assessment data. The clustering results are illustrated in Fig. 3, where the numbers of clusters are selected as 2 and 3, respectively. In Fig. 3(a), two clusters are represented by using the scatter plot on the valence-arousal plane from 32 participants (1280 points). Each cluster can be considered as a preference of the self-assessment. One observation is that the point (5, 5) is approximately located on the boundary between two clusters and the mean vector of the centers of the two clusters is also around (5, 5). That is, when assuming there are two different rating preferences, the location of (5, 5) is suitable to differentiate the scores into two categories. It suggests that the target emotion classes can be predetermined as two classes, i.e., *low* or *high* according to the threshold of 5 on each dimension. On the other hand, when three different preferences are considered in Fig. 3(b), three clusters are derived and indicate the preferences of (1) high-valence and low-arousal, (2) low valence, and (3) high-valence and high-arousal. Note that none of clusters is located in the center of the plot in Fig. 3(b). It implies the neutral state is not a common preference when *k* is limited. To this end, the neutral level is not considered as an emotional class in the current study. Instead, the binary emotion classification on the arousal and valence dimensions is employed and the target emotion class is the same for all feature vectors belonging to the same trial.

2.3. Ensemble deep learning model for emotion recognition

2.3.1. Stacked autoencoder

A deep-learning based emotion classifier hypothesizes that a hierarchy of intermediate representations from the physiological features is necessary to characterize the underlying feature properties related to different emotional states. In many real-world applications, SAE has shown its competence and is benefit from a deep-structure-based learning mechanism with multiple hidden layers achieved by a pre-training strategy for unsupervised feature learning and a fine-tuning stage based on the pre-trained network parameters [49].

The basic component of SAE is termed as autoencoder (AE) [50]. AE is built via a single-hidden-layer neural network of feed-forward architecture with the same input-output layers. We denote its input and hidden activations as $\mathbf{x} \in R^D$ and $\mathbf{x}_h \in R^d$, where D

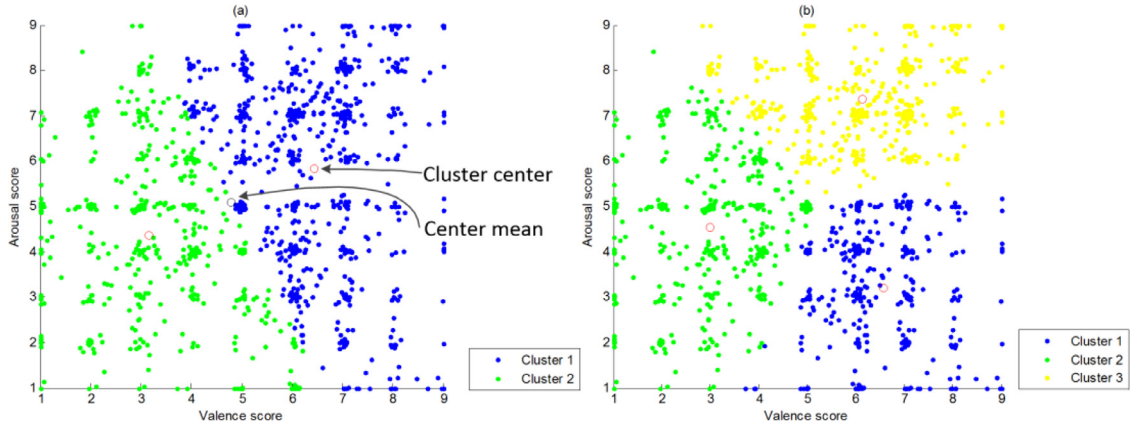

Fig. 3. Clustering results of the self-assessment data on the valence-arousal plane by using k-means clustering method: (a) $k=2$ for two emotion clusters, (3) $k=3$ for three emotion clusters.

and d indicate the input and abstracted dimensionalities of physiological features, respectively. The transformation from $\mathbf{x}$ to $\mathbf{x}_h$ is achieved by,

$$\mathbf{x}_h = s(\mathbf{W}\mathbf{x}+\mathbf{b}), \tag{1}$$

with $\mathbf{W} \in R^{d\times D}$, $\mathbf{b} \in R^d$ and $s(\bullet)$ denoting weight matrix, bias vector and sigmoid function $s(z)=1/(1+e^{-z})$. The input vector $\mathbf{x}$ are then reconstructed as $\tilde{\mathbf{x}} \in R^D$ from hidden to output neurons with tied weights, $\tilde{\mathbf{W}} = \mathbf{W}^T$,

$$\tilde{\mathbf{x}} = s(\tilde{\mathbf{W}}\mathbf{x}_h+\tilde{\mathbf{b}}) = s(\tilde{\mathbf{W}}\cdot s(\mathbf{W}\mathbf{x}+\mathbf{b})+\tilde{\mathbf{b}}). \tag{2}$$

The AE parameters $\mathbf{W}$, $\tilde{\mathbf{W}}$, $\mathbf{b}$ and $\tilde{\mathbf{b}}$ are learned via BP algorithm with squared-error cost function,

$$L(\mathbf{x},\tilde{\mathbf{x}}) = \sum_{i=1}^{N}\left\|\mathbf{x}^{(i)}-\tilde{\mathbf{x}}^{(i)}\right\|^2, \tag{3}$$

where N is the number of the training instances. The trained AE is denoted as,

$$\mathbf{x}_h{}^{(1)} = s(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)}), \tag{4}$$

with $\mathbf{W}^{(1)}, \tilde{\mathbf{W}}^{(1)}, \mathbf{b}^{(1)}$, and $\tilde{\mathbf{b}}^{(1)}$ indicating the optimal parameters, i.e.,

$$\{\mathbf{W}^{(1)},\tilde{\mathbf{W}}^{(1)},\mathbf{b}^{(1)},\tilde{\mathbf{b}}^{(1)}\} = \arg\min\frac{1}{N}L(\mathbf{x},s(\tilde{\mathbf{W}}\cdot s(\mathbf{W}\mathbf{x}+\mathbf{b})+\tilde{\mathbf{b}})). \tag{5}$$

Considering $\mathbf{x}_h{}^{(1)}$ as the input of another AE, the higher representations of the physiological feature abstractions $\mathbf{x}_h{}^{(n)}$ could be hierarchically computed by stacking multiple AEs to elicit a SAE network,

$$\mathbf{x}_h{}^{(n)} = s(\mathbf{W}^{(n)}\ldots s(\mathbf{W}^{(2)}s(\mathbf{W}^{(1)}\mathbf{x}+\mathbf{b}^{(1)})+\mathbf{b}^{(2)})\ldots+\mathbf{b}^{(n)}) = s^{(n)}_{\hat{\mathbf{W}},\hat{\mathbf{b}}}(\mathbf{x}). \tag{6}$$

The procedure for determining the SAE parameters layer by layer is called *pre-training*. By adding an additional two-neuron output layer that corresponds to binary emotions, a deep classifier is derived as,

$$\mathbf{y} = s(\mathbf{U}\mathbf{x}_h{}^{(n)}+\mu) = s(\mathbf{U}s^{(n)}_{\hat{\mathbf{W}},\hat{\mathbf{b}}}(\mathbf{x})+\mu) = s_{sae}(\mathbf{x}), \tag{7}$$

where $\mathbf{y}=[1\ \ 0]^T$ or $\mathbf{y}=[0\ \ 1]^T$ indicates the low or high levels of each emotion dimension, $\mathbf{U}$ and $\boldsymbol{\mu}$ are the weight matrix and bias vector of the decision layer, $s^{(n)}_{\hat{\mathbf{W}},\hat{\mathbf{b}}}(\mathbf{x})$ denotes the n-hidden-layer network with model parameters $\hat{\mathbf{W}}$ and $\hat{\mathbf{b}}$. Then, BP algorithm is used again for *fine-tuning* the pre-trained SAE parameters.

2.3.2. SAE-based ensemble deep learning model

Considering that the physiological features may share different hidden properties across different CNS and PNS modalities, we implement the SAE network on homogenous feature subset independently. The higher abstractions from multiple feature subsets could be separately elicited with flexible structures. Then, the feature abstractions are properly fused to achieve better classification performance. The network architecture of the MESAE is illustrated in Fig. 4 and the following three steps are required to build the emotion classifier.

1) Initialize member SAEs

Let the overall feature set be defined as ${F_0}^{D_0}$ with $D_0=425$, q non-overlapped physiological feature subsets are given as follows,

$$\begin{aligned}&{F_1}^{D_1},\quad {F_2}^{D_2},\cdots,{F_q}^{D_q}\\ &s.t.\quad {F_1}^{D_1}\cup{F_2}^{D_2}\cup\cdots\cup{F_q}^{D_q}={F_0}^{D_0},\\ &D_1+D_2+\cdots+D_q=D_0,\\ &\text{and}\quad F_i^{D_i}\cap F_j^{D_j}=\emptyset,\quad \text{for}\quad i,j\in\{1,2,\cdots q\}\end{aligned} \tag{8}$$

Regarding the heterogeneity of multi-channel EEG PSDs in different cortical areas, the power features of 32 channels could be grouped into three subsets, ${F_1}^{D_1}$, ${F_2}^{D_2}$ and ${F_3}^{D_3}$, indicating left (14 channels), central (4 channels) and right (14 channels) scalp, respectively. Another subset ${F_4}^{D_4}$ is built by EEG power differences. The EEG means and variances constructs subset ${F_5}^{D_5}$ while the features indicating EEG complexity degree (zero-crossing rate and approximate entropy) build subset ${F_6}^{D_6}$. The EOG and EMG features are combined as ${F_7}^{D_7}$ since both of them are highly correlated with the activities of facial muscles. In the end, ${F_8}^{D_8}$,${F_9}^{D_9}$,${F_{10}}^{D_{10}}$, and ${F_{11}}^{D_{11}}$ are built based on skin temperature, GSR, blood volume pressure, and respiration features, respectively.

Define feature vectors belonging to feature subset $F_i^{D_i}$ as,

$$\mathbf{x}(F_i^{D_i})\in F_i^{D_i}, i,j\in\{1,2,\cdots q\}, q=11. \tag{9}$$

We build q SAE-based deep models for describing the hidden feature abstractions of each physiological feature subset, i.e.,

$$\begin{cases}\mathbf{y}_1 = s(\mathbf{U}s^{(n_1)}_{\hat{\mathbf{W}}_1,\hat{\mathbf{b}}_1}(\mathbf{x}({F_1}^{D_1}))+\mu_1) = s^{(1)}_{sae}(\mathbf{x})\\ \mathbf{y}_2 = s(\mathbf{U}s^{(n_2)}_{\hat{\mathbf{W}}_2,\hat{\mathbf{b}}_2}(\mathbf{x}({F_2}^{D_2}))+\mu_2) = s^{(2)}_{sae}(\mathbf{x})\\ \ldots\\ \mathbf{y}_{11} = s(\mathbf{U}s^{(n_{11})}_{\hat{\mathbf{W}}_{11},\hat{\mathbf{b}}_{11}}(\mathbf{x}({F_{11}}^{D_{11}}))+\mu_{11}) = s^{(11)}_{sae}(\mathbf{x}),\end{cases} \tag{10}$$

where $s^{(n_1)}_{\hat{\mathbf{W}}_1,\hat{\mathbf{b}}_1}(\mathbf{x}({F_1}^{D_1}))$,$s^{(n_2)}_{\hat{\mathbf{W}}_2,\hat{\mathbf{b}}_2}(\mathbf{x}({F_2}^{D_2}))$, ..., and $s^{(n_{11})}_{\hat{\mathbf{W}}_{11},\hat{\mathbf{b}}_{11}}(\mathbf{x}({F_{11}}^{D_{11}}))$ denote the higher feature abstractions of each feature subset.

2) Model structure identification for member SAEs

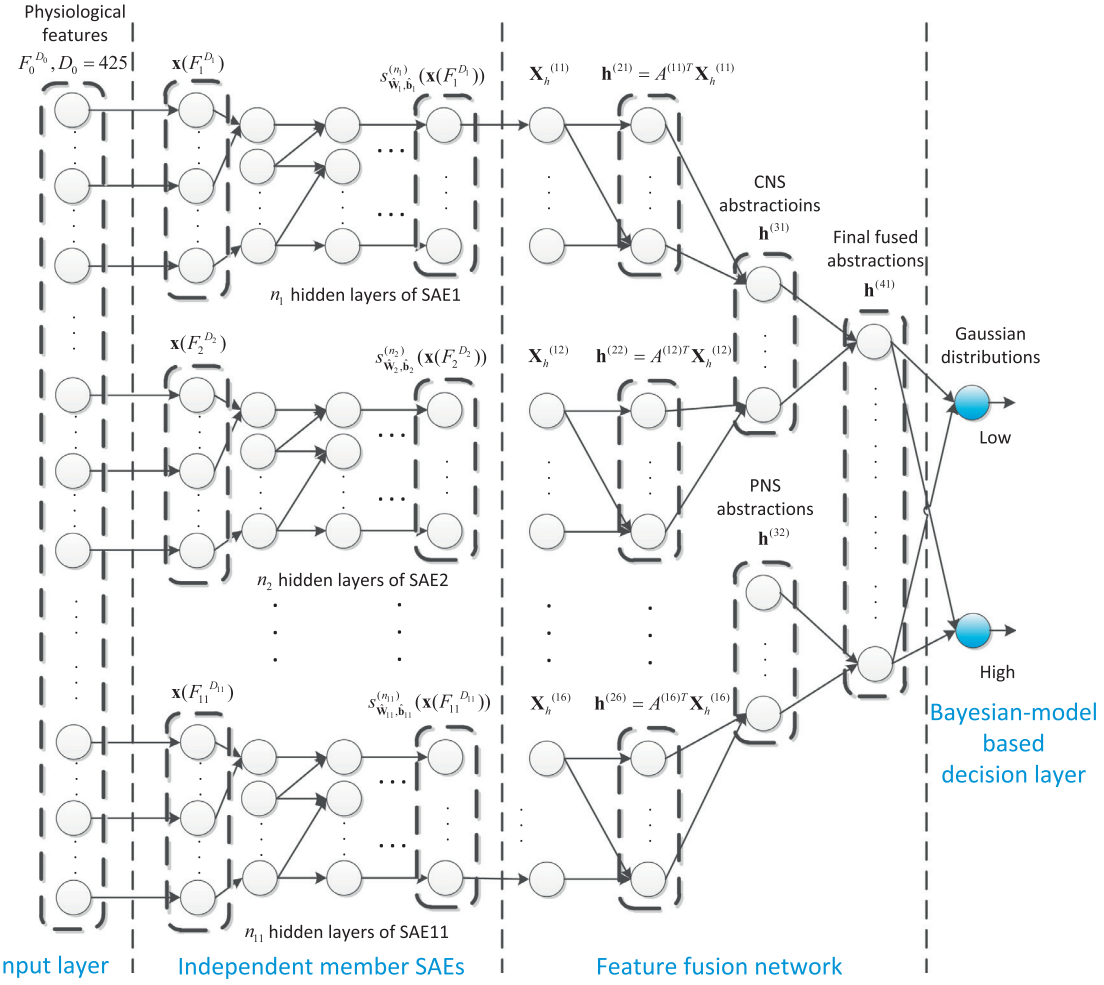

Fig. 4. Architecture of multiple-fusion-layer based ensemble classifier of SAEs, where $\mathbf{x}(F_i^{D_i})$ denotes the feature vector from the feature subset F_i with dimensionality D_i, $s^{(n_i)}_{\hat{\mathbf{W}}_i,\hat{\mathbf{b}}_i}(\mathbf{x}(F_i^{D_i}))$ denotes the corresponding feature abstraction elicited by member SAE $s^{(n_i)}_{\hat{\mathbf{W}}_i,\hat{\mathbf{b}}_i}$, $\mathbf{X}_h^{(1j)}$ denotes the combined SAE abstractions, and $\mathbf{h}^{(jk)}$ is the abstraction fusion at different levels.

Prior to learning network weights and bias, the structure (or hyper-parameters) of each SAE must be predetermined, i.e., the number of hidden layers and that of neurons in each hidden layer. Here, we proposed a model structural identification index based on the intrinsic dimensionality analysis to find a parsimonious emotion classifier. That is, the minimum hidden layers and the neurons would be selected with the minimal loss of the local geometric information from the input to abstracted data space. The structural loss function (SLF) accounting for the geometric information is defined as,

$$\sigma_R^2(\mathbf{D_X}, \mathbf{D_H}) = \lambda_1(1 - \rho_{1_{\mathbf{D_X}\mathbf{D_H}}}^2) + (1 - \lambda_1)(1 - \rho_{2_{\mathbf{D_X}\mathbf{D_H}}}^2), \tag{11}$$

where $\mathbf{D_X}$ and $\mathbf{D_H}$ denote the similarity matrices of inputs and outputs of the mapping, λ_1 is a weighting parameter, $\rho_{1_{\mathbf{D_X}\mathbf{D_H}}}^2$ and $\rho_{2_{\mathbf{D_X}\mathbf{D_H}}}^2$ are defined by the following equations.

$$\rho_{1_{\mathbf{D_X}\mathbf{D_H}}} = \frac{\sum_{k=1}^{N}(\mathbf{v}_{\mathbf{D_X}}(k) - \bar{v}_{\mathbf{D_X}})(\mathbf{v}_{\mathbf{D_H}}(k) - \bar{v}_{\mathbf{D_H}})}{\sqrt{\sum_{k=1}^{N}(\mathbf{v}_{\mathbf{D_X}}(k) - \bar{v}_{\mathbf{D_X}})^2}\sqrt{\sum_{k=1}^{N}(\mathbf{v}_{\mathbf{D_H}}(k) - \bar{v}_{\mathbf{D_H}})^2}}, \tag{12}$$

$$\rho_{2_{\mathbf{D_X}\mathbf{D_H}}} = \frac{\sum_{k=1}^{N}(\mathbf{a}_{\mathbf{D_X}}(k) - \bar{a}_{\mathbf{D_X}})(\mathbf{a}_{\mathbf{D_H}}(k) - \bar{a}_{\mathbf{D_H}})}{\sqrt{\sum_{k=1}^{N}(\mathbf{a}_{\mathbf{D_X}}(k) - \bar{a}_{\mathbf{D_X}})^2}\sqrt{\sum_{k=1}^{N}(\mathbf{a}_{\mathbf{D_H}}(k) - \bar{a}_{\mathbf{D_H}})^2}}. \tag{13}$$

In Eqs. (12), and (13), $\mathbf{D_X}$ is for the SAE inputs, where the element is the Euclidean distance between each two instances, while $\mathbf{D_H}$ is for the abstracted feature vectors from a SAE hidden layer. The values of $\rho_{1_{\mathbf{D_X}\mathbf{D_H}}}$ and $\rho_{2_{\mathbf{D_X}\mathbf{D_H}}}$ represent the Pearson product-moment correlation coefficient and Spearman's rank correlation coefficient. It is noted that $\mathbf{v}(k)$ and $\mathbf{a}(k)$ denote the k^{th} entry of the matrices under the two cases (Pearson or Spearman correlation coefficients) while $\bar{v}$ and $\bar{a}$ denote the mean values of the all entries. The two terms could separately evaluate linear or nonlinear correlation between input physiological features and feature abstractions. The computation of $\rho_{1_{\mathbf{D_X}\mathbf{D_H}}}$ or $\rho_{2_{\mathbf{D_X}\mathbf{D_H}}}$ is achieved by taking over all entries of $\mathbf{D_X}$ and $\mathbf{D_H}$ in the original order $\mathbf{v}$ or a ranked order $\mathbf{a}$. The smallest σ_R^2 suggests that the original and abstracted physiological features are highly correlated and possess similar local geometric structure. Since the prior knowledge of the weights between linear and nonlinear correlation coefficients is insufficient, we simply selected the value of λ_1 as 0.5. In such case, Eq. (11) can be alternatively formularized as $\sigma_R^2(\mathbf{D_X}, \mathbf{D_H}) = (1 - \rho_{1_{\mathbf{D_X}\mathbf{D_H}}}^2)(1 - \rho_{2_{\mathbf{D_X}\mathbf{D_H}}}^2)$, i.e., use the product instead of the sum. On the other hand, if there is a possibility that the linear correlation coefficient incorrectly reflects the mapping property of the model, a smaller value of λ_1 can be more suitable. The pseudo codes of the SAE structural identification and learning algorithms are listed in Table 2. We also use data augmentation to improve the training stability of each SAE. The augmented instances are generated by repeatedly superposing a standard Gaussian distributed noise (scaled by 0.1) upon the original physiological features.

Table 2
Pseudo codes for model structural identification and training member SAEs.

Start model structural identification
Set the training set S_{TR} for each member SAE
for $i = 1 : q$
$S_{TR}(:,:,i) = \{(\mathbf{x}_1(F_i^{D_i}), y_1), (\mathbf{x}_2(F_i^{D_i}), y_2), \ldots, (\mathbf{x}_N(F_i^{D_i}), y_N)\}$
$\mathbf{x}(F_i^{D_i}) \in R^{D_i}, S_{TR}(:,:,i) \in R^{N\times(D_i+1)}$
End for
for $i = 1 : q,\ k_{\min} = d_j$
for $j = 1 : n_q$
for $k = 1 : k_{\min}$
Initialize a SAE network $SAE_{(i,j,\mathbf{k}_{\min}(j)=k)}$
of j hidden layers and $\mathbf{k}_{\min} \in R^j$ neurons in these layers
Pre - train and fine - tune the $SAE_{(i,j,\mathbf{k}_{\min}(j)=k)}$ via $S_{TR}(:,:,i)$ cross - validation
Compute average σ_R^2 between $\mathbf{x}(F_i^{D_i})$ and $s^{(j)}_{\hat{\mathbf{W}}_i,\hat{\mathbf{b}}_i}(\mathbf{x}(F_i^{D_i}))$
End for
$k_{\min} = \arg\min(\sigma_R^2\ (k))$
$\mathbf{k}_{\min}(j) = k_{\min}$
End for
$\mathbf{k}_{\min}$ represents the numbers of neurons in $SAE_{(i,n_i,\mathbf{k}_{\min})}$
End for
End model structural identification

Start training member SAEs
Initialize q $SAE_{(i,n_i,\mathbf{k}_{\min})}$ with $\mathbf{k}_{\min} \in R^{n_i}$ neurons of each layer $(i = 1, 2, \cdots q)$
Start instance augementation for S_{TR}
for $z = 1 : 7, i = 1 : q,\ j = 1 : n_i, k = 1 : D_i$
$\tilde{S}_{TR}^{(z)}(j,k,i) = S_{TR}(j,k,i) \pm p_{st} \cdot rand(0,1)$, with $p_{st} = 0.1$
End for
Update $S_{TR} = S_{TR} \cup \tilde{S}_{TR}^{(1)} \cup \tilde{S}_{TR}^{(2)} \cup \cdots \cup \tilde{S}_{TR}^{(7)}$
End instance augemenation for S_{TR}
Add a top layer with two neurons and build $FNN_{(i,j,\mathbf{k}_{\min}(j)=k)}$
Pre - train all $SAE_{(i,n_i,\mathbf{k}_{\min})}$ and finely tune $FNN_{(i,n_i,\mathbf{k}_{\min})}$
End training member SAEs

Remark 1. The geometrical information (GI) of physiological data can be defined as the relative distance (or the similarity) between each two instances in the feature space of each modality. In this study, the Euclidean distance metric is employed. In practice, GI reflects the data distribution via the relative geometrical position of physiological instances in a high dimensional space. The GI is incorporated and measured for SAE structure identification in the following way. The similarity matrices of $\mathbf{D_X}$ and $\mathbf{D_H}$ of the whole 1280 instances from 32 participant are computed. The computation of $\mathbf{D_X}$ and $\mathbf{D_H}$ has quantified GI across original physiological features and their deep abstractions. The high correlation between $\mathbf{D_X}$ and $\mathbf{D_H}$ indicates GI of physiological data has been well preserved between multiple feed-forward mappings in the deep model. On the other hand, low correlation suggests the loss of physiological feature abstraction is severe and the potential overfitting caused by too complex model structure may arise. Another advantage for employing GI is its computation can be also achieved via unsupervised pre-training of SAEs. To conclude, the GI is measured to keep the useful information of the physiological features for the deep model selection.

3) Construct hierarchical feature fusion network

After the network structures from $s_{sae}^{(1)}(\mathbf{x})$ to $s_{sae}^{(11)}(\mathbf{x})$ are determined, we obtain the higher abstractions of the physiological feature $\mathbf{X}_h$,

$$\mathbf{X}_h = \left[\mathbf{x}_h{}^{(n_1)T} \quad \mathbf{x}_h{}^{(n_2)T} \quad \cdots \quad \mathbf{x}_h{}^{(n_{11})T}\right]^T. \tag{14}$$

where $\mathbf{x}_h{}^{(n_i)}$ denotes the coded feature abstractions from the member SAE $s^{(n_i)}_{\hat{\mathbf{W}}_i,\hat{\mathbf{b}}_i}$. Instead of feeding $\mathbf{X}_h$ directly to a decision layer, we use a hierarchical feature fusion network to improve the inter-class discrimination. Considering the multimodality existing in $\mathbf{X}_h$, six fusion sublayers are first built in parallel and the input of each sublayer is,

$$\begin{cases} \mathbf{X}_h{}^{(11)} = \left[\mathbf{x}_h{}^{(n_1)}T \quad \mathbf{x}_h{}^{(n_2)}T \quad \mathbf{x}_h{}^{(n_3)}T\right]^T \\ \mathbf{X}_h{}^{(12)} = \mathbf{x}_h{}^{(n_4)}T \\ \mathbf{X}_h{}^{(13)} = \left[\mathbf{x}_h{}^{(n_5)}T \quad \mathbf{x}_h{}^{(n_6)}T\right]^T \\ \mathbf{X}_h{}^{(14)} = \mathbf{x}_h{}^{(n_7)}T \\ \mathbf{X}_h{}^{(15)} = \left[\mathbf{x}_h{}^{(n_8)}T \quad \mathbf{x}_h{}^{(n_9)}T\right]^T \\ \mathbf{X}_h{}^{(16)} = \left[\mathbf{x}_h{}^{(n_{10})}T \quad \mathbf{x}_h{}^{(n_{11})}T\right]^T, \end{cases} \tag{15}$$

where $\mathbf{X}_h{}^{(11)}$ is the higher abstractions of EEG PSD features of all cortical areas, $\mathbf{X}_h{}^{(12)}$ represents that of EEG power differences, $\mathbf{X}_h{}^{(13)}$ is for EEG time-domain features, $\mathbf{X}_h{}^{(14)}$ indicates the abstraction from EOG and EMG, $\mathbf{X}_h{}^{(15)}$ is fused from skin temperature and resistance, and $\mathbf{X}_h{}^{(16)}$ presents the abstraction of the blood pressure and respiration.

The feature fusion approach is motivated by our previous work [20, 46] and a hierarchical version is further developed. Six adjacent graphs $G^{(11)}$, $G^{(12)}$, $G^{(13)}$, $G^{(14)}$, $G^{(15)}$, and $G^{(16)}$ are first built for N training instances in the form of $\mathbf{X}_h{}^{(11)}$, $\mathbf{X}_h{}^{(12)}$, $\mathbf{X}_h{}^{(13)}$, $\mathbf{X}_h{}^{(14)}$, $\mathbf{X}_h{}^{(15)}$, and $\mathbf{X}_h{}^{(16)}$, respectively. For each graph with N nodes, there is an edge for each two instances if they are "emotionally" close to each other. That is, all training instances belonging to the same emotional class share an edge, otherwise there is no edge. The weight matrices $E^{(11)}$,$E^{(12)}$,$E^{(13)}$,$E^{(14)}$,$E^{(15)}$, and $E^{(16)}$ are then constructed for all graphs where the element is $e_{ij}=1$ if two instances are connected, otherwise $e_{ij}=0$.

Then, the feature fusion mapping is derived by solving following generalized eigenvector problem, i.e.,

$$X_h^{(1k)}L^{(1k)}X_h^{(1k)T}\mathbf{c}^{(1k)} = \nu^{(1k)}X_h^{(1k)}P^{(1k)}X_h^{(1k)T}\mathbf{c}^{(1k)}, \quad \text{for } k = 1, 2, \cdots, 6. \tag{16}$$

where the ith column of the instance matrix $X_h^{(1k)}$ is the feature vector in the form of $\mathbf{X}_{h,\,i}{}^{(1k)}$, $P^{(1k)}$ denotes a diagonal matrix with

the diagonal element $P_j^{(1k)} = \sum_i E_{i,j}^{(1k)}$, and $L^{(1k)} = P^{(1k)} - E^{(1k)}$. After the eigenvalues are sorted as $\nu_0^{(1k)} < \nu_1^{(1k)} < \cdots < \nu_{d^{(1k)}-1}^{(1k)}$, the solutions of Eq. (15) are $\mathbf{c}_0^{(1k)}, \mathbf{c}_1^{(1k)}, \cdots, \mathbf{c}_{d^{(1k)}-1}^{(1k)}$. In Eq. (16), the Eigen Vector-Eigen Value mapping is motivated by the Linear Discriminating Analysis (LDA) to improve inter-class discriminating capacity and reduce the intra-class divergence of the abstractions. Simultaneously, the relative distances between instances of the same class can be preserved since the adjacent graph is used to initialize the weight matrix. The transformation matrix of feature mapping is derived as,

$$\begin{aligned} &\mathbf{h}_i^{(2k)} = A^{(1k)T}\mathbf{X}_{h,i}{}^{(1k)}, \\ &A^{(1k)T} = [\mathbf{c}_0^{(1k)}, \mathbf{c}_1^{(1k)}, \cdots, \mathbf{c}_{d^{(1k)}-1}^{(1k)}], \quad \text{for} \quad k = 1, 2, \cdots, 6. \end{aligned} \tag{17}$$

By using the above mapping, the closer (belonging to the same emotion class) physiological abstraction vectors $\mathbf{X}_{h,\,i}{}^{(1k)}$ are also geometrically closer to each other in their mapped representation $\mathbf{h}_i{}^{(2k)}$. That is, the local geometrical information of the fused feature abstractions is properly preserved via Eq. (17).

The feature fusion mapping above can be stacked into a hierarchical structure. We use another two fusion sublayers in parallel with the input of each sublayer as,

$$\begin{cases} \mathbf{X}_h{}^{(21)} = \left[\mathbf{h}_i{}^{(11)T} \quad \mathbf{h}_i{}^{(12)T} \quad \mathbf{h}_i{}^{(13)T}\right]^T \\ \mathbf{X}_h{}^{(22)} = \left[\mathbf{h}_i{}^{(14)T} \quad \mathbf{h}_i{}^{(15)T} \quad \mathbf{h}_i{}^{(16)T}\right]^T \end{cases}, \tag{18}$$

where $\mathbf{X}_h{}^{(21)}$ and $\mathbf{X}_h{}^{(22)}$ denote the feature abstractions of CNS and PNS, respectively. Accordingly, the following transformation matrices are elicited in the same way,

$$\mathbf{h}_i{}^{(3k)} = A^{(2k)T}\mathbf{X}_{h,i}{}^{(2k)}, \quad \text{for} \quad k = 1, 2. \tag{19}$$

The final fused feature vector $\mathbf{X}_h{}^{(31)}$ is computed by the last fusion layer as,

$$\mathbf{X}_{h,i}{}^{(41)} = A^{(41)T}\left[\mathbf{h}_i{}^{(31)T} \quad \mathbf{h}_i{}^{(32)T}\right]^T. \tag{20}$$

In the end, the decision layer of the deep learning network is based on a Bayesian model to find underlying Gaussian distributions $P(\mathbf{X}_{h,\,i}{}^{(41)}|\mathbf{y})$ for the final physiological feature abstractions $\mathbf{X}_h{}^{(41)}$ with the emotion class $\mathbf{y}$, i.e.,

$$\tilde{\mathbf{y}} = \arg\max_y P(\mathbf{X}_{h,i}{}^{(41)}|\mathbf{y}) = \arg\max_y \left[\varepsilon \prod_1^N P(\mathbf{y}|\mathbf{X}_{h,i}{}^{(41)})P(\mathbf{y})\right] \tag{21}$$

where ε is a normalization constant and $\tilde{\mathbf{y}}$ is the predicted emotional state.

3. Results

3.1. Model structure of member SAEs

The scheme of the data split for model structure identification, training and testing under two cases of the feature extraction are defined as follows.

For Case 1, the classifier structure is determined based on the two-fold cross-validation of all 1280 instances from 32 participants in a participant-generic manner due to the limited sample size. On the other hand, the classifier is trained and tested via 10-fold cross validation technique with a participant-specific style based on the same dataset. That is, for each participant, his/her physiological data (40 instances from all 1280 ones) are divided into 10 subsets with 9 for training (36 instances) and the remaining 1 (4 instances) for testing. Such procedure will repeat 10 times until all subsets are tested.

Table 3
Model structural identification results for each member SAE.

	Case 1			*Case 2*		
	$\mathbf{k}_{\min}(1)$	$\mathbf{k}_{\min}(2)$	$\mathbf{k}_{\min}(3)$	$\mathbf{k}_{\min}(1)$	$\mathbf{k}_{\min}(2)$	$\mathbf{k}_{\min}(3)$
$s_{sae}^{(1)}$	47	31	30	32	28	22
$s_{sae}^{(2)}$	20	16	14	18	14	14
$s_{sae}^{(3)}$	69	55	46	68	51	39
$s_{sae}^{(4)}$	56	55	51	56	45	38
$s_{sae}^{(5)}$	54	22	10	54	34	14
$s_{sae}^{(6)}$	64	39	27	64	43	24
$s_{sae}^{(7)}$	20	20	19	19	16	6
$s_{sae}^{(8)}$	6	3	2	6	5	3
$s_{sae}^{(9)}$	23	15	15	23	17	12
$s_{sae}^{(10)}$	7	5	4	6	3	3
$s_{sae}^{(11)}$	14	14	9	19	16	10
Sum	380	275	227	365	272	185

Note: $s_{sae}^{(i)}$ represents the i-th member SAE, $\mathbf{k}_{\min}(j)$ denotes the number of hidden neurons in the j-th hidden layer.

For Case 2, the classifier structure is determined based on the two-fold cross-validation of 3200 (i.e., 25%) instances from 32 participants (12,800 in total). The classifier is trained and tested via the participant-specific 10-fold cross validation based on the remaining 9600 (i.e., 75%) instances (300 for each participant). The large size of the instances in Case 2 facilitate the data used for model structure identification, training and testing to be non-overlapped.

The hyper parameters of the proposed MESAE classifier include the learning rate L_{le}, batch size B_{bs}, number of hidden layers n_q and the number of hidden neurons $\mathbf{k}_{\min}$ in each hidden layer. The values of L_{le} is selected as 1 according to the work from Liu et al. [43] while $B_{bs} = 10$ is adopted due to the small size of the training set under Case 1. In particular, we use the objective function σ_R^2 defined Eq. (11) and the cross-validation technique to optimize n_q and $\mathbf{k}_{\min}$. More specifically, we fix $n_q = 1$ and increase $\mathbf{k}_{\min}(n_q)$ from 1 to d_j according to the algorithm presented in Table 2. Finally, the average $\sigma_R^2(\mathbf{k}_{\min}(n_q))$ is computed, which explores the cases of different number of neurons with one hidden layer. Similarly, multiple hidden layer cases are investigated with $n_q > 1$ and the search range of $\mathbf{k}_{\min}(n_q)$ becomes $[1, \mathbf{k}_{\min}(n_q - 1)]$. The optimal n_q and $\mathbf{k}_{\min}$ are determined by exhausting all cases of $\sigma_R^2(\mathbf{k}_{\min}(n_q))$ values.

The detail results of three member SAEs for left, central, and right scalp EEG-PSD features of Case 2 are shown in Fig. 5. Taken Fig. 5(a) as an example, a SAE with a single hidden layer for 70 PSD features (i.e., 70 input neurons) of left scalp is first initialized while the SLF for different number of hidden neurons (from 1 to 70) is computed according to the algorithm in Table 2. In particular, the value of SLF first decreases with the increase of hidden neurons and start to converge after employing more than 20 neurons. The line plot shows that the minimum of SLF is achieved at $\mathbf{k}_{\min}(1) = 32$ hidden neurons. That is, the optimal number of neurons in hidden layer 1 for the SAE of left scalp EEG-PSD is 32. For hidden layer 2, $\mathbf{k}_{\min}(2) = 28$ is elicited in the same way with $\mathbf{k}_{\min}(1) = 32$ fixed while $\mathbf{k}_{\min}(3) = 22$ is derived with both $\mathbf{k}_{\min}(1) = 32$ and $\mathbf{k}_{\min}(2) = 28$ fixed. For Fig. 5(b) and (c), the values of $\mathbf{k}_{\min}$ are calculated in the same manner.

On the other hand, when using three hidden layers, the variation of SLF becomes unstable. It suggests the training stability of the deep model is undermined because of the overfitting induced by too deep model structure. The similar results are also found for the remaining eight SAEs for both of the two cases. Therefore, the final hidden layer number is selected as three for all member SAEs. The numbers of hidden neuron of each SAE based on two cases are listed in Table 3. The total numbers of feature abstractions of

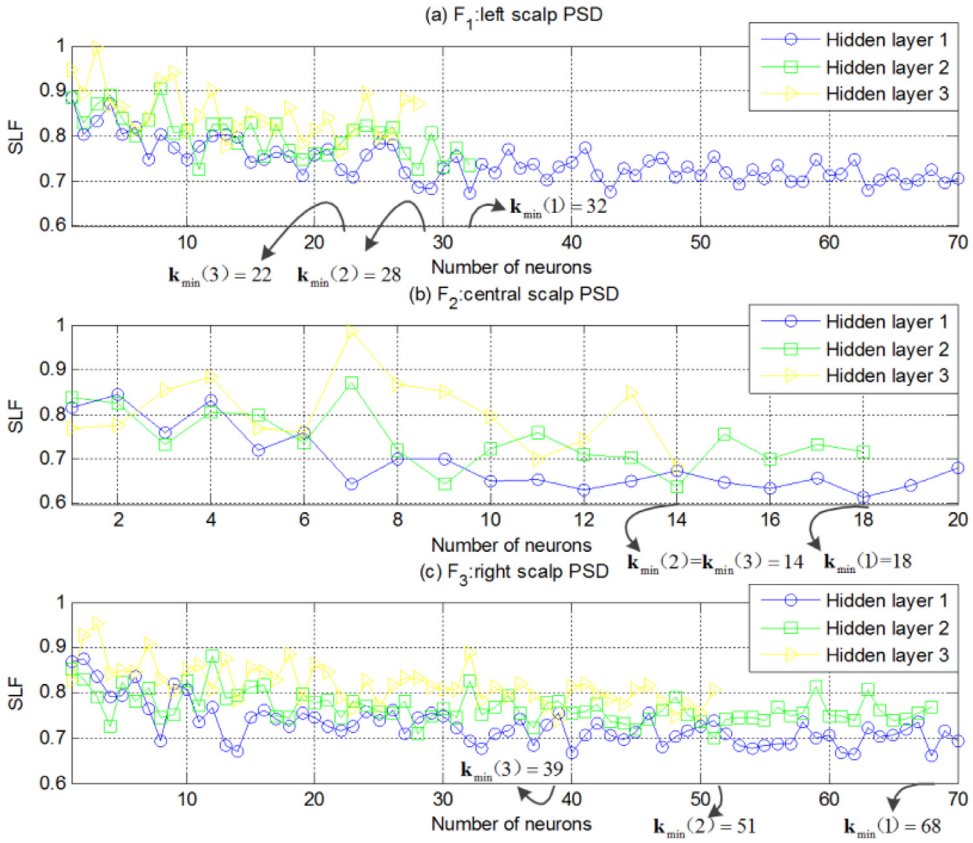

Fig. 5. Change of SLF vs. different number of neurons for each hidden layer. Results of three member SAEs for (a) left, (b) central, and (c) right scalp EEG-PSD features of Case 2 are shown.

the last SAE hidden for the two cases are reduced to 227 and 185, respectively. It indicates the intrinsic dimensionality of the multimodal feature sets is much smaller than 425. The difference between the model identification results are caused by different computation and spilt schemes of the physiological features.

3.2. Comparison of classification results against existing emotion classifiers

After the structural parameters for all member SAEs have been optimized, the generalizability of the MESAE is investigated and compared with several existing emotion classifiers of shallow structure. The metrics for classification performance are defined as follows. The classification accuracy for low arousal or valence level (denoted as *sensitivity* P_{sen}) is derived by,

$$P_{sen} = n_{TP}/(n_{TP} + n_{FN}), \tag{22}$$

where n_{TP} is the number of correctively predicted instances belonging to the low class and n_{FN} is the number of misclassified low-class instances. Moreover, the precision for recognizing the low-class instances is defined as *precision* P_{pre},

$$P_{pre} = n_{TP}/(n_{TP} + n_{FP}), \tag{23}$$

with n_{FP} denoting the number of misclassified high-level instances. The overall classification accuracy (P_{acc}) is computed by,

$$P_{acc} = (n_{TN} + n_{TP})/(n_{TN} + n_{FN} + n_{TP} + n_{FP}). \tag{24}$$

with n_{TN} denoting the number of correctively predicted high-level instances. Considering the instance imbalance, *F*1-score (denoted as P_f) of low emotion class is also employed,

$$P_f = 2P_{pre}P_{sen}/(P_{pre} + P_{sen}). \tag{25}$$

Finally, P_{acc} and P_f are used to evaluate the classification performance.

The classification performance of the MESAE is first compared against seven shallow classifiers, i.e., KNN, LR, LSSVM, NB, PCA-NB, LE-NB, and NPE-NB. KNN denotes the *K*-nearest neighbor classifier with the selected parameter of *K* from a candidate set {1, 2,..., 30}, LR represents the logistic regression method, and LSSVM denotes the least square support vector machine using the linear kernel. We found that using nonlinear kernels such as radial basis function achieves very low P_f so that those kernels are excluded from the comparison. The regularization parameter of LSSVM is selected by using grid search in a candidate set $\{2^{-4}, 2^{-5}, ..., 2^{-10}\}$. Moreover, the naive Bayesian classifier (denoted as NB) and its hierarchical variants are investigated. For the latter, PCA-NB denotes the physiological features are processed via PCA. LE-NB and NPE-NB represent the physiological features are processed by two nonlinear dimensionality reduction techniques, i.e., Laplacian eigenmaps [51] and neighbor preserving embedding [52]. The hyper-parameters of KNN, PCA-NB, LE-NB, and NPE-NB, i.e., the number of the nearest neighbors *K*, the percentage of the total variance for selecting the principal components, the number of the output dimensionalities of LE and NPE, are properly selected via two-fold cross validation from candidate sets {1, 2,..., 30}, {0.01, 0.02,..., 0.99}, and {1, 2,..., 425}, respectively. The datasets and the cross-validation schemes used for hyper-parameter determination and performance evaluation for all classifiers under two cases are as same as those for MESAE defined in Section 3.1.

The classification results of arousal dimension of Case 2 are illustrated in Fig. 6(a) and (b) in terms of P_{acc} and P_f, respectively. MESAE achieves the highest P_{acc} and P_f values for most of the participants. For other classifiers, LR elicits the poorest performance. The performances of KNN, LSSVM, NB, PCA-NB, LE-NB, and NPE-

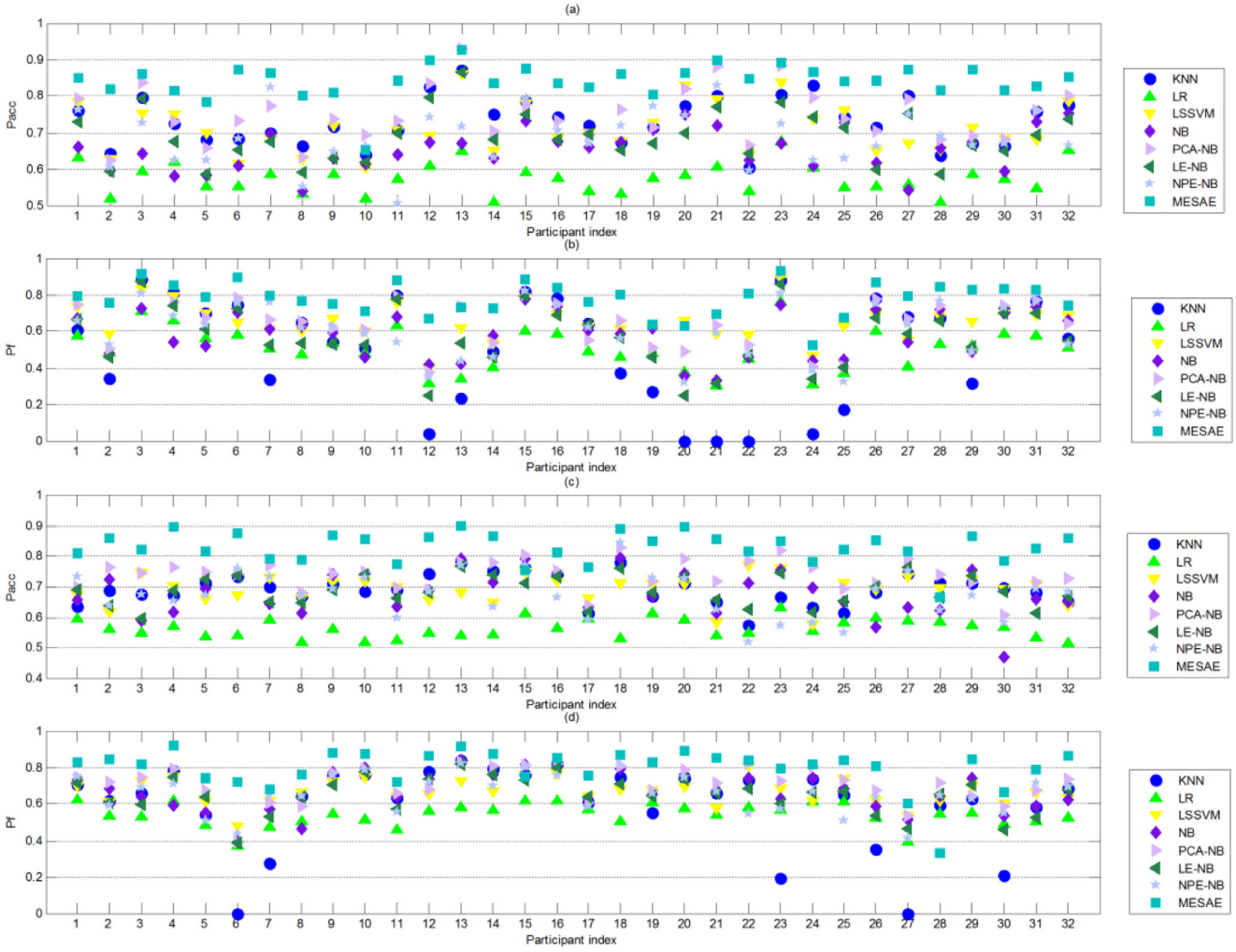

Fig. 6. Emotion classification performance of Case 2. (a) and (b): classification accuracy and *F*1-score of arousal dimension; (c) and (d): classification accuracy and *F*1-score of valence dimension.

NB, are comparable. A remarkable fact is that participant 12, 20, 21, 22, and 24, achieve very low P_f for KNN while the corresponding P_{acc} values are all above 0.5. The classification results of valence scale under Case 2 are shown in Fig. 6(c) and (d). The MESAE also achieves the best P_{acc} and P_f values for most participants.

The detailed classification performance comparisons on arousal and valence dimensions under Case 1 and 2 are illustrated by boxplots in Figs. 7 and 8, respectively. Each column data represents the results of 32 participants. Each classification metric of MESAE is compared with that of other seven classifiers via two tailed pair *t*-test. The statistical comparison is performed repeatedly across seven combinations, i.e., MESAE vs. KNN, vs. LSSVM, vs. NB, vs. PCA-NB, vs. LE-NB, and vs. NPE-NB. The significant improvements are found for all comparisons in Case 1 with $p < 0.001$ and most comparisons in Case 2 with $p < 0.05$.

To further validate the effectiveness of the proposed method, the performance comparison of the MESAE against SAE, DBN, and their variants is carried out since SAE and DBN were investigated by several recent published works on the physiological data of the DEAP database [40–45]. In total, four additional deep classifiers are constructed. For Case 1, the first two classifiers are the DBN and the SAE based deep networks with the same settings in [41] and [45], respectively. As suggested by Jirayucharoensak et al., a PCA layer has been added and linked to the input layer of the deep models [44]. Accordingly, two additional classifier variants denoted by PCA-SAE and PCA-DBN are generated by the same settings in [44]. For Case 2, the number of the hidden neurons in each layer of DBN, SAE, PCA-SAE and PCA-DBN is reselected via a candidate set of {1, 2,..., 425} due to the large size of the available instances.

The classification results of Case 1 are shown in Fig. 9((a)-(d)) via box-whisker plots. For arousal dimension, the MESAE achieves the highest medians of P_{acc} and P_f, respectively. For the former, the paired *t*-test results suggest that the MESAE significantly outperforms SAE, DBN, PCA-SAE, and PCA-DBN with $t = 10.6$ $p < 0.001$, $t = 7.3$ $p < 0.001$, $t = 10.0$ $p < 0.001$, and $t = 6.05$ $p < 0.001$, respectively. The significant superiority is seen for P_f with $t = 7.8$ $p < 0.001$, $t = 9.3$ $p < 0.001$, $t = 7.5$ $p < 0.001$, and $t = 9.4$ $p < 0.001$, respectively. For the remaining four deep classifiers, PCA-SAE achieves the highest performance. Although PCA-DBN possesses high P_{acc} and P_{spe} values, the corresponding P_f is very low. For valence dimension, the MESAE also achieves the highest medians of all classification metrics. In particular, the significant superiority is observed for P_{acc} and P_f with $p < 0.001$ according to the paired *t*-test. Fig. 9((e)-(h)) illustrates the classification performance of all deep models under Case 2. Similarly, the MESAE achieves the highest values of P_{acc} and P_f with significant improvements against other four deep classifiers for both of the arousal and valence dimensions ($p < 0.001$). In particular, the classification performance elicited under Case 2 is much higher than that of Case 1 due to the large size of training instances.

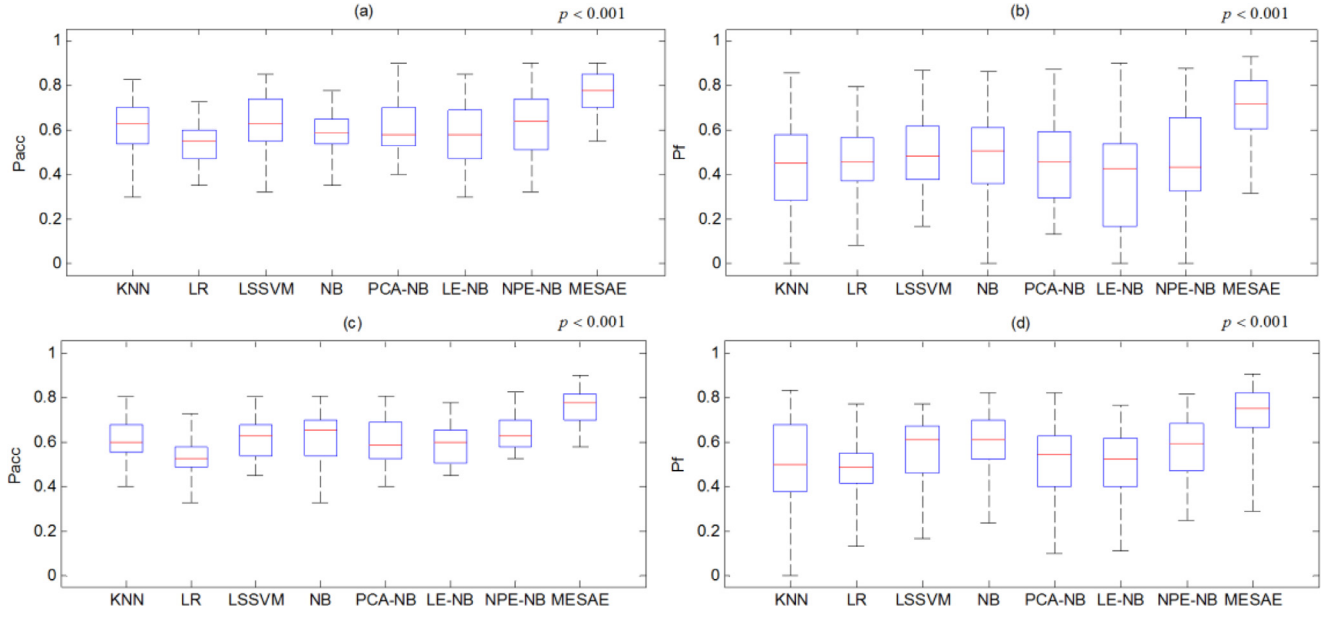

Fig. 7. Box-whisker plots for emotion classification results of arousal ((a)-(b)) and valence ((c)-(d)) dimension under Case 1, $p < 0.001$ denotes the degree of significance regarding the improvement of the classification performance between MESAE and other classifiers via paired *t*-test.

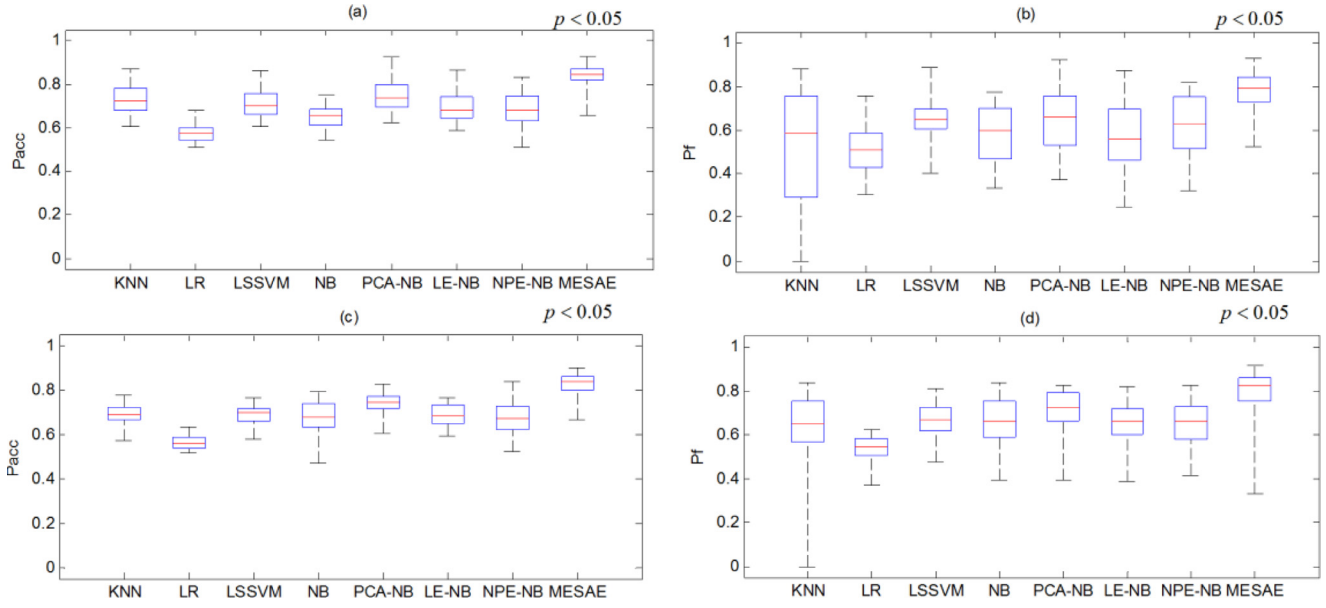

Fig. 8. Box-whisker plots for emotion classification results of arousal ((a)-(b)) and valence ((c)-(d)) dimension under Case 2, $p < 0.05$ denotes the degree of significance regarding the improvement of the classification performance between MESAE and other classifiers via paired *t*-test.

Table 4 compares the participant-average classification performance between the MESAE classifier under Case 1 and several reported studies on the same database with the same size of the available instances. For all studies in the reported works, the method for labeling the ground truth of the physiological features is the same, i.e., a threshold of 5 is employed for labeling the low and high arousal/valence classes. In particular, except for Atkinson and Campos [32], the same subject-independent 10-fold cross-validation is used. In [32], the subject-independent cross validation was performed via 8 folds. The CNS, PNS, and multimedia content analysis (MCA) features are analyzed by Koelstra et al. [48]. The highest P_f value is achieved by the fusion of CNS and MCA features for arousal classification (0.6310) and the fusion of PNS and MCA features for valence classification (0.6520). Liu and Olga [53], Naser and Saha [54], Chen et al. [55] as well as Yoon and Chung [56] have developed threshold-based detection algorithm, dual-tree complex wavelet packet transformation method, C4.5 decision tree algorithm and Bayesian weighted-log-posterior classifier based on the EEG data of DEAP, respectively. Atkinson and Campos [32] combined mutual information minimization based feature extraction algorithm and SVM classifier. The results of two DBN variants from Li et al. and Wang and Shang are also included. From the figure, we found that the MESAE achieves the highest values for both P_{acc} and P_f compared to the reported studies.

In summary, the performance of the MESAE emotion classifier regarding the classification accuracy and the F-score is superior to several state-of-the art shallow classifiers as well as the deep classifiers with the prior knowledge based structure. In particular, the large sample size (Case 2) elicits higher classification performance for both of the arousal and valence dimensions. The above observa-

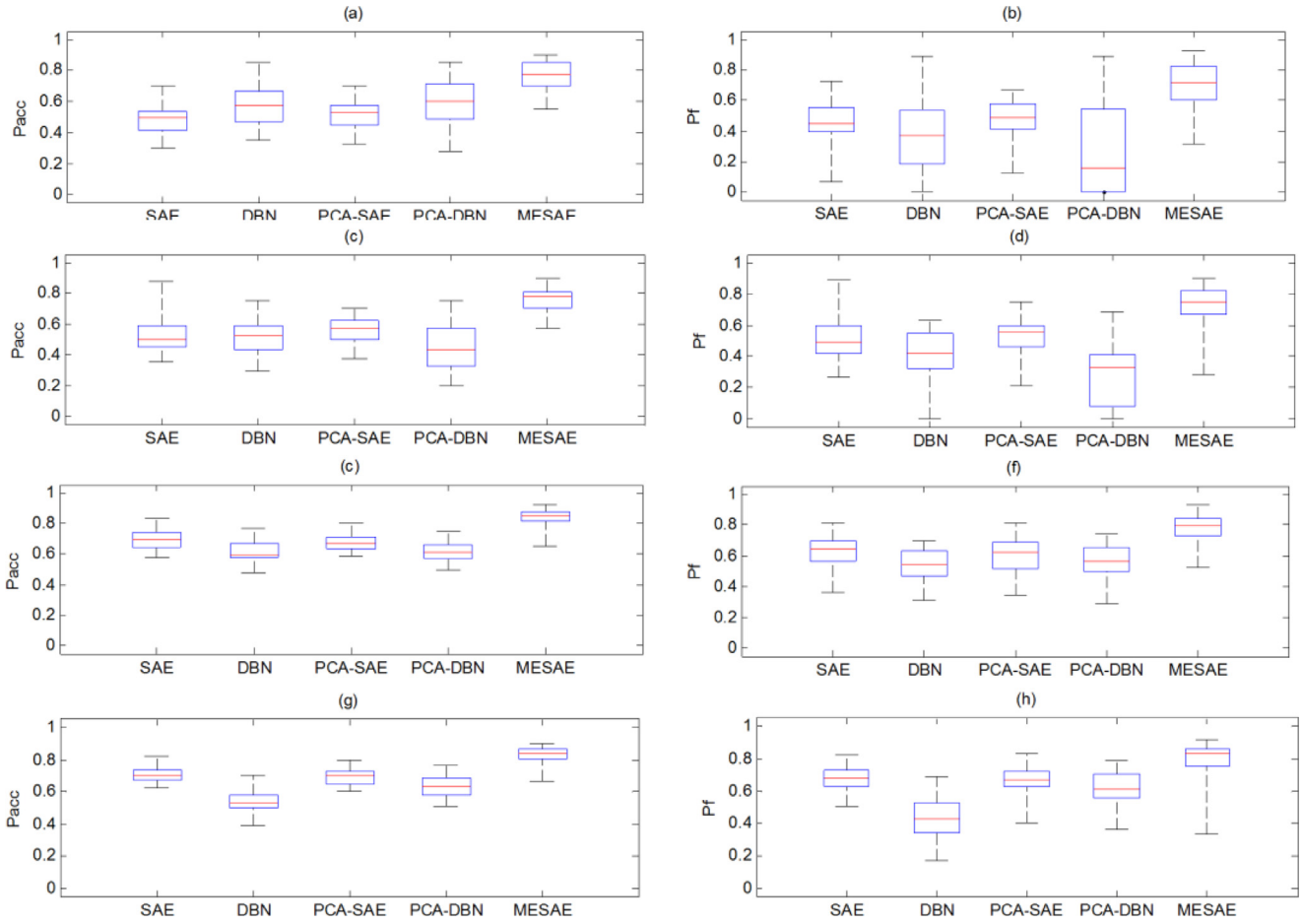

Fig. 9. Box-whisker plots for emotion classification results of arousal (Case 1: (a)-(b), Case 2: (e)-(f)) and valence dimensions (Case 1: (c)-(d), Case 2: (g)-(h)) using deep classifiers.

Table 4
Participant-average classification performance comparison between MESAE and several reported studies using shallow or deep classifiers.

	Arousal		*Valence*	
	P_{acc}	P_f	P_{acc}	P_f
CNS feature based single modality (Koelstra et al. [48])	0.6200	0.5830	0.5760	0.5630
PNS feature based single modality (Koelstra et al. [48])	0.5700	0.5330	0.6270	0.6080
CNS and MCA feature based arousal classification, PNS and MCA feature based valence classification (Koelstra et al. [48])	–	0.6310	–	0.6520
CNS, PNS and MCA feature based three modalities (Koelstra et al. [48])	–	0.6180	–	0.6470
Liu and Sourina [53]	0.7651	–	0.5080	–
Naser and Saha [54]	0.6620	–	0.6430	–
Chen et al. [55]	0.6909	0.6896	0.6789	0.6783
Atkinson and Campos [32]	0.7306	–	0.7314	–
Yoon and Chung [56]	0.7010	–	0.7090	–
Li et al. [41]	0.6420		0.5840	
Wang and Shang [40]	0.5120		0.6090	
MESAE	**0.7719**	**0.6901**	**0.7617**	**0.7243**

tions indicate the deep model can benefit from sufficient training and validating instances as well as the proper hyper-parameters of the network structure. Moreover, the MESAE is also competitive against several existing emotion classifiers reported in the literature. By comparing with the best method, the mean of classification rate and F-score of MESAE improves by 5.26%.

3.3. Hierarchical abstractions of physiological features

The six hidden layers of MESAE (three hidden layers in each SAE and three fusion layers) lead to a non-transparent classification model. To give an insight of how such ensemble network process physiological features, the activations (i.e., the output values) for part of the neurons in each hidden layer are visualized based on the data from participant 1 under Case 2 as an example. Moreover, the discrimination capability of the activations from the last hidden layer of each SAE and the first two fusion layers are evaluated on the Bayesian model. The motivation of the analysis is to examine whether the generalization capability can benefit from the hierarchical deep structure or not.

In Fig. 10, the physiological feature abstractions of participant 1 under Case 2 from three representative neurons in each hidden layer of $s_{sae}^{(1)}(\mathbf{x})$ and fusion layers of the arousal classifier are shown based on six 3-D scatter plots. In Fig. 10(a) of the 1st hidden layer, the instances of both low and high arousal classes are

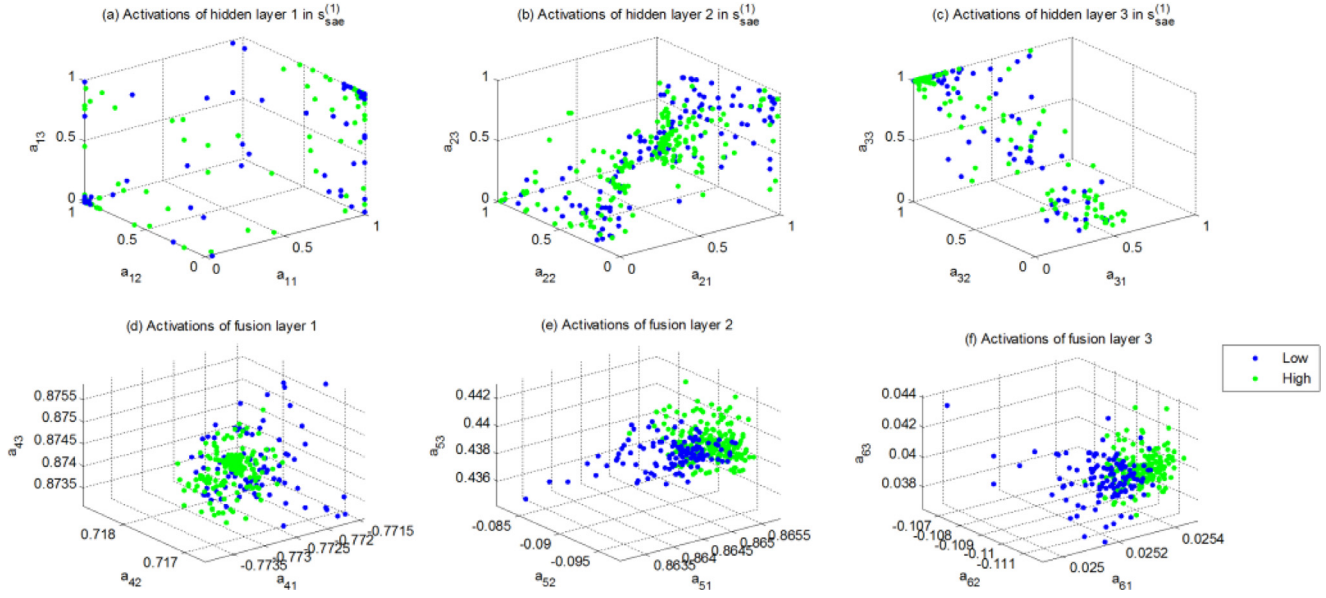

Fig. 10. 3-D scatter plots of the neuron activations of each hidden layer in MESAE for arousal classification for participant 1 under Case 2. The notation of a_{ij} indicates the j^{th} neuron activation in the i^{th} layer of MESAE.

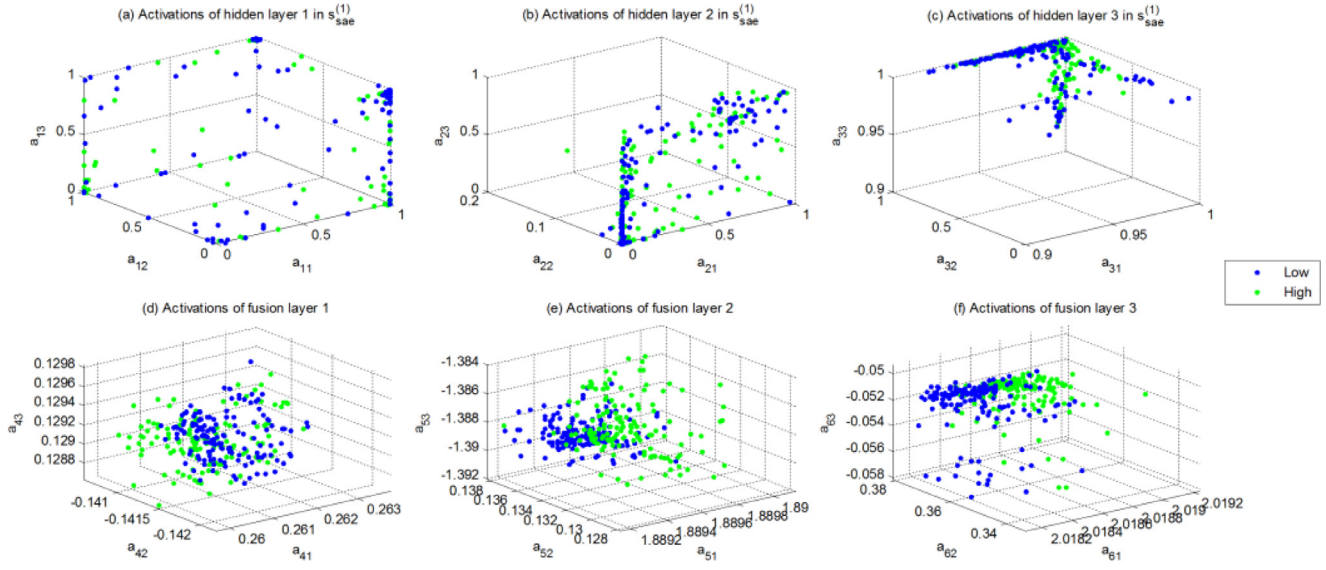

Fig. 11. 3-D scatter plots of the neuron activations of each hidden layer in MESAE for valence classification for participant 1 under Case 2.

distributed on the edges. In Fig. 10(b), the feature abstractions become more concentrated. In Fig. 10(c), all low-arousal instances lie in the center and are rounded by high class instances. When the higher abstractions of the EEG power features are further processed by the 1st fusion layer (Fig. 10(d)), the discrimination is obviously increased. In Fig. 10(e) of the 2nd fusion layer, all EEG features are employed and two distinct clusters are generated with slightly overlapping. From the final fusion layer (Fig. 10(f)), the CNS and PNS features are merged into two clear clusters. In Fig. 11, MESAE has been trained and tested for valence classification. The same neurons are selected for generating the scatter plots. The derived activation values of each layer are different from those in Fig. 10. The reason is that the fine-tuning and the construction of the adjacent graph are based on different training targets. Similar to Fig. 10, the inter-class discrimination has been successfully improved via three hidden layers and three feature fusion layers.

The arousal and valence classification performance derived by specific type of physiological feature abstractions under Case 1 is illustrated in Fig. 12 based on participant-specific 10-fold cross validation. In total, 19 cases are examined. The first 11 columns of each subfigure show the P_{acc} and P_f for all participants based on the final abstractions of 11 member SAEs, respectively. In Fig. 12(a) and (b), the highest medians of P_{acc} and P_f values are obtained by $\mathbf{x}_h^{(n_4)}$, i.e., the abstractions of the differences of EEG powers, and it implies $\mathbf{x}_h^{(n_4)}$ is the most important feature abstractions regarding discriminating binary arousal levels. For valence classification, the highest median of P_{acc} in abstractions from member SAEs is also achieved by $\mathbf{x}_h^{(n_4)}$. However, the highest median of P_f is elicited by the abstractions of means and variances of EEG signals. The classification performance of feature abstractions from six sublayers in the 1st, 2nd, and 3rd fusion layer are also compared in the last eight columns of each subfigure. The P_{acc} and P_f of CNS feature abstractions ($\mathbf{h}_{31}$) achieve the highest performance and are significantly higher than those of PNS ($\mathbf{h}_{32}$) for arousal dimension

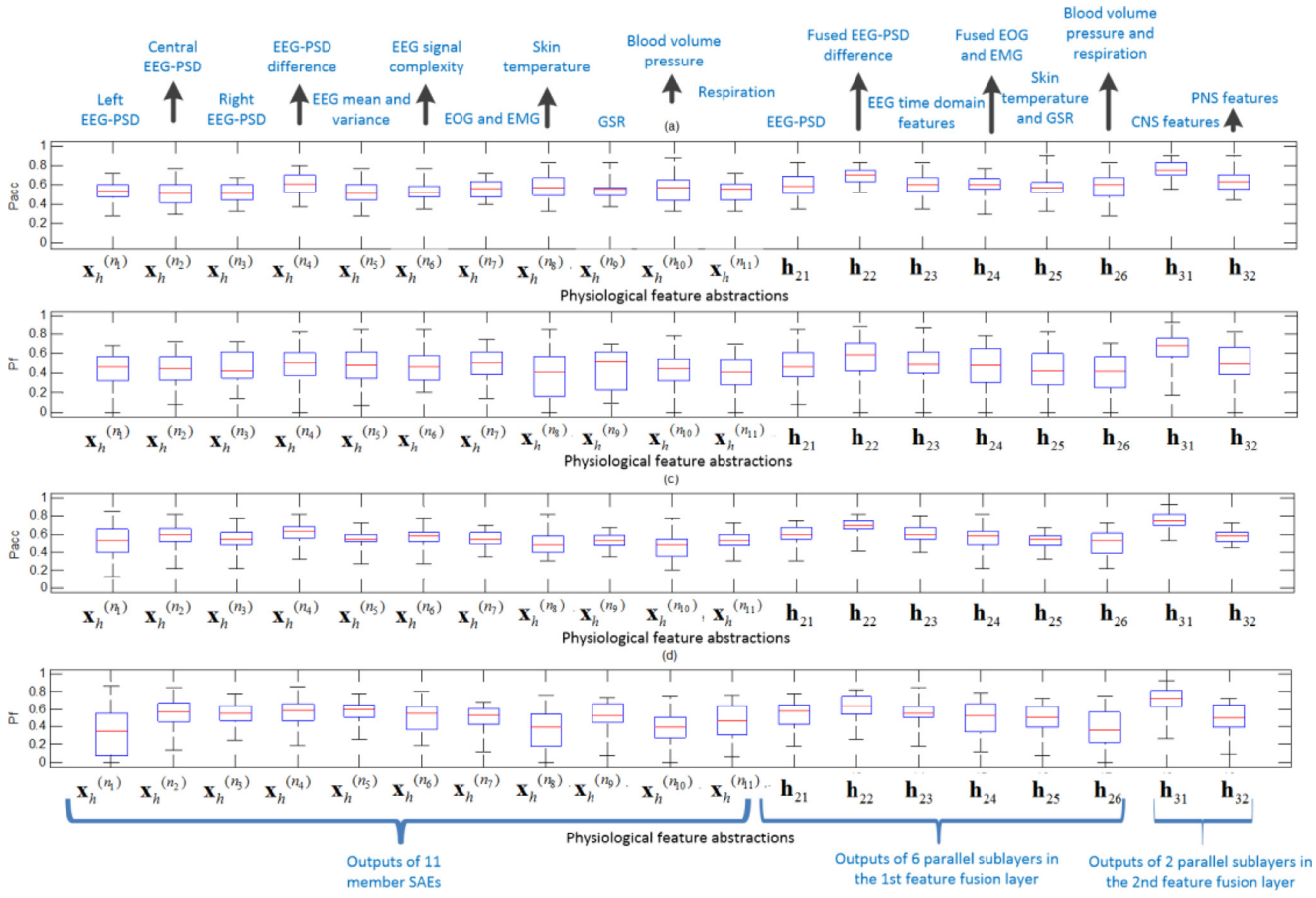

Fig. 12. Box-whisker plots for arousal ((a)-(b)) and valence (c)-(d)) classification performance using specific type of physiological feature abstractions under Case 1.

($t = 8.9$, $p < 0.001$, $t = 4.7$, $p < 0.001$) and valence dimension ($t = 9.1$, $p < 0.001$, $t = 6.6$, $p < 0.001$), respectively.

The feature-abstraction based classification results under Case 2 are shown in Fig. 13. For the arousal and valence dimensions, the highest P_{acc} medians computed by the abstractions from member SAEs are achieved by $\mathbf{x}_h^{(n_5)}$ (the values of EEG mean and variance in each segment) and $\mathbf{x}_h^{(n_1)}$ (left EEG-PSD), respectively. It indicates the large size of the training data improve the discriminating capability for specific physiological abstractions. The similar observation can be also found for $\mathbf{h}_{21}$ (the fusion of EEG-PSD abstractions) that achieves the highest P_{acc} and P_f values among $\{\mathbf{h}_{21},\mathbf{h}_{22},\ldots,\mathbf{h}_{26}\}$. For all 19 abstraction, the values of P_{acc} and P_f for CNS feature abstractions ($\mathbf{h}_{31}$) also achieves the highest performance. The significant superiority against PNS abstractions ($\mathbf{h}_{32}$) can be found for both of the arousal ($t = 11.8$, $p < 0.001$, $t = 7.4$, $p < 0.001$) and valence ($t = 11.2$, $p < 0.001$, $t = 9.9$, $p < 0.001$) dimensions.

Since the deep learning classifier is considered as a non-transparent model, we try to visualize the representative neuron activations (i.e., feature abstractions) for the testing data in each hidden layer of the MESAE in Figs. 10 and 11. The hidden states in the deep model can be thus unfolded to show how the method processes the physiological features in a hierarchical way. By observing the two figures, the clear discrimination between the low and high arousal or valence classes is found with the abstractions from the deeper layers. It indicates the MESAE with the selected parsimonious structure for fusing multiple modalities could filter the unwanted noise in shallow layers and elicits useful emotion indicators in deeper layers. The classification performance presented in Figs. 12 and 13 quantifies the discriminating capacity of the physiological abstractions at different levels. Since the highest performance is achieved by the deepest abstractions, the selected deep structure of the MESAE is validated to be suitable for each participant.

3.4. Classification performance of MESAE using different ensemble schemes

Let us denote the MESAE of three fusion layers presented in Sect. 3.2 as scheme 1. In this section, we investigate the classification performance of MESAE with a single fusion layer and denote such classifier as scheme 2.

For MESAE scheme 2, 11 member SAEs and a feature fusion layer with six sublayers are constructed. The data split and cross-validation schemes for hyper-parameter selection, training and testing follow the same way defined in Sect. 3.1 and 3.2. Then, each fusion sublayer is linked to a Bayesian model and elicits six member classifiers. They are divided into CNS and PNS groups with each containing three classifiers (corresponding to $\{\mathbf{h}_{21}, \mathbf{h}_{22}, \mathbf{h}_{23}\}$ and $\{\mathbf{h}_{24}, \mathbf{h}_{25}, \mathbf{h}_{26}\}$) and the ensemble classification committee is built by selecting the best T_1 and T_2 classifiers. The majority voting is used again to obtain final classifier predictions, i.e.,

$$\tilde{y} = \arg\max \sum_{j=1}^{T_1+T_2} \nu_j^c(\mathbf{x}(F_0^{D_0})), \nu \in \{0,1\}, c \in \{1,2\} \tag{28}$$

where $\nu_j^c = 1$ indicates that the member classifier j classifies $\mathbf{x}(F_0^{D_0})$ as class c and $\nu_j^c = 0$ denotes the opposite case. The feature extraction and data split approaches under Case 1 and 2 are separately applied to the MESAE scheme 2.

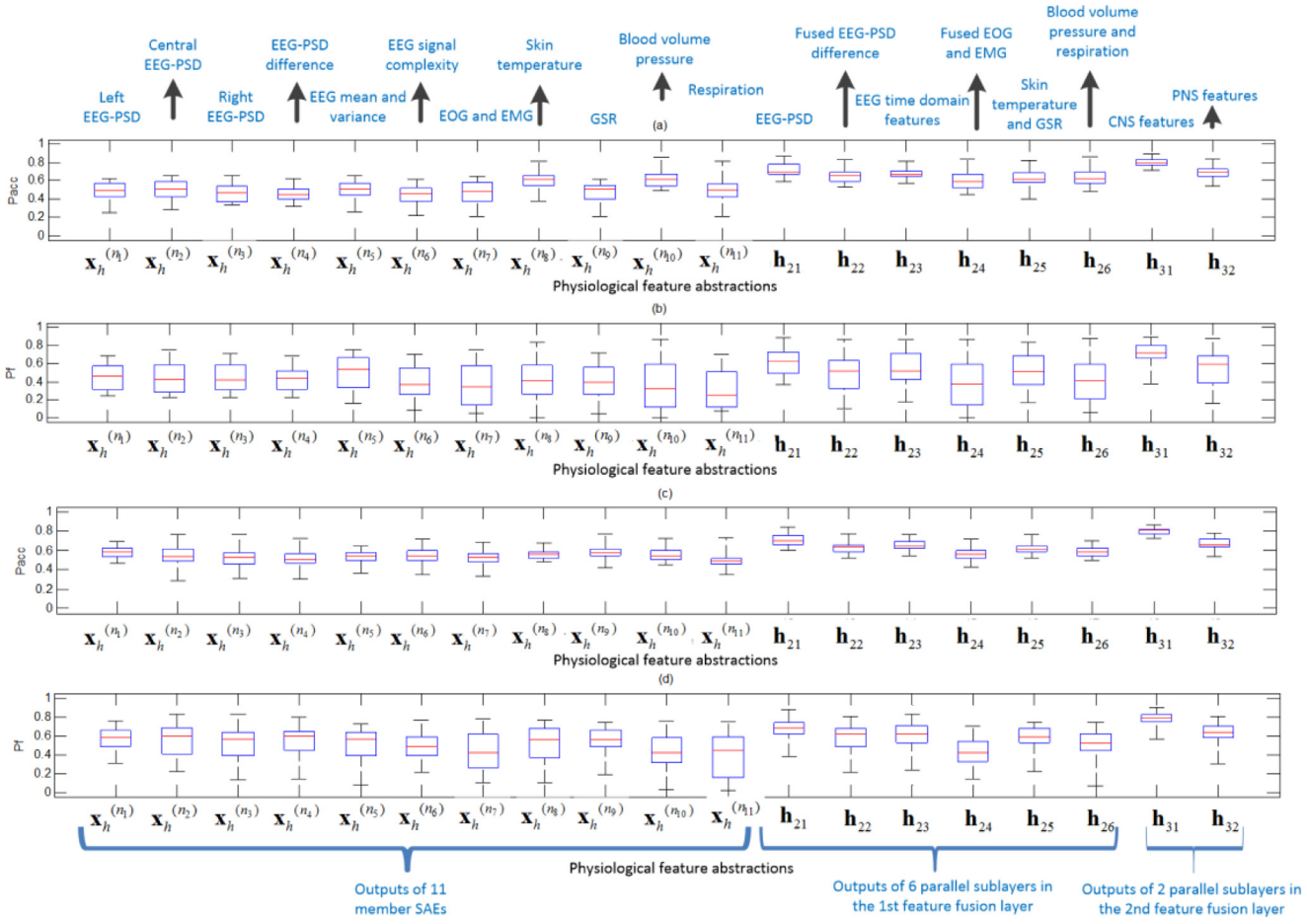

Fig. 13. Box-whisker plots for arousal ((a)-(b)) and valence (c)-(d)) classification performance using specific type of physiological feature abstractions under Case 2.

Table 5
Participant-average arousal classification performance of the MESAE scheme 2 under Case 1.

			Number of member classifiers using PNS feature abstractions (T_2)			
			0	*1*	*2*	*3*
Number of member classifiers using CNS feature abstractions (T_1)		*0*	–	–	–	*0.6117*
	P_{acc}	*1*	–	–	0.6648	0.6516
		2	–	**0.6719**	0.6633	0.6695
		3	0.6625	0.6695	0.6625	0.6594
		0	–	–	–	0.5179
	P_f	*1*	–	–	0.5941	*0.4838*
		2	–	**0.6045**	0.5141	0.5897
		3	0.5901	0.5295	0.5878	0.5134

The participant-average arousal and valence classification performances of the MESAE scheme 2 under Case 1 are shown in Tables 5 and 6, respectively. In Table 5, the highest P_{acc} (0.6719) and P_f (0.6045) are achieved by the same classifier combination, i.e., ($T_1 = 2$, $T_2 = 1$). It indicates that the classifiers based on fused abstractions of EEG-PSD differences, EEG time domain features, and EOG/EMG features construct the optimal classification committee. The lowest P_{acc} (0.6117) and P_f (0.4838) are derived with the cases of ($T_1 = 0$, $T_2 = 3$) and ($T_1 = 1$, $T_2 = 3$), respectively. In Table 6, the highest P_{acc} (0.6828) and P_f (0.6381) are obtained with the same case of ($T_1 = 3$, $T_2 = 0$), i.e., using all member classifiers from CNS feature abstractions, while the lowest values (0.5523 and 0.4761) are derived with ($T_1 = 0$, $T_2 = 3$) and ($T_1 = 1$, $T_2 = 3$) for P_{acc} and P_f, respectively. Tables 7 and 8 summarize the classification performances of the MESAE scheme 2 under Case 2. Similarly, the higher classification performance benefits more from CNS member classifiers. For arousal dimension, the highest P_{acc} (0.7218) and P_f (0.6501) values are all achieved at ($T_1 = 3$, $T_2 = 0$). For arousal dimension, the highest P_{acc} and P_f values are achieved at ($T_1 = 3$, $T_2 = 2$) and ($T_1 = 3$, $T_2 = 0$), respectively. In general, for both emotion dimensions and cases, setting $T_1 > T_2$ can yield better P_{acc} and P_f while the classification performance of under Case 2 is superior to that of Case 1 due to larger size of the training instances.

The classification results of MESAE in ensemble scheme 1 and 2 (the best results shown in Tables 5–8) as well as CNS feature abstraction fusion case (the best results shown in Figs. 12 and 13) under Case 1 and 2 are summarized in Table 9. From the table, the MESAE scheme 1 achieves the highest performance for all cases. It indicates the higher performance of MESAE benefits from the comprehensive fusion of multimodal feature abstractions with multiple fusion layers. The MESAE performance could be undermined when the fusion layers or the modalities of the abstractions are insuffi-

Table 6
Participant-average valence classification performance of the MESAE scheme 2 under Case 1.

			Number of member classifiers using PNS feature abstractions (T_2)			
			0	*1*	*2*	*3*
Number of member classifiers using CNS feature abstractions (T_1)		*0*	–	–	–	*0.5523*
	P_{acc}	*1*	–	–	0.6273	0.6148
		2	–	0.6680	0.6469	0.6430
		3	**0.6828**	0.6492	0.6555	0.6500
		0	–	–	–	0.5082
	P_f	*1*	–	–	0.5877	*0.4761*
		2	–	0.6229	0.5301	0.6016
		3	**0.6381**	0.5318	0.6121	0.5502

Table 7
Participant-average arousal classification performance of the MESAE scheme 2 under Case 2.

			Number of member classifiers using PNS feature abstractions (T_2)			
			0	*1*	*2*	*3*
Number of member classifiers using CNS feature abstractions (T_1)		*0*	–	–	–	0.6546
	P_{acc}	*1*	–	–	0.6633	0.6704
		2	–	0.6815	0.6871	0.6873
		3	**0.7218**	0.7032	0.7078	0.7054
		0	–	–	–	0.5807
	P_f	*1*	–	–	0.5986	0.5300
		2	–	0.6025	0.5646	0.6092
		3	**0.6501**	0.5799	0.6394	0.5959

Table 8
Participant-average valence classification performance of the MESAE scheme 2 under Case 2.

			Number of member classifiers using PNS feature abstractions (T_2)			
			0	*1*	*2*	*3*
Number of member classifiers using CNS feature abstractions (T_1)		*0*	–	–	–	0.6253
	P_{acc}	*1*	–	–	0.6470	0.6421
		2	–	0.6615	0.6636	0.6794
		3	0.7099	0.6850	**0.7108**	0.6990
		0	–	–	–	0.5780
	P_f	*1*	–	–	0.6092	0.5222
		2	–	0.6192	0.5631	0.6451
		3	**0.6910**	0.5957	0.6840	0.6284

Table 9
Participant-average classification performance of MESAE using different ensemble classification schemes.

	Arousal		*Valence*	
	P_{acc}	P_f	P_{acc}	P_f
MESAE (CNS feature abstraction fusion, Case 1)	0.7570	0.6388	0.7602	0.7093
MESAE (Ensemble scheme 2, Case 1)	0.6719	0.6045	0.6828	0.6381
MESAE (Ensemble scheme 1, Case 1)	**0.7719**	**0.6901**	**0.7617**	**0.7243**
MESAE (CNS feature abstraction fusion, Case 2)	0.7986	0.7101	0.8009	0.7731
MESAE (Ensemble scheme 2, Case 2)	0.7218	0.6501	0.7108	0.6910
MESAE (Ensemble scheme 1, Case 2)	**0.8418**	**0.7798**	**0.8304**	**0.7950**

Note: The highest value for each case is marked in bold type.

cient. The higher P_{acc} and P_f values under Case 2 imply the MESAE performance can be further improved by using larger size of the training instances.

4. Discussions

In order to improve the effectiveness in learning the optimal model parameters with limited training instances as well as high-dimensional physiological features (i.e., 40×425 or 400×425 data matrix for each participant), the following three schemes are incorporated for modeling the MESAE emotion classifier.

(1) Instead of concatenating all features as a whole vector, we employ 11 sub-vectors based on the type of the physiological modality to reduce the input dimensionality for each SAE. More specifically, the numbers of the input neurons of the 11 SAEs are 70, 20, 70, 56, 64, 64, 21, 6, 25, 7, and 22, respectively. According to the 11 low-dimensional feature subspaces, the reliability for both of the retraining and the fine-tuning algorithm of the deep model was expected to be improved.

(2) The intrinsic dimensionality of each feature subspace was estimated via the proposed SLF, i.e., the loss function for determining the minimal network structure. The minimum number of neurons of each hidden layer has been selected for preserving the geometrical information in physiological features of each modality. Based on that, the member SAE was derived with fewer parameters required to be learned. Such approach enhanced the robustness of the AE architec-

ture to filter the noise components. Consequently, the efficiency of the pre-training stage of the deep model has been further ensured.

(3) Finally, the data augmentation was applied in the fine-tuning stage to generate artificial training instances. To avoid the overfitting, the artificial feature vector was computed by superposing a disturbance sampled from the standard Gaussian distribution on the original feature value. The reliability for the convergence of the BP algorithm could be improved.

The significant improvement of generalization capability of the MESAE has been found regarding all 32 participants compared against several single and shallow emotion classifiers. The *F*1-score is more representative than the classification accuracy to evaluate classification performance considering the imbalance between low vs. high emotion classes. The significant inter-participant variations have been observed. For specific participants, the perfect recognition for both arousal and valence dimensions were achieved. According to the classification details presented by confusion matrices, MESAE has a better robustness against the data imbalance. The reason behind is that both of the data augmentation and pre-training strategy of the deep network has the capability to obtain more reliable data abstractions in high-dimensional space.

The inter-class discrimination capability could be gradually optimized when eliciting hierarchical representations of physiological features across different SAE hidden layers and feature fusion layers. The classification performance solely based on the CNS feature is significantly better than that of the PNS feature for both of arousal and valence dimensions. The potential reason is that the EEG activity has a more rapid response within 60 sec. in reflecting the emotional states than GSR or the skin temperature.

According to the results comparisons between different fusion schemes of MESAE, scheme 1 that uses three hierarchical fusion layers possess the best classification accuracy and *F*1-score among all ensemble schemes. The potential reason is that the majority voting based on the single fusion layer cannot provide much improvement upon the classification performance compared with an optimized single classifier when severe intra- and inter-participant variations both exist in physiological signals [46].

To conclude, the results presented in Sect. 3.2 validated that the MESAE significantly outperforms several conventional classifiers as well as the hybrid classification approaches that adopted two layers of mappings to process physiological features. It suggests a sufficiently deep architecture of the model should be implemented to improve the emotion classification performance, e.g., six hidden layers. In Sect. 3.3, the MESAE has been applied on different feature spaces of the physiological data. The results indicate the discrimination power of the feature abstractions can be stably improved by increasing the number of modalities. The observation implies the deep classifier ensemble is more suitable to cope with the high-dimensional problem of the multimodality than the single modality. The results in Sect. 3.4 show how the numbers of input modalities and fusion layers affect the generalizability of the MESAE. Consequently, the deep structure based ensemble classifier is quite competitive for emotion recognition via multimodal physiological signals when both of its structure and abstraction fusion scheme are carefully optimized.

Finally, the limitations in current work and corresponding further work may include:

(1) The MESAE emotion classifier presents in the manuscript is designed by the subject-specific paradigm. That is, for each new user, both of the structure and the weights of the deep network must be determined again to fit the data distribution of the target physiological data. The model trained by the data from one subject may be not adaptable to the testing data from another subject. Hence, it is necessary to investigate whether the common or shared deep structure for all subjects exists or not in the future work.

(2) Moreover, the performance of the MESAE has been undermined when both of the training and validating data are limited. The potential reason is that the deep model requires lager size of the training samples than the shallow classifier since additional structural hyper-parameters are required to be selected. In the future work, we will investigate the performance of the deep classifier by using huge volumes of physiological data as well as the effective data augmentation technique to artificially generate feature vectors.

(3) After examining the classifier for specific subjects whose classification performance is quite low, we found high skewness of the instance size between low and high arousal/valence classes. It indicates the performance of the MESAE is still limited with the extremely imbalanced data. Another potential reason is that the feature extraction may be failed for the data of these subjects. Both of issues should be investigated in our future work by incorporating cost-sensitive training scheme or improve discriminating capability of the extracted features.

5. Conclusion

The physiological-data-driven approaches for learning deep network structure and fusing high-level abstractions presented in this study are effective for designing more accurate deep emotion classifiers. Based on the optimal geometrical information preservation between the physiological features and feature abstractions, the parsimonious structure of the MESAE can be properly identified and lead to a higher generalization capability than the shallow emotion classifiers and the state-of-the-art deep learning models. Such structural learning framework can avoid the uncertainty of the empirical determination of the hidden-neuron numbers in the deep emotion classifier. On the other hand, the feasibility of the MESAE with multiple fusion layers is supported by the visualization and quantitative evaluation of discrimination capacity for intermediate representations of physiological features at different levels. The reliability of the structural learning and the multilayer based fusion methods in MESAE is validated by the stable superiority on different sizes of the available physiological instances (i.e., small or large size) extracted from DEAP database.

Conflict of interests

None declared.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 61673276, No. 11502145, No. 61603256, the Foundation of Shanghai Municipal Education Commission and the Faculty Innovation Ability Development Project of University of Shanghai for Science and Technology.

References

[1] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multi-modal database for affect recognition and implicit tagging, IEEE Trans. Affect. Comput. 3 (2012) 42–55.

[2] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, L. Van Gool, A 3-D audio-visual corpus of affective communication, IEEE Trans. on Multimedia 12 (2010) 591–598.

[3] Z. Yin, J. Zhang, Operator functional state classification using least-square support vector machine based recursive feature elimination technique, Comput. Methods Prog. Biomed. 113 (2014) 101–115.

[4] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 39–58.

[5] J. Kim, E. Andre, Emotion recognition based on physiological changes in music listening, IEEE Trans. Pattern Anal. Mach. Intell. 30 (2008) 2067–2083.

[6] A. Hanjalic, L.-Q. Xu, Affective video content representation and modeling, IEEE Trans. Multimedia 7 (2005) 143–154.
[7] C. Brunner, C. Vidaurre, M. Billinger, C. Neuper, A comparison of univariate, vector, bilinear autoregressive, and band power features for brain-computer interfaces, Med. Biol. Eng. Comput. 49 (2011) 1337–1346.
[8] N. Birbaumer, Breaking the silence: brain-computer interfaces (BCI) for communication and motor control, Psychophysiology 43 (2006) 517–532.
[9] J. Zhang, Z. Yin, R. Wang, Recognition of mental workload levels under complex human-machine collaboration by using physiological features and adaptive support vector machines, IEEE Trans. Hum. Mach. Syst. 45 (2015) 200–214.
[10] F. Agrafioti, D. Hatzinakos, A.K. Anderson, ECG pattern analysis for emotion detection, IEEE Trans. Affect. Comput. 3 (2012) 102–115.
[11] R.W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: analysis of affective physiological state, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 1175–1191.
[12] R. Cowie, E. Douglas-Cowie, Emotion recognition in human-computer interaction, IEEE Signal Process. Mag. 18 (2001) 32–80.
[13] J. Kim, E. Andre, Emotion-specific dichotomous classification and feature-level fusion of multichannel biosignals for automatic emotion recognition, in: Proceedings of IEEE Inter-national Conference on Multisensor Fusion and Integration for Intelligent Systems, Seoul, Korea, August 20-22, 2008.
[14] H. Lee, A. Shackman, D. Jackson, R. Davidson, Test-retest reliability of voluntary emotion regulation, Psychophysiology 46 (2009) 874–879.
[15] I.C. Christie, B.H. Friedman, Autonomic specificity of discrete emotion and dimensions of affective space: a multivariate approach, Int. J. Psychophysiol. 51 (2004) 143–153.
[16] G. Chanel, C. Rebetez, M. Bétrancourt, T. Pun, Emotion assessment from physiological signals for adaptation of game difficulty, IEEE Trans. Syst. Man Cybern. A Syst. Humans 41 (2011) 1052–1063.
[17] O. AlZoubi, S.K. D'Mello, R.A. Calvo, Detecting naturalistic expressions of nonbasic affect using physiological signals, IEEE Trans. Affect. Comput. 3 (2012) 298–310.
[18] L. Brown, B. Grundlehner, J. Penders, Towards wireless emotional valence detection from EEG, in: *Proceedings of Engineering in Medicine and Biology Society*, EMBC, 2011 Annual International Conference of the IEEE 2011.
[19] R.J. Davidson, Affective neuroscience and psychophysiology: toward a synthesis, Psychophysiology 40 (2003) 655–665.
[20] Z. Yin, J. Zhang, Identification of temporal variations in mental workload using locally-linear-embedding -based EEG feature reduction and support-vector-machine-based clustering and classification technique, Comput. Methods Prog. Biomed. 115 (2014) 119–134.
[21] G. Lee, M. Kwon, S. Kavuri-Sri, M. Lee, Emotion recognition based on 3D fuzzy visual and EEG features in movie clips, Neurocomputing 144 (2014) 560–568.
[22] J. Cannon, P.A. Krokhmal, Y. Chen, R. Murphey, Detection of temporal changes in psychophysiological data using statistical process control methods, Comput. Methods Prog. Biomed. 107 (2012) 367–381.
[23] N. Martini, D. Menicucci, L. Sebastiani, R. Bedini, A. Pingitore, N. Vanello, A. Gemignani, The dynamics of EEG gamma responses to unpleasant visual stimuli: from local activity to functional connectivity, NeuroImage 60 (2012) 922–932.
[24] D. Nie, X.-W. Wang, L.-C. Shi, B.-L. Lu, EEG-based emotion recognition during watching movies, In: *Proceedings of the 5th International IEEE/EMBS Conference on Neural Engineering (NER)* 2011.
[25] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, F. Babiloni, Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness, Neurosci. Biobehav. Rev. 44 (2014) 44.
[26] G.K. Verma, U.S. Tiwary, Multimodal fusion framework: a multiresolution approach for emotion classification and recognition from physiological signal, NeuroImage 102 (2014) 162–172.
[27] M. Balconi, C. Lucchiari, EEG correlates (event-related desynchronization) of emotional face elaboration: a temporal analysis, Neurosci. Lett. 392 (2006) 118–123.
[28] C.A. Frantzidis, C. Bratsas, C.L. Papadelis, E. Konstantinidis, C. Pappas, P.D. Bamidis, Toward emotion aware computing: an integrated approach using multichannel neurophysiological recordings and affective visual stimuli, IEEE Trans. Inf. Technol. Biomed. 14 (2010) 589–597.
[29] M. Balconi, G. Mazza, Brain oscillations and BIS/BAS (behavioral inhibition/activation system) effects on processing masked emotional cues: ERS/ERD and coherence measures of alpha band, Int. J. Psychophysiol. 74 (2009) 158–165.
[30] E.I. Konstantinidis, C.A. Frantzidis, C. Pappas, P.D. Bamidis, Real time emotion aware applications: a case study employing emotion evocative pictures and neuro-physiological sensing enhanced by graphic processor units, Comput. Methods Prog. Biomed. 107 (2012) 16–27.
[31] M. Khezri, M. Firoozabadi, A.R. Sharafat, Reliable emotion recognition system based on dynamic adaptive fusion of forehead biopotentials and physiological signals, Comput. Methods Prog. Biomed. 122 (2015) 149–164.
[32] J. Atkinson, D. Campos, Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers, Expert. Syst. Appl. 47 (2016) 35–41.
[33] S. Liu, D. Zhang, M. Xu, H. Qi, F. He, X. Zhao, P. Zhou, L. Zhang, D. Ming, Randomly dividing homologous samples leads to overinflated accuracies for emotion recognition, Int. J. Psychophysiol. 96 (2015) 29–37.
[34] C. Li, C. Xu, Z. Feng, Analysis of physiological for emotion recognition with the IRS model, Neurocomputing 178 (2016) 103–111.
[35] J. Christensen, J. Estepp, G. Wilson, C. Russell, The effects of day-to-day variability of physiological data on operator functional state classification, NeuroImage 59 (2012) 57–63.
[36] F. Laurent, M. Valderrama, M. Besserve, M. Guillard, J.-P. Lachaux, J. Martinerie, G. Florence, Multimodal information improves the rapid detection of mental fatigue, Biomed. Signal. Process. Control. 8 (2013) 400–408.
[37] Y. Shin, S. Lee, M. Ahn, H. Cho, S.C. Jun, H.-N. Lee, Simple adaptive sparse representation based classification schemes for EEG based brain–computer interface applications, Comput. Biol. Med. 66 (2015) 29–38.
[38] D. Iacoviello, A. Petracca, M. Spezialetti, G. Placidi, A real-time classification algorithm for EEG-based BCI driven by self-induced emotions, Comput. Methods Prog. Biomed. 122 (2015) 293–303.
[39] R.M. Mehmood, H.J. Lee, A novel feature extraction method based on late positive potential for emotion recognition in human brain signal patterns, Comput. Electr. Eng. 53 (2016) 444–457.
[40] D. Wang, Y. Shang, Modeling physiological data with deep belief networks, Int. J. Inf. Educ. Technol. 3 (5) (2013) 505–511.
[41] X. Li, P. Zhang, D. Song, G. Yu, Y. Hou, B. Hu, EEG based emotion identification using unsupervised deep feature learning, SIGIR2015 Workshop on Neuro-Physiological Methods in IR Research, 13 August 2015.
[42] K. Li, X. Li, Y. Zhang, A. Zhang, Affective state recognition from EEG with deep belief networks, IEEE 13th International Conference on Bioinformatics and Biomedicine, 2013.
[43] X. Jia, K. Li, X. Li, A. Zhang, A Novel semi-supervised deep learning framework for affective state recognition on EEG signals, IEEE 14th International Conference on Bioinformatics and Bioengineering, 2014.
[44] S. Jirayucharoensak, S. Pan-Ngum, P. Israsena, EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation, Scientific World J. 1 (2014) 1:10.
[45] E. Acar, F. Hopfgartner, S. Albayrak, Understanding affective content of music videos through learned representations, 20th Anniversary International Conference, January 2014.
[46] J. Zhang, Z. Yin, R. Wang, Pattern classification of instantaneous cognitive task-load through GMM clustering, laplacian eigenmap and ensemble SVMs, IEEE/ACM Trans. Comput. Biol. Bioinf. (May 2016) Available online, doi:10.1109/TCBB.2016.2561927.
[47] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy layer-wise training of deep networks, Adv. Neural. Inf. Process. Syst. 19 (2007) 153–160.
[48] S. Koelstra, C. Muehl, M. Soleymani, J.-S. Lee, A. Yazdani, T.E. Ebrahimi, T. Pun, A. Nijholt, I.Y. Patras, DEAP: a database for emotion analysis using physiological signals, IEEE Trans. Affect. Comput. 3 (2012) 18–31.
[49] Y. Liu, X. Feng, Z. Zhou, Multimodal video classification with stacked contractive autoencoders, Sig. Process. 120 (2016) 761–766.
[50] J. Li, Z. Struzik, L. Zhang, A. Cichocki, Feature learning from incomplete EEG with denoising autoencoder, Neurocomputing 165 (2015) 23–31.
[51] M. Belkin, P. Niyogi, Laplacian Eigenmaps and spectral techniques for embedding and clustering, Adv. Neural. Inf. Process. Syst. (2002) 585–591.
[52] L. van der Maaten, E. Postma, J. van den Herik, Dimensionality Reduction: A Comparative Review, Tilburg University, 2009 *Technical Report 2009-005*.
[53] Y. Liu, O. Sourina, EEG-based valence level recognition for real-time applications, in: IEEE International Conference on Cyberworlds (CW), 2012, pp. 53–60.
[54] D.S. Naser, G. Saha, Recognition of emotions induced by music videos using DT-CWPT, in: Indian Conference on Medical Informatics and Telemedicine (ICMIT) IEEE, 2013, pp. 53–57.
[55] J. Chen, B. Hua, P. Moore, X. Zhang, X. Ma, Electroencephalogram-based emotion assessment system using ontology and data mining techniques, Appl. Soft. Comput. 30 (2015) 663–674.
[56] H.J. Yoon, S.Y. Chung, EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm, Comput. Biol. Med. 43 (2013) 2230–2237.