# Emotion Distribution Learning Based on Peripheral Physiological Signals

Yezhi Shu, Pei Yang, Niqi Liu, Shu Zhang, Guozhen Zhao, and Yong-Jin Liu, *Senior Member, IEEE*

**Abstract**—Emotion analysis based on peripheral physiological signals has attracted increasing attention recently in affective computing. Previous works usually predict emotional states using a single emotion label for each discrete time. However, in real-world scenarios, it is not sufficient due to the fact that the real-world emotional state is usually a mixture of basic emotions. In this paper, we formulate the emotion analysis as an emotion distribution learning (EDL) problem and make two contributions. First, we establish a standardised dataset containing four negative emotions (anger, disgust, sadness, fear) and three positive emotions (tenderness, joy, amusement), which could be a useful benchmark for the EDL task. Second, we propose an emotion distribution prediction system which has the following distinct characteristics: (1) after processing raw peripheral physiological signals, we compute totally 89 representative features from four channels, i.e., GSR, SKT, ECG and HR, (2) an adaptive feature selection strategy based on recursive feature elimination (RFE) is used to select the most significant features in our EDL task, and (3) we design a dedicated EDL model based on convolution neural networks that takes information from both the feature correlation and the time domain into consideration. Experiments were conducted to validate our proposed system, and the results indicated that (1) the proposed feature selection strategy effectively selects significant features and improves algorithmic performance, and (2) the proposed EDL model can obtain good results in terms of six evaluation measures and outperform existing methods.

**Index Terms**—Emotion recognition, distribution learning, peripheral physiological signals, feature selection

## 1 Introduction

Emotion recognition has attracted increasing interests in recent years in affective computing and human-computer interaction (HCI). Emotion is the psychological and physiological human response raised by neurophysiological changes, and various kinds of measures for emotion recognition have been studied [1]. These measures can be broadly classified into audio-visual based and physiological based categories [2].

Audio-visual based measures detect emotions from behavioral signals, such as speech sequences, facial expressions and gestures. Compared to behavioral signals, physiological signals are difficult to conceal. For example, people can disguise a smile during negative emotional experience [3]; however, it is unlikely to control physiological reactions such as electroencephalography (EEG), heart rate, skin temperature, etc. Moreover, according to Levenson's research on Americans and the Minangkabau of West Sumatra, there is cross-cultural consistency of autonomic nervous system (ANS) representation between different nations [4]. These findings indicate the advantage of using physiological signals for emotion recognition. As summarized in Section 2, both brain signals (e.g., EEG) and peripheral physiological signals have been studied for emotion analysis. Due to the population of portable peripheral devices such as smart wrist watches, in our study, we pay attention to peripheral physiological signals including electrocardiogram (ECG), heart rate (HR), galvanic skin response (GSR) and skin temperature (SKT).

To characterize emotions, two representative emotion models exist [5]: the discrete model and the dimensional model. The former uses a fixed number of basic emotions and the six-basic-emotion form [6] is such a typical example. The latter describes emotion in a continuous 2D or 3D space in which every point represents a specified subtle emotion. Widely used dimensional models includes the 2D valence-arousal (VA) model and the 3D valence-arousal-dominance (VAD) model. According to the cognitive theory [7], [8], valence indicates whether emotion is positive or negative, arousal indicates the intensity of emotion, and dominance refers to the degree of human control over the emotion. Although the dimensional model is easier to describe much more emotion types, when two or more discrete emotions are close in the VA or VAD space, the dimensional model becomes difficult to distinguish them, while the discrete model is easy to characterize them by their unique behavioral responses [9]. In our study, we use the discrete model to study a mixture of positive emotions (tenderness, joy and amusement) and negative emotions (anger, disgust, sadness and fear), which are frequently induced in real-world scenarios.

- *Yezhi Shu, Pei Yang, Niqi Liu, Shu Zhang, and Yong-Jin Liu are with the BNRist, Department of Computer Science and Technology, MOE-Key Laboratory of Pervasive Computing, Tsinghua University, Beijing 100084, China. E-mail: {shuyz19, yangpei20, liunq18}@mails.tsinghua.edu.cn, {zhangshu2020, liuyongjin}@tsinghua.edu.cn.*
- *Guozhen Zhao is with the CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing 100101, China, and also with the Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: zhaogz@psych.ac.cn.*

*Manuscript received 23 May 2021; revised 18 Mar. 2022; accepted 27 Mar. 2022. Date of publication 30 Mar. 2022; date of current version 13 Sept. 2023.*
*This work was supported in part by the Natural Science Foundation of China under Grants U1736220, 61725204 and in part by Tsinghua University Initiative Scientific Research Program.*
*(Corresponding author: Yong-Jin Liu.)*
*Recommended for acceptance by W. Zheng.*
*Digital Object Identifier no. 10.1109/TAFFC.2022.3163609*

Many previous researches recognize emotions using a single emotion label for each discrete time. However, in our daily life, the real induced emotions are unlikely to be a single pure emotion state, but rather a mix of several emotions with different intensities (especially for the positive emotions). Therefore, in this paper, we formulate the emotion recognition as an emotion distribution learning (EDL) problem. In the learned emotion distribution, each component of the distribution corresponds to an intensity of a basic emotion in the discrete model. To the best of authors' knowledge, there is no publicly available dataset of peripheral physiological signals with emotion distribution labels. In this paper, we construct such a dataset and propose an emotion distribution learning (EDL) model based on convolution neural networks. The contributions of this paper are two folds:

- We construct a standardized emotion distribution dataset with peripheral physiological signals. In order to evoke reliable target-specific emotions, realistic movie clips with high ecological validity are selected and used as inducing materials. We label the dataset according to the subjects' self-reports and transform these labels into emotion distributions; therefore our dataset provides a useful benchmark for the EDL task.
- We proposed an emotion distribution prediction system which incorporates a novel feature selection strategy and a 2D convolution neural network model for EDL. Experiments show that our proposed method can achieve good results and outperforms existing methods.

## 2 Related Work

### 2.1 Emotion Induction

Emotion induction is a necessary prerequisite in emotion analysis, in which the selection of stimulus materials plays an important role. Visual materials are one of the most commonly used stimulus materials, and some well-known datasets of visual materials include Affective Norms for English Words (ANEW), Affective Norms for English Text (ANET), International Affective Picture System (IAPS) [10], [11], [12]. Another form is auditory materials, and one representative dataset is International Affective Digital Sounds (IADS) [13], [14], which collects birdsong, explosive sound and the sound of rain, etc. Many studies have shown that music evokes emotional responses [15], [16]. Some researchers also use music to elicit emotional state [17], [18], [19]. In recent years, many studies (e.g., [20]) use the combination of audio-visual stimuli (such as film clip and music video) to induce emotions.

In this paper, we follow the work [20] to use film clips for eliciting emotions, by considering the following reasons. First, films share the most advantages that pictures offer and they are dynamic which are more similar to our realistic life [21]. Second, real-world emotions are often elicited by dynamic visual and auditory stimuli which are similar to films, and therefore, films have high degree of ecological validity [22]. Finally, the dynamic combined audio-visual stimuli can induce emotions without deception [20], [22].

### 2.2 Emotion Recognition Using Physiological Signals

Emotion analysis based on physiological signals is a hot topic in affective computing. At present, many studies focus on discrete emotion recognition using EEG signals. Ruiz-Padial *et al.* [23] used Higuchi fractal dimension (HFD) of EEG and heart rate variability (HRV) of ECG as features, to investigate their relationships with four discrete emotions (i.e., disgust, fear, humor and neutrality). They found that brain complexity changed similarly as HRV in response to different video clips which induce specific emotions. Liu *et al.* [20] built a real-time discrete emotion system based on EEG, and they used movie clips to induce eight emotions including joy, amusement, tenderness, anger, disgust, fear, sadness and neutrality. In this work, EEG data were collected by using the Emotiv EPOC system, and the accuracy was up to 60.55% by importing prior knowledge into their proposed 3-layer support vector machine (SVM) based recognition system.

Deep learning technology has also been used for EEG-based emotion recognition. Zheng *et al.* [24] employed deep belief networks (DBNs) to recognize three valences of emotions (i.e., positive, neutral and negative emotions), and their results showed the superiority of DBN over SVM, logistic regression (LR), and $k$ nearest neighbor (kNN). Tang *et al.* [25] introduced a Bimodal Long Short-Term Memory (LSTM) network which takes temporal information of multimodal signals into consideration for emotion recognition. In addition, in order to explore the relations among different EEG channels and minimize the feature distribution shift between different sessions and/or subjects, Du *et al.* proposed an attention-based LSTM model named ATDD-LSTM [26]. By introducing both attention mechanism and domain discriminator, ATDD-LSTM achieves state-of-the-art performance on subject-dependent, subject-independent and cross-session evaluation.

Besides EEG signals, peripheral physiological signals have also been used for emotion recognition. Peripheral physiological signals including cardio activity, skin conductance, etc., were used in [27] to recognize amusement and sadness for representing positive and negative emotions, respectively. However, only recognizing two emotion types seriously limited the scope of applications. Kragel *et al.* [28] collected electrodermal, cardiac, respiratory and gastric data using the BIOPAC MP150 system, and these signals were used together with self-report measures in a seven-emotion recognition task. The accuracy is up to 58.0% by using only peripheral physiological signals, and their results showed that the discrete model is better to characterize emotions in peripheral physiological signals than the dimensional model. Observing that emotion is usually expressed in multiple modalities, Zhang *et al.* [29] collected 3D dynamic imaging, 2D video, thermal videos and peripheral physiological signals. In their work, peripheral data including heart rate, blood pressure, respiration and skin conductivity (Electrodermal activity, EDA) were captured by the BIOPAC MP150 system, and extracted multi-modal features were fed into SVM for classifying five discrete emotions.

All above works only classified or predicted a single emotion type for a specific signal sample (collected in a short time). To make the classification more closer to real-world

scenarios, in this paper, we propose a model which predicts a distribution of basic emotions.

### 2.3 Label Distribution Learning

In the real world, the human emotion state at every moment is often a mix of basic emotions, and then it is not sufficient to use a pure emotion to describe the emotional state. To remedy this defect, multi-label learning (MLL) was proposed (e.g., [30] in text emotion analysis), which assigns multiple emotion labels to an instance. However, MLL cannot provide a weight for each emotion label, i.e., it cannot solve the emotion label ambiguity problem [31]. In order to describe the weight or role of each label, a new learning paradigm called *label distribution learning* (LDL) was proposed [32], [33]. In LDL, each label was assigned a non-negative real number (defined as *description degree*) which represents the degree of how the label describes the instance. For any instance, the sum of description degrees of all labels is 1, indicating a full description of this instance [32]. Compared to MLL, LDL is more suitable to deal with the emotion ambiguity.

So far, several representative LDL algorithms have been proposed for various specific applications [32], [33], [34], [35], [36]. In the work [33], Geng *et al.* proposed three strategies to design LDL algorithms. The first strategy, called problem transformation, is to transform a LDL problem into a single label learning (SLL) problem. The key is to construct a single label training set (from original training samples with distribution labels) using resampling technology. To do so, two representative algorithms are PT-Bayes and PT-SVM. The second strategy is to use algorithm adaptation which extends some existing learning methods to deal with label distribution. For this strategy, algorithms including AA-kNN and AA-BP can be used. The third strategy is to use specialized algorithms such as the conditional probability neural network (CPNN) proposed in [32] which learns the conditional probability density function.

Yang *et al.* [35] improved CPNN by introducing binary encoding for label and distribution augmentation strategies. Their proposed BCPNN and ACPNN were used for image sentiment distribution learning. Furthermore, Yang *et al.* [34] and Zhang *et al.* [36] proposed multi-task frameworks for image sentiment distribution learning and text emotion distribution, respectively. Different from the existing label distribution learning tasks such as image and text sentiment distribution mentioned above, in this paper, by treating emotion types as general labels, we study the emotion distribution learning (EDL) problem for human beings using peripheral physiological signals.

Due to the fact that emotion distribution studied in this paper is significantly dependent on the individuals, our task is more challenging compared to the the sentiment analysis on image or sentence. Furthermore, the datasets, features and algorithms for analyzing peripheral physiological signals are significantly less than those used for images or text sentiment analysis, which makes our task more difficult. In order to predict the emotion distribution of individuals using peripheral physiological signals, we constructed a dataset and proposed a emotion distribution prediction system which includes a novel feature selection strategy and a convolution neural network model that takes feature correlation information and time domain information into consideration.

## 3 Constructing Dataset of Peripheral Physiological Signals

One great challenge in our study is the lack of a dataset of peripheral physiological signals with emotion distribution labels. In this section, we construct such a dataset, which provides a useful benchmark for the EDL task.

### 3.1 Emotion Induction by Movies

Ge *et al.* [37] established a standardised database of Chinese emotional film clips, which has been shown to be effective in inducing emotional states according to the validation on a large sample [38]. To construct the database, nine trained research assistants first collected more than 1,000 Chinese film excerpts, from which 111 clips were selected by three cognitive psychologists by evaluating the potential to successfully elicit the target emotion of each clip. Then, 39 effective film clips were further selected by 315 undergraduate and graduate students based on two objective criterion (hit rate and intensity) and eventually 22 clips were selected by another 147 undergraduate students (using the success index) to build the standardised database. In our study, different from labeling a single emotion, it is more likely to get confused when labelling the emotion distribution. In order to maintain the consistency of emotion distribution estimated by different subjects, we employed an expert with rich experience in emotion assessment to further evaluate each clip, and a total of 14 emotional movie clips including 7 emotion categories (anger, disgust, sadness, fear, tenderness, joy and amusement) were selected from the standardized database of Chinese emotional film clips. Among these 14 emotional movie clips, 7 were positive emotions and the other 7 were negative emotions. The used 7 emotion categories (including 4 negative and 3 positive emotions) were carefully selected as good representatives to form the emotion label distributions by comprehensive literature review and subjective reports in [37] (see more details in Appendix A), which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TAFFC.2022.3163609.

Anger was induced by two clips which depicted the violent assault and massacre of Chinese girls, and lasted for 81 and 155 seconds respectively. Disgust was induced by one clip which described the plot of the boy being invaded by an old eunuch and lasted for 159 seconds. Sadness was elicited by two clips which showed scenes of a son losing his father and a father losing his son, and lasted for 140 and 116 seconds respectively. Fear was induced by two clips which were about the protagonist being chased and robbed of children by a strange woman and meeting a ghost in the bathroom respectively, and these two clips lasted for 102 and 68 seconds. As for positive movies, three tender clips showed the pictures of the master and servant playing together (99 seconds), the father and the son playing together (141 seconds), and the first meeting of the leading roles (91 seconds) respectively. Three joyful clips showed the successful horse vaulting of the protagonist (120 seconds), the lovely alien creatures (92 seconds) and the celebration of the Spring Festival in rural China (85 seconds) respectively. In addition,

one amusing clip described a very absurd war scene which lasted for 67 seconds.

### 3.2 Acquisition Procedure

Thirty eight students (undergraduates, postgraduates or doctoral students) participated in the experiment with an average age of 23.95 years (SD = 1.56). Among them, 17 were male and 21 were female. All the subjects were right-handed and healthy (without anxiety, depression or any other mental illness), with a normal or corrected-to-normal vision. The subjects were informed of the whole process of the experiment before they filled in the informed con-sent form. They were told that they could terminate the experiment at any time, and all the experimental data were anonymous and only used for scientific research. The day before the experiment, the subjects were asked not to take in alcohol, smoke, drink coffee or take medicine, and to ensure adequate sleep. Before formal experiment, biosensors were attached to subject with the help of experimenters (details are described in Section 3.3).

The experimental program was presented on the 15-inch LCD screen (1024 $\times$ 768, 60 Hz) with the script program written in VB. In the experiment, the subtitles of all the movie segments were removed and the resolution was set to 720 $\times$ 576. Since the within-subject design was adopted, each participant watched 14 video clips in random order. Within the trial of each video clip, an instruction was presented first. Then a fixation point appeared (1 second), followed by a 1-min go/no go task (when subjects see '1' on the screen, they press the key, and when '9' appears, they do not press the key), which served as a distraction operation to eliminate the effects of previous emotions. After that, subjects were asked to rest for 80 seconds and the video clip was presented. The above process would be repeated until all the video clips were played. During the whole experiment, peripheral physiological signals were recorded by an MP150 data recording system (BIOPAC Systems Inc.). The emotion distribution labels for these peripheral physiological signals were specified in Section 3.4.

### 3.3 Biosensors

For ECG and HR measurement, ECG100C amplifier was used. Experimenters wiped the skin surface of the left and right inner ankles and the right carotid artery of subject with alcohol first. Then, three Ag-AgCl pre-gelled electrodes were attached to these three places respectively. Finally, the VIN + line of ECG was connected to the electrode patch of the left lower limb, the VIN – line to the electrode patch of the right carotid artery, and the GND line to the electrode patch of the right lower limb.

For GSR acquisition, the distal parts of the right index finger and middle finger were taken. Experimenters wiped these two parts of subject with medical alcohol, and then apply conductive paste around the positive and negative poles. The sensor receives the recorder and transmits the signal data to the GSR100C amplifier. SKT was measured by SKT sensor placed on the thumb of subject connected with SKT100C amplifier.

### 3.4 Emotion Distribution Labels

As introduced above, we selected 14 emotional movie clips including 7 emotion classes. We used these 14 videos to induce emotions on 38 subjects and recorded their peripheral physiological signals. After collecting these signals, we conducted data preprocessing, feature extraction and feature selection sequentially as described in Section 4.

For each emotion-eliciting video, all subjects were asked to perform a self-assessment to indicate (1) their feeling on the degree of each of seven discrete emotions and (2) the degree of familiarity with this video. A nine-level Likert scale (1-9) was used in self-assessment, where 1 represents "not at all" and 9 represents "extremely strong" (on feeling seven discrete emotions or the familiarity degree with the video).

Let $D^x = \{d_1^x, d_2^x, \ldots, d_k^x\}$ be the emotion distribution label of $x$th emotion-eliciting video, where $k$ is the number of emotion classes and $x$ is the index of the video. $\hat{r}_{ij}^x$ denotes the chosen scale of the $j$th emotion of the $i$th subject and $f_i^x$ denotes the familiarity degree of $i$th subject with this video. We can compute corresponding emotion distribution label $D^x$ using the following steps. We first normalized $\hat{r}_{ij}^x$ for each subject as $r_{ij}^x = \frac{\hat{r}_{ij}^x}{\sum_k \hat{r}_{ik}^x}$ to reduce personal bias [39]. We followed the assumption in [40], that the more familiar the subject knows the video, the more accurate the emotional assessment is made. Then the description intensity of the $j$th emotion in $x$th video can be computed as

$$d_j^x = \frac{\sum_i f_i^x \cdot r_{ij}^x}{\sum_k \sum_i f_i^x \cdot r_{ik}^x}. \tag{1}$$

Taking different partitioning methods of subjects in training and testing phases into account, the formulation of emotion description intensity becomes

$$d_{j\_train}^x = \frac{\sum_{i \in train\_set} f_i^x \cdot r_{ij}^x}{\sum_k \sum_{i \in train\_set} f_i^x \cdot r_{ik}^x} \tag{2}$$

$$d_{j\_test}^x = \frac{\sum_{i \in test\_set} f_i^x \cdot r_{ij}^x}{\sum_k \sum_{i \in test\_set} f_i^x \cdot r_{ik}^x}. \tag{3}$$

## 4 The Proposed EDL Model

In this section, we propose a solution to the emotion distribution learning (EDL) task formulated in Section 4.4. The flowchart of the whole system is shown in Fig. 1. The system consists of two sub-processes: training and testing. The EDL model is trained (using the training set in the training phase) to predict emotion distribution of unseen data (using the testing set in the testing phase). The first three stages of training and testing processes are the same, i.e., preprocessing (Section 4.1), feature extraction (Section 4.2) and feature selection (Section 4.3).

The preprocessing stage removes noise and drift from the raw physiological data. Then we calculate totally 89 features from four-channel peripheral physiological signals in the feature extraction stage, and select the most significant features using the recursive feature elimination (RFE) algorithm. After formulating the EDL problem (Section 4.4), these selected features are used to train the proposed EDL model which is designed as a deep convolution network (Section 4.5).

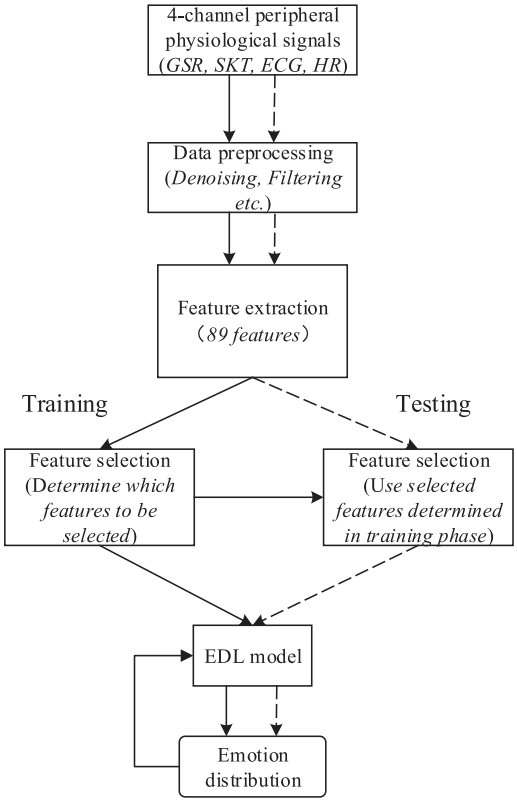



Fig. 1. The flowchart of emotion distribution prediction system based on peripheral physiological signals. The system consists of two sub-processes: training (solid lines) and testing (dashed lines). The two sub-processes share the steps of data preprocessing, feature extraction and feature selection.

### 4.1 Data Preprocessing

Since the acquisition process may easily involve noise, preprocessing is essential to regain clean peripheral physiological signals. For example, respiration, body movements, body temperature change and perspiration can lead to baseline drift [41], [42]. In order to remove or reduce the influence of noise as much as possible, according to [43], we use a second order Butterworth filter for galvanic skin response (GSR), skin temperature (SKT) and electrocardiogram (ECG) to help eliminate direct current (DC) offset and baseline drift. Fig. 2 shows one example about the waveforms of an ECG signal segment before and after filtering. It is clearly observed that after drift and DC offset removal, the baseline is stable at zero.

### 4.2 Feature Extraction

Following the work in [44], [45], we extract features from peripheral physiological signals in both temporal and frequency domains. Note that in literature (e.g., Table 5 in [44]), frequency features for peripheral physiological signals are computed with small wave bands, usually in [0, 5] Hz. Considering both real-time requirement and feature variety, we set the length of extraction unit to be 10 seconds for all frequency features, leading to the frequency resolution of 0.1 Hz. The overlap of two consecutive units is 9 seconds. We extract totally 89 features for each unit from four channels; i.e., 39 features from galvanic skin reaction (GSR), 4 features from skin temperature (SKT), 39 features from electrocardiogram (ECG) and 7 features from heart rate (HR). Full details of all 89 features are presented in Appendix B, available in the online supplemental material.

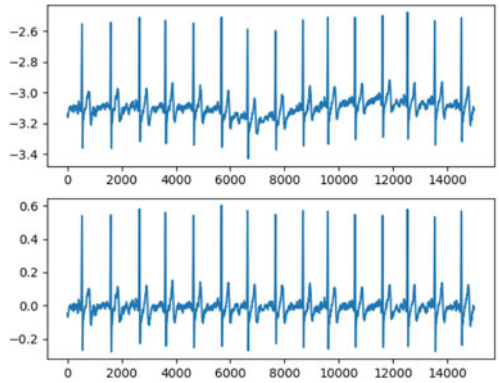



Fig. 2. Waveforms of a 15s segment of original ECG signal (top row) and corresponding filtered (bottom row) ECG signal. After filtering, the baseline is stablized at zero.

### 4.3 Feature Selection

We collect a large amount of features that are used in literatures from four channels (GSR, SKT, ECG and HR) to offer a superset of potential good features. We observe that not all of these 89 features are useful for our EDL task defined in Section 4.4; that is, some features may weakly correlate with the emotion distribution, and some features may be highly correlated with each other and thus redundant. Therefore, it is necessary to perform a feature selection operation.

The feature selection algorithms can be broadly categorized into filter and wrapper methods[46]. Compared to filter methods, wrapper methods usually perform better[47]. Some representative wrapper methods include genetic algorithm (GA), sequential forward selection (SFS) and sequential backward selection (SBS). Following [46], we use the SBS method due to its superior performance. In particular, we pay attention to a specific SBS method called *recursive feature elimination* (RFE), which ranks features by recursively removing the feature that has the smallest ranking criterion computed by a trained classifier [48].

In order to determine the optimal number of selected features, we proposed a feature selection strategy by combining linear regression RFE (RFE_LR) and support vector machine RFE (RFE_SVM). RFE_LR analyzes the correlation between features and labels based on a regression model which is appropriate in our EDL task, because the labels of emotion distribution learning are density distributions rather than discrete values. On the other hand, in order to highlight the effect of the dominant emotion for describing the emotional state, we employ a classifier based RFE (i.e., RFE_SVM) as an auxiliary discriminator to feature selection. By taking advantages of these two different representations, we propose the following feature selection strategy.

Denote the set of all features as $F$. Let $F_t^L$ and $F_t^S$ be subsets of $F$ which contain the top-$t$, $t \in \{1, 2, \ldots, 89\}$, features ranked by RFE_LR and RFE_SVM respectively, and $S_t = F_t^L \bigcap F_t^S$. We use $F_{t^*}^L$ as the set of selected features, where $t^*$ is the minimal value of $t$ that satisfies $|S_t|/|F_t^L| \geq \gamma$, where $|\cdot|$ is the cardinality of the set and $\gamma$ is a threshold. Note that since RFE_LR is more appropriate than RFE_SVM in our EDL task,

we use $|F_t^L| = t$ as a divisor and the constraint can be rewritten as $|S_t|/t \geq \gamma$. A large $\gamma$ may select insignificant features, while a small $\gamma$ may not be able to select enough significant features. Once the $\gamma$ is determined, we finally select $t^*$ features in our feature selection procedure. More details of the selected features are studied in Sections 5.2 and 5.3.

### 4.4 Emotion Distribution Learning

In this section, we formulate the emotion distribution learning (EDL) task and introduce baseline algorithms. In Section 4.5, based on the convolution neural network, we propose a novel EDL model for this task.

#### 4.4.1 Formulation of EDL

We generate 89 features and from them select $t^*$ important features to represent input peripheral physiological signals. Denote the feature space (each point in this space representing a $t^*$-dimension feature vector) as $s$ and the discrete emotion set as $y$, i.e., $y_j \in y$ is the $j$th emotion class, $j \in \{1, 2, \ldots, C\}$, where $C = 7$ denotes the total seven emotion classes in the dataset constructed in Section 3. For any feature vector $s_i \in s$ in the space $s$, we use $d_{s_i}^y$ to describe the intensity degree in the emotion set $y$ for this $s_i$, in particular, let the description degree of the $j$th emotion class for $s_i$ be $d_{s_i}^{y_j}$. Then the emotion distribution (in terms of description degrees) corresponding to $s_i$ is defined as $D_{s_i} = \{d_{s_i}^{y_1}, d_{s_i}^{y_2}, \ldots, d_{s_i}^{y_C}\}$. The distribution $D_{s_i}$ satisfies two constraints: (1) $0 \leq d_{s_i}^{y_j} \leq 1$; (2) $\sum_{j=1}^{C} d_{s_i}^{y_j} = 1$. We note that $d_{s_i}^{y_j}$ is the deterministic description degree (but not the statistical probability) of the $j$th emotion class for the feature vector $s_i$, while in the mathematical form, these two constraints allow the value of $d_{s_i}^{y_j}$ to be operated as probabilities.

Given the above notations and definitions, we can formulate the EDL problem as follows. Given a training set $S = \{(s_i, D_{s_i}) | i \in \{1, 2, \ldots, N\}\}$, where $N$ represents the number of training samples, our goal is to learn a mapping function that maps a specific feature vector $s_i$ to an emotion distribution $D_{s_i}$. In the form of mathematical description, we can regard $d_{s_i}^{y_j}$ as a conditional probability $p(y_j|s_i)$. Then the target of EDL is equivalent to learn the conditional probability mass function $p(y|s_i)$ from the training set $S$.

We suppose $p(y|s_i)$ is a parameter model, denoted as $p(y|s_i;\theta)$. Then the EDL problem is transformed into finding the optimal parameter $\theta^*$ such that given $s_i$, the parameter model can output a distribution which is as close as possible to $D_{s_i}$. We denote the output distribution $\{p(y_1|s_i;\theta), p(y_2|s_i;\theta), \ldots, p(y_C|s_i;\theta)\}$ as $\tilde{D}_{s_i}$, and denote the distance measure between two distributions as $Dist(\tilde{D}_{s_i}, D_{s_i})$. Then we can compute the optimal parameter $\theta^*$ by solving the following problem:

$$\theta^* = \arg\min_{\theta} \sum_i Dist(\tilde{D}_{s_i}, D_{s_i}). \tag{4}$$

#### 4.4.2 Baseline Algorithms

Several solutions to the classic label distribution learning (LDL) problem [33] can be extended to solve our EDL problem, and we treat them as the baseline algorithms which are compared in Section 5.

Three strategies were proposed in [33] to solve the LDL problem. The first strategy is called problem transformation (PT), which transforms an LDL problem into a single label learning (SLL) problem. Two typical methods in this strategy is PT-Bayes and PT-SVM, which solve the corresponding SLL problem by Bayes and SVM classifiers respectively. To extend these two methods to solve our EDL problem, each training sample $(s_i, D_{s_i})$ is first changed into $C$ single-label samples $(s_i, y_j)$ with weight $d_{s_i}^{y_j}$, $j = 1, 2, \ldots, C$ and $i = 1, 2, \ldots, N$. Then the training set is resampled according to the weight of each sample (by following the same resampling rules in [33]), and the size of the training set remains the same. The resampled training set can be used to train a classifier (e.g., Bayes or SVM classifier). To predict the emotion distribution of a feature vector $s_i'$, the learnt classifier needs to output the description degree for each emotion class $y_j$, $j = 1, 2, \ldots, C$. To adapt the PT-Bayes method, the posterior probability of each emotion class computed by Bayes rule can be regarded as the corresponding description degree. To adapt the SVM method, we use an improved implementation of Platt's posterior probabilities [49] to compute the probability of each binary SVM, and estimate the probability of each emotion class (served as description degree) by a pairwise coupling multi-class method [50].

The second strategy is to use algorithm adaptation and two adapted algorithms, namely AA-kNN and AA-BP, were used in [33]. To extend AA-kNN to solve the EDL problem, the description degree of $y_j$ for a feature vector $s_i'$ can be defined as the mean of the distribution of all the $k$ nearest neighbours as follow:

$$p(y_j|s_i') = \frac{1}{k} \sum_{i \in N_k(s_i')} d_{s_i}^{y_j}, (j = 1, 2, \ldots, C), \tag{5}$$

where $N_k(s_i')$ is the index set of the $k$ nearest neighbours of $s_i'$ in the training set. To make use of AA-BP in our EDL problem, the softmax activation function can be applied to each unit of the output layer of the AA-BP network. Denoting the $j$th output unit as $\xi_j$, the description degree can be computed as follows.

$$p(y_j|s_i') = \frac{\exp(\xi_j)}{\sum_{k=1}^{C} \exp(\xi_k)}. \tag{6}$$

The third strategy is to design specialized algorithms (SA) that directly solve the LDL or EDL problem. In this paper, we utilize two specialized algorithms for EDL which are SA-IIS [33] and SA-CPNN [32]. SA-IIS is based on the well-known optimization strategy called improved iterative scaling (IIS), which uses the maximum entropy model as the parametric model. The description degree in SA-IIS is defined as

$$p(y_j|s_i;\theta) = \frac{1}{Z} \exp\left(\sum_k \theta_{y_j,k} g_k(s_i)\right), \tag{7}$$

where $Z = \sum_{y_j} \exp(\sum_k \theta_{y_j,k} g_k(s_i))$ is the normalization factor, $\theta_{y_j,k}$ is the parameter corresponding to the emotion class $y_j$ and $k$th feature of the feature vector $s_i$, and $g_k(s_i)$ is the $k$th feature of $s_i$. Finally, we obtain the optimal parameter $\theta^*$ by maximizing the parameter model defined as follows.

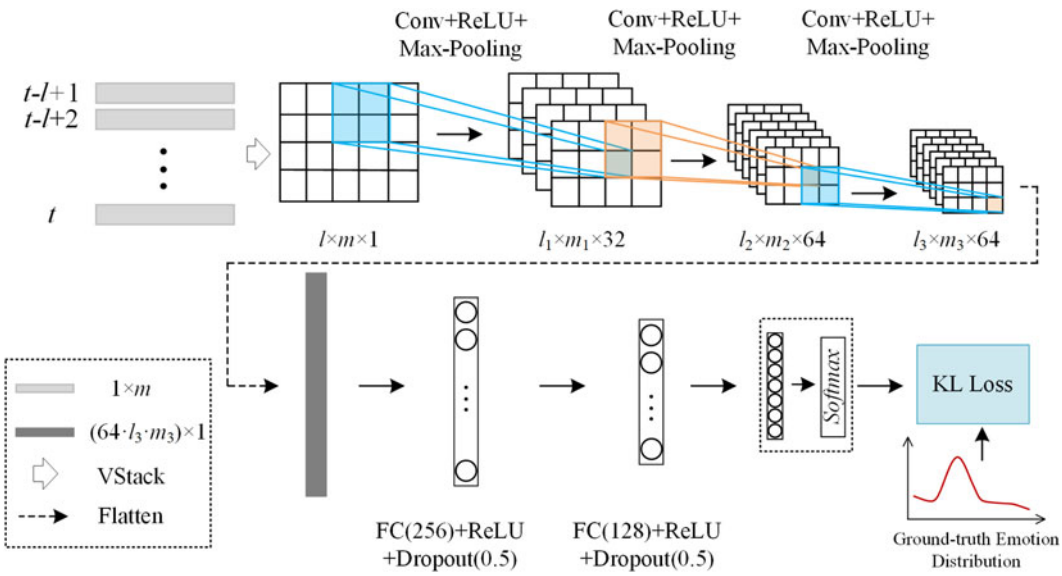



Fig. 3. The proposed EDL model. This model consists of three convolution layers, two dense layers and one softmax layer. $l$ represents the number of discrete time points in a continuous period of time, $m$ is the number of selected features in a feature vector. The value of parameters $l_i, m_i,\ i = 1, 2, 3$, can be determined given the specific configuration of the convolution kernel and stride size, etc.

$$\theta^* = \arg\max_{\theta} \sum_{i,j} d_{s_i}^{y_j} \left( \sum_k \theta_{y_j,k} g_k(s_i) - \ln \sum_t \exp\left( \sum_k \theta_{y_t,k} g_k(s_i) \right) \right). \tag{8}$$

SA-CPNN is a three layer neural network, which takes both feature vector $s_i$ and discrete emotion set $y$ as input. The description degree, which is the output of the network, is defined as

$$p(y_j|s_i;\theta) = \exp(b(\theta) + f(s_i, y_j;\theta)), \tag{9}$$

where $b(\theta)$ represents the bias, which ensures $\int p(s)ds = 1$, and is defined as

$$b(\theta) = -\ln\left( \sum_{y_j} \exp(f(s_i, y_j;\theta)) \right). \tag{10}$$

The Kullback-Leibler divergence is used as the distance measure in SA-CPNN. By utilizing the Eq. (4), we obtain the optimal parameter as

$$\begin{aligned} \theta^* &= \arg\min_{\theta} \sum_i \sum_j (d_{s_i}^{y_j}) \ln\left( \frac{d_{s_i}^{y_j}}{p(y_j|s_i;\theta)} \right) \\ &= \arg\max_{\theta} \sum_i \sum_j (d_{s_i}^{y_j}) \ln(p(y_j|s_i;\theta)) \\ &= \arg\max_{\theta} \sum_i \sum_j (d_{s_i}^{y_j})(b(\theta) + f(s, y_j;\theta)). \end{aligned} \tag{11}$$

### 4.5 A New CNN-Based EDL Model

In this section, we propose a new solution to the EDL problem formulated in Section 4.4.1. Different from the baseline algorithms summarized in Section 4.4.2, our solution considers both feature correlation and temporal cues by using a two-dimensional feature vector stacking. By stacking multiple consecutive feature vectors into a matrix form, our method can make use of two dimensional convolution, which has the advantage that each feature (e.g., the $i$th element in a feature vector) is aligned in the matrix (e.g., the $i$th column). Fig. 3 shows our solution which is a model based on the convolution neural network (CNN).

After the feature selection step in Fig. 2, $t^*$ features are selected and we use a $t^*$-dimension feature vector to store their values. Denote a $m$-dimensional feature vector at time $t$ as $X_t \in \mathbb{R}^m$, $m = t^*$. To take the temporal information into account, our model predicts the emotion distribution at time $k$, based on $l$ feature vectors $X_{k-l+1}, X_{k-l+2}, \ldots, X_k$ in a continuous period of time before $k$. We stack these $l$ feature vectors into a two-dimensional feature map $\mathcal{X}_t \in \mathbb{R}^{l \times m}$, such that each row is an $m$-dimensional feature vector. We use this feature map as input to our CNN model and then the convolutions of this feature map contain the information of both feature correlation and temporal cues.

In order to capture both local and global cues of the input feature map, we use 2D convolution (each convolution layer is followed by a Max-Pooling layer) and dense layers (each layer is a fully connected layer) to construct the network. As shown in Fig. 3, the proposed network architecture consists of three convolution layers, two dense layers and one softmax layer. Detailed configurations of each convolution layer and downsampling layer are listed in Table 1. We also use activation functions to introduce non-linearity into the model; in particular, we incorporate rectifier linear units (ReLUs) as the activation function for each convolution layer. The formulation of ReLU is as follow:

$$f(x) = \max(0, x). \tag{12}$$

After three convolution and downsampling operations, we obtain a set of feature maps, which will be further flattened into a vector before inputting them into dense layers. There are totally two dense layers, and they have 256 and 128 neurons, respectively. The activation functions used in these two dense layers are also ReLUs. In addition, we apply the

TABLE 1
Detailed Configurations of Convolution Layers and Max-Pooling Layers

| Layer | Input → Output Shape | Layer Information |
|---|---|---|
| Conv1 | $(l, m, 1) \rightarrow (l, m-2, 32)$ | (N32,K1×3, S1×3), ReLU |
| Max-Pooling1 | $(l, m-2, 32) \rightarrow (l, m-4, 32)$ | K1×3,S1×3 |
| Conv2 | $(l, m-4, 32) \rightarrow (l, m-6, 64)$ | (N64,K1×3, S1×3), ReLU |
| Max-Pooling2 | $(l, m-6, 64) \rightarrow (l, m-8, 64)$ | K1×3,S1×3 |
| Conv3 | $(l, m-8, 64) \rightarrow (l, m-10, 64)$ | (N64,K1×3, S1×3), ReLU |
| Max-Pooling3 | $(l, m-10, 64) \rightarrow (l, m-12, 64)$ | K1×3,S1×3 |

*$l$ and $m$ are the same as in Fig. 3, N represents the number of output channels, K stands for the kernel size, S is the stride size.*

dropout strategy on each dense layer to randomly disconnect some neurons (for avoiding overfitting), and the dropout rate is set to be 0.5 in all our experiments. The output of the second dense layer is then passed to the softmax layer for emotion distribution prediction. The softmax layer in our model is a linear layer with $C = 7$ neurons (corresponding to seven emotion classes), and the output $o_i (i = 1, 2, \ldots, C))$ of these seven neurons are than normalized using the softmax function defined as follows:

$$\hat{d}_{\mathcal{X}_t}^{y_j} = \frac{\exp(o_j)}{\sum_{k=1}^{C} \exp(o_k)}, \tag{13}$$

where $C = 7$ is the number of the emotion classes, and $\hat{d}_{\mathcal{X}_t}^{y_j}$ indicates the description degree of the $j$th emotion class predicted from $\mathcal{X}_t$.

For the parameter learning, we use the Kullback-Leibler divergence (KLD) as the loss function to measure the similarity between the predicted distribution and the ground truth. Given the training set denoted as $(\mathcal{X}_t, D_{\mathcal{X}_t}), t \in \{1, 2, \ldots, N\}$, where $D_{\mathcal{X}_t} = \{d_{\mathcal{X}_t}^{y_1}, d_{\mathcal{X}_t}^{y_2}, \ldots, d_{\mathcal{X}_t}^{y_C}\}$, the EDL loss in the proposed model is formulated as follows:

$$\mathcal{L} = \frac{1}{N} \sum_{t=1}^{N} \sum_{j=1}^{C} d_{\mathcal{X}_t}^{y_j} \log \frac{d_{\mathcal{X}_t}^{y_j}}{\hat{d}_{\mathcal{X}_t}^{y_j}}. \tag{14}$$

Given the loss function (14), we can compute the optimal parameters $\theta^*$ by solving the optimization problem below:

$$\begin{aligned} \theta^* &= \arg\min_{\theta} \frac{1}{N} \sum_{t=1}^{N} \sum_{j=1}^{C} d_{\mathcal{X}_t}^{y_j} \log \frac{d_{\mathcal{X}_t}^{y_j}}{\hat{d}_{\mathcal{X}_t}^{y_j}} \\ &= \arg\min_{\theta} \frac{1}{N} \sum_{t=1}^{N} \sum_{j=1}^{C} d_{\mathcal{X}_t}^{y_j} \log d_{\mathcal{X}_t}^{y_j} \\ &\quad - \frac{1}{N} \sum_{t=1}^{N} \sum_{j=1}^{C} d_{\mathcal{X}_t}^{y_j} \log \hat{d}_{\mathcal{X}_t}^{y_j} \\ &= \arg\min_{\theta} - \frac{1}{N} \sum_{t=1}^{N} \sum_{j=1}^{C} d_{\mathcal{X}_t}^{y_j} \log \hat{d}_{\mathcal{X}_t}^{y_j}. \end{aligned} \tag{15}$$

We use the Adam algorithm to solve the above optimization problem and find the optimal parameters $\theta^*$. More details about the selection of parameter learning rate and $l$ can be found in Appendix C, available in the online supplemental material.

# 5 Experiments

To evaluate the performance of our proposed EDL model (Section 4.5), we compare it with six baseline algorithms which are introduced in Section 4.4.2, i.e., AA-BP [33], AA-kNN [33], PT-Bayes [33], PT-SVM [33], SA-IIS [33] and SA-CPNN [32]. To ensure the validity and fairness of the comparisons, we apply two evaluation protocols, whose details are as follows.

## 5.1 Evaluation Setup

*Evaluation Protocols.* We adopt two classic evaluation protocols in our comparisons.

*Subject-Dependent Evaluation.* For each subject, we split the video that he/she watched (along with the related peripheral physiological signals) into the training set $V_1$ and the testing set $V_2$ separately according to the following rules: (1) If the number of videos which induce the same target dominant emotion is larger than 1, we randomly choose one video into $V_2$, and the remaining videos are in $V_1$; (2) If there is only one video corresponding to the inducing target dominant emotion, we separate this video in two, and put the first half into $V_1$ and second half into $V_2$. In this way, the videos (watched by the same subject) that induced the same target emotion are evenly divided into the training and testing sets. Furthermore, our final results are averaged on the testing set $V_2$.

*Subject-Independent Evaluation.* We apply the leave-one-subject-out cross-validation for subject-independent evaluation. In each validation, we just use data of one subject as the testing set and data of the other 37 subjects as the training set. The final results are averaged on 38-fold cross-validation in which the data of each subject is used once for test.

*Evaluation Details.* After dividing the training and testing sets, we follow the steps introduced in Section 4 to obtain the selected features and corresponding distribution labels. In particular, we only use the training set to accomplish emotion distribution label generation, feature selection and EDL model training, both for subject-dependent and subject-independent evaluations. Considering that our method uses both feature correlation and temporal cues, for fair comparison, we also offer temporal cues in the baseline algorithms. In details, data is formed as a temporal sequence in baseline algorithms, i.e., samples in $l$ adjacent time units are concatenated into a new sequence along the feature dimension. We set the parameters of baseline methods the same as in previous work [32], [33] which has been demonstrated to be effective. We use the same parameter setting in each cross-validation iteration.

*Evaluation Metrics.* Since the output of the EDL task is a discrete distribution, the traditional evaluation indexes such as accuracy and precision-recall curve are no longer suitable. We followed [33] to use six distribution-based measures for quantitative evaluation in our experiments. Given two discrete distributions $D = \{d_1, d_2, \ldots, d_C\}$ and $\widehat{D} = \{\widehat{d}_1, \widehat{d}_2, \ldots, \widehat{d}_C\}$, where $C$ indicates the number of discrete types or classes, the six measures are summarized in Table 2, in which $\downarrow$ indicates "the smaller the better", and $\uparrow$ indicates "the larger the better". Four measures (i.e., Chebyshev, Clark, Canberra and Kullback-Leibler) are distance metrics and the other two measures (i.e., Consine and Intersection) are similarity metrics.

TABLE 2
Six Distribution Distance/Similarity Measures

| Measure | Type | Formula |
|---|---|---|
| Chebyshev ↓ | Dist | $Dis_1(D,\widehat{D}) = \max_j \lvert d_j - \widehat{d}_j \rvert$ |
| Clark ↓ | Dist | $Dis_2(D,\widehat{D}) = \sqrt{\sum_{j=1}^{C} (\frac{d_j - \widehat{d}_j}{d_j + \widehat{d}_j})^2}$ |
| Canberra ↓ | Dist | $Dis_3(D,\widehat{D}) = \sum_{j=1}^{C} (\frac{\lvert d_j - \widehat{d}_j \rvert}{d_j + \widehat{d}_j})$ |
| Kullback-Leibler ↓ | Dist | $Dis_4(D,\widehat{D}) = \sum_{j=1}^{C} d_j \ln \frac{d_j}{\widehat{d}_j}$ |
| Cosine ↑ | Sim | $Dis_5(D,\widehat{D}) = \frac{\sum_{j=1}^{C} d_j \widehat{d}_j}{\sqrt{\sum_{j=1}^{C} d_j^2} \sqrt{\sum_{j=1}^{C} \widehat{d}_j^2}}$ |
| Intersection ↑ | Sim | $Dis_6(D,\widehat{D}) = \sum_{j=1}^{C} \min(d_j, \widehat{d}_j)$ |

*The abbreviations of Dist and Sim stand for distance and similarity, respectively. ↓ indicates "the smaller the better", and ↑ indicates "the larger the better".*

Additionally, pair-sample $t$-tests are used to evaluate whether statistically significant differences existed in each distribution-based metric between our method and the best method in baseline algorithms. We use the $p$-value ($^{***}p < .001$, $^{**}p < .01$,$^{*}p < .05$) to measure the difference.

## 5.2 Subject-Dependent Evaluation

### 5.2.1 Quantitative and Qualitative Evaluation

As introduced in Section 4.3, we use RFE_LR and RFE_SVM in our EDL feature selection. In our subject-dependent evaluation, we partition the dataset into training and testing sets, and set $\gamma = 50\%$. The relation between values of $|S_t|/t$ and $t$ over all the training data is shown in Fig. 5, which indicates the optimal value of $t$ is $t^* = 50$. So in the testing process, we select the top 50 features from all 89 features.

Table 3 summarized the best results that our method achieved with $l = 10$. In addition to the six measures in Table 2, we also used the mean value of rankings of all six measures (denoted as *Average Rank*) to indicate an overall performance. In Table 3, the best performance in each measure was highlighted in boldface, and the corresponding ranking is given in parentheses. These results showed that our method outperforms all baseline algorithms.

To further explore the effectiveness of our proposed method, we visualized the emotion distributions predicted by all the methods and some examples were illustrated in Fig. 4. These qualitative results showed that our results (the last column) were the closest to the ground truth distribution.

### 5.2.2 Feature Selection Analysis and Ablation Study

In our proposed emotion distribution prediction system (Fig. 1), feature selection is an important step. Since in subject-dependent protocol, the training and testing sets are fixed, the effectiveness of feature selection can be clearly examined by using this protocol.

By applying the feature selection strategy (Section 4.3), 50 features were selected from 89 features extracted from four channels, and the selected features were listed in Table 5. Among the 50 selected features, there were 22 GSR features, 25 ECG features and 3 HR features. The feature selection ratios of the four channels from high to low were ECG (64.1%), GSR (56.4%), HR (42.9%) and SKT (0%). In addition, Table 5 also showed the rank of each feature in the parentheses, and the average ranks of selected GSR, ECG and HR features were 21.1, 26.6, 49.0 respectively. we did not give the average rank of SKT features since no SKT features were selected.

First, to verify the optimization of the feature number determined by the feature selection strategy, we conducted experiments using different feature numbers. The performance of our proposed EDL method when using 40, 50 and 60 features (in order to roughly cover the recommended range of the threshold $\gamma$) was shown in Table 7. The results showed that when using 50 features, our method achieved the best average rank as 1.2 and also achieved the best performance in five measure evaluations. More experiments about feature selection is referred to Appendix C, available in the online supplemental material.

To further validate the effectiveness of the feature selection step, we compared the performance of our proposed method and six baseline algorithms, with or without the feature selection strategy. In Table 4, we summarized the change values of six measures and the relative improvement ratios by adding the feature selection step, where the change value of each measure is defined as $S_{fs} - S_{all}$, $S_{fs}$ and $S_{all}$ representing measure values with and without feature selection respectively, and the relative improvement ratio of each measure is defined as $\frac{S_{fs} - S_{all}}{S_{all}}$. The results showed that the changes corresponding to measures Chebyshev, Clark, Canberra and Kullback-Leibler are all negative, while changes corresponding to measures Cosine and Intersection are all positive, indicating that all methods benefit from the feature

TABLE 3
Subject-Dependent Experimental Results of Our Method and Six Baseline Algorithms on Different Measures With 50 Selected Features, $l = 10$, Where $l$ is the Number of Consecutive Discrete Time Points in a Time Sequence Defined in Section 4.5

| Algorithm | Chebyshev ↓ | Clark ↓ | Canberra ↓ | Kullback-Leibler ↓ | Cosine ↑ | Intersection ↑ | Average Rank |
|---|---|---|---|---|---|---|---|
| AA-BP[33] | 0.2245 (6) | 1.0657 (6) | 2.4484 (6) | 0.4255 (6) | 0.7359 (5) | 0.6537 (5) | 5.7 (6) |
| AA-kNN[33] | 0.1919 (2) | 0.8819 (3) | 2.0920 (2) | 0.2931 (4) | 0.7739 (4) | 0.6870 (4) | 3.3 (3) |
| PT-Bayes[33] | 0.7568 (7) | 2.4055 (7) | 6.2638 (7) | 12.0230 (7) | 0.3862 (7) | 0.2182 (7) | 7.0 (7) |
| PT-SVM[33] | 0.2193 (5) | 1.0603 (5) | 2.4477 (5) | 0.4034 (5) | 0.7276 (6) | 0.6463 (6) | 5.3 (5) |
| SA-IIS[33] | 0.1970 (4) | 0.8849 (4) | 2.1441 (3) | 0.2626 (3) | 0.7979 (3) | 0.6898 (3) | 3.3 (3) |
| SA-CPNN[32] | 0.1940 (3) | 0.8808 (2) | 2.1452 (4) | **0.2413 (1)** | 0.8056 (2) | 0.6912 (2) | 2.3 (2) |
| Ours | **0.1857**$^{***}$ **(1)** | **0.8495**$^{***}$ **(1)** | **2.0154**$^{***}$ **(1)** | 0.2426$^{***}$ (2) | **0.8065**$^{***}$ **(1)** | **0.7049**$^{***}$ **(1)** | **1.2 (1)** |

*The difference in distribution prediction between our method and the best compared method are highlighted with stars in each measurement ($^{***}p < .001$, $^{**}p < .01$,$^{*}p < .05$).*

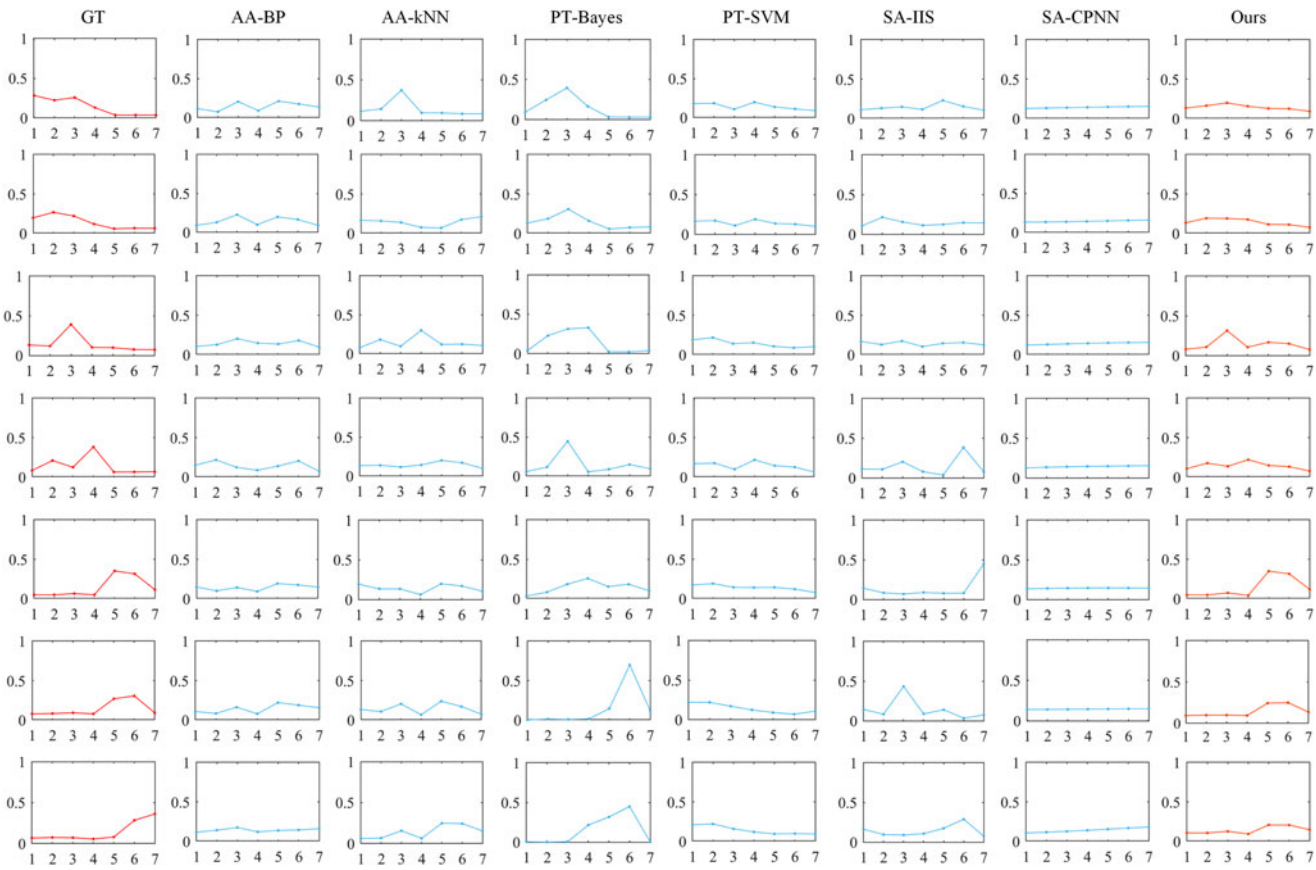



Fig. 4. Predicted emotion distributions using baseline algorithms and our proposed method in subject-dependent validation. The images in each column except the first one are result of algorithm specified by the text at top. "GT" indicates ground truth. The numbers 1 to 7 correspond to emotions anger, disgust, sadness, fear, tenderness, joy and amusement. The first four rows are corresponding to distribution predict results of four negative emotions (anger, disgust, sadness, fear), and the last three rows are corresponding to distribution predict results for three positive emotions (tenderness, joy, amusement).

selection. In particular, for the PT-Bayes method, all the measure values are improved by more than 10% (the minimal improvement is 13.99% for Cosine, and the maximum is 33.46% for Kullback-Leibler). For our proposed method, feature selection increase the value of Kullback-Leibler measure by 6.51%. All these results demonstrated the effectiveness of feature selection.

### 5.3 Subject-Independent Evaluation

Noting that the value of $t^*$ varies in each fold of cross-validation, we directly select $t$-top features in each validation with $t = 40, 50, 60$.

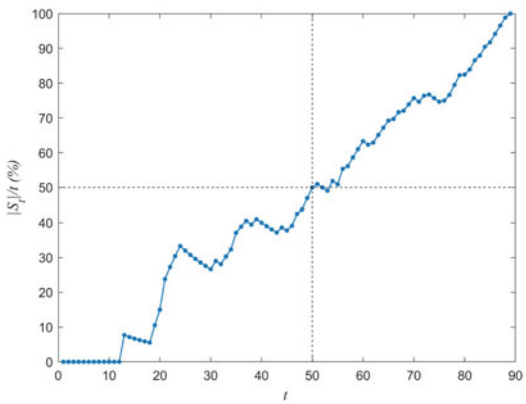



Fig. 5. Relations between values of $|S_t|/t$ and $t$ over all the training data in subject-dependent evaluation.

Table 6 shows the best results using our proposed method and six baseline methods, with $l = 10$, $t = 40$ ($t = 50$ and 60 are summarized in Table C3 in Appendix C), available in the online supplemental material. Though a large disparity exists among different subjects, our subject-independent results are close to the subject-dependent results shown in Table 3, which indicates the effectiveness of our method. We also study the effect of different feature numbers and different consecutive discrete time points in a time sequence in Appendix C, available in the online supplemental material. All the results show that among different combinations of feature numbers $t$ and $l$, our method can stably obtain good performance.

To further analyze feature selection results, we collect the common selected features shared in all folds of subject-independent evaluation with $t = 50$. These common selected features are summarized in Table 9. Only 20 GSR features and 9 ECG features are shared in all folds. These results are consistent with the observation that physiological signals have high variance between different subjects [51]. Both time and frequency features in GSR and ECG play important roles in subject-independent evaluation.

### 5.4 Comparison on Emotion Classification Task

Although our method is designed for the EDL problem, in this section, we show that our method also has good performance on the traditional single emotion label classification problem. We assign an emotion label to each feature instance in the training and test set using the dominant emotion in the

TABLE 4
Results of the Change Values and Relative Improvement Ratios (in Parentheses) of Six Measures, by Adding the Feature Selection Step

| Measure | AA-BP | AA-kNN | PT-Bayes | PT-SVM | SA-IIS | SA-CPNN | Ours |
|---|---|---|---|---|---|---|---|
| Chebyshev ↓ | −0.0029 (−1.50%) | **−0.0011 (−0.56%)** | −0.1340 (−32.63%) | −0.0004 (−0.19%) | −0.0010 (−0.43%) | −0.0001 (−0.05%) | −0.0007 (−0.38%) |
| Clark ↓ | −0.0144 (−1.59%) | −0.0042 (−0.47%) | −0.2592 (−17.16%) | −0.0046 (−0.51%) | −0.0478 (−4.85%) | −0.0075 (−0.85%) | −0.0034 (−0.40%) |
| Canberra ↓ | −0.0259 (−1.22%) | **−0.0119 (−0.56%)** | −0.6381 (−17.64%) | −0.0062 (−0.29%) | −0.1281 (−5.62%) | −0.0205 (−0.95%) | −0.0241 (−1.18%) |
| Kullback-Leibler ↓ | **−0.0132 (−4.77%)** | −0.0013 (−0.42%) | **−0.5475 (−33.46%)** | **−0.0057 (−1.97%)** | **−0.0515 (−13.06%)** | **−0.0054 (−2.20%)** | **−0.0169 (−6.51%)** |
| Cosine ↑ | +0.0065 (+0.82%) | +0.0015 (+0.20%) | +0.0797 (+13.99%) | +0.0043 (+0.56%) | +0.0047 (+0.64%) | +0.0046 (+0.57%) | +0.0098 (+1.23%) |
| Intersection ↑ | +0.0030 (+0.43%) | +0.0018 (+0.26%) | +0.0860 (+17.55%) | +0.0022 (+0.32%) | +0.0114 (+1.73%) | +0.0043 (+0.63%) | +0.0045 (+0.64%) |

ground truth emotion distribution, which leads to a single label dataset for seven-category classification.

We first compare our method with four representative classification algorithms which are SVM, Random Forest (RF), kNN and Deep Belief Network (DBN), using the subject-dependent setting. We use the scikit-learn library to implement SVM, RF and kNN. DBN is implemented in MATLAB based on DBNToolbox. For SVM, we set the max iteration to 50,000 and search the best penalty parameter C using logspace function in numpy from $10^{-10}$ to $10^{10}$. For Random Forest, we search the best estimator number in range [1, 50] and set the max depth parameter to 50. We test the performance of kNN using different number $k$ of nearest neighbors, and find that $k = 2$ reaches the best accuracy in our task. For DBN, we construct a two hidden-layer network and search different number of neurons (ranging from 10 to 90). We also consider different learning rates (ranging from 0.1 to 0.9) for both unsupervised and supervised training to better optimize the model.

TABLE 5
The 50 Selected Features From Four Channels

| Channel | Features |
|---|---|
| GSR (22) | standard deviation (15), variance (14), median (11), average (10), average of derivative for negative values only (44), proportion of negative samples in the derivative versus negative samples in all samples (1), number of local minima in GSR signal (19), average rising time (33), 3rd order moment (2), first degree difference (43), second degree difference (20), spectral power in 0-2.4Hz (34), max psd on 0-2.4Hz (21), min psd on 0-2.4Hz (18), variance of psd on 0-2.4Hz (22), mean of rise time (41), mean of latency (35), mean of amplitude (23), mean of EDAatApex (13), first derivative of rise time (17), first derivative of latency (16), first derivative of SCRWidth (12) |
| SKT (0) | - |
| ECG (25) | median (40), standard deviation (6), minimum (36), maximum (47), spectral power in 0.1-0.2Hz (32), power (5), spectral power in 0.2-0.3Hz (31), number of NNI (46), mean of NNI (27), minimum of NNI (26), maximum of NNI (30), mean of first degree difference of NNI (29), maximum of first degree difference of NNI (37), SDNN (38), RMSSD (3), SDSD (28), NN50 (24), pNN50 (25), NN20 (42), pNN20 (45), relative power of VLF (9), relative power of LF (7), relative power of HF (8), sd1 of Poincare' plot (4), sd2 of Poincare' plot (39) |
| HR (3) | standard deviation (48), spectral power in 0.15-1Hz (49), ratio of spectral power in 0.05-0.15Hz to spectral power in 0.15-1Hz (50) |

*For each feature, its rank is summarized in parentheses.*

The results are summarized in Table 8, showing that our method (i.e., Ours$_{de}$) has the best performance. Even for subject-independent protocol, our method can obtain the results (i.e., Ours$_{in}$) closed to Ours$_{de}$. Note that for a seven-category classification problem, the chance level is $1/7 = 14.29\%$ and the results in Table 8 are all slightly better than it. This is possibly because our dataset is originally designed for EDL problem and simply converting it into a dataset of single labels (by choosing the dominant emotion label whose intensity may be very close to other emotions in the distribution) makes the seven-category classification problem extremely challenging.

### 5.5 More Comparisons

*Comparisons With Random Tests*. To further verify the performance of our method, we conducted two random tests in two ways. First, we randomly shuffle the ground-truth distribution labels of the test set as the estimated distribution labels; we denote this case as R1. Second, we randomly shuffle the train data to scramble its origin timing order, meanwhile, the correspondence between training data and training labels is also disturbed; we denote this case as R2. The test of both cases is repeated 100 times in our subject-dependent and subject-independent evaluation. The averaged results are summarized in Table 10, showing that our method is much better than random guessing, because our method can learn useful relation between data and labels which is important in the EDL problem.

### 5.6 Discussion

In this paper, we proposed an EDL model that predicts emotion distribution from peripheral physiological signals. In our system, we extracted features from four channels (GSR, SKT, ECG and HR), and applied a feature selection strategy based on RFE. The study in [17] showed that ECG signals are significantly useful for correct classification of valence/arousal differentiation, and features from time or frequency domains of the HRV time series are decisive for classification of emotions including joy, anger, sadness and pleasure. Moreover, the experimental results in [52] indicated the positive correlation between skin conductance

TABLE 6
Subject-Independent Experimental Results of Our Method and Six Baseline Algorithms on Different Measures With $t = 40$, $l = 10$

| Algorithm | Chebyshev ↓ | Clark ↓ | Canberra ↓ | Kullback-Leibler ↓ | Cosine ↑ | Intersection ↑ | Average Rank |
|---|---|---|---|---|---|---|---|
| AA-BP[33] | 0.2156 (5) | 1.0134 (5) | 2.3642 (5) | 0.3842 (6) | 0.7531 (5) | 0.6622 (5) | 5.2 (5) |
| AA-kNN[33] | 0.1969 (2) | 0.8649 (2) | 2.0515 (2) | 0.2868 (4) | 0.7766 (4) | 0.6895 (2) | 2.7 (2) |
| PT-Bayes[33] | 0.7170 (7) | 2.3580 (7) | 6.0691 (7) | 8.9546 (7) | 0.3906 (7) | 0.2389 (7) | 7.0 (7) |
| PT-SVM[33] | 0.2267 (6) | 1.0265 (6) | 2.3925 (6) | 0.3815 (5) | 0.7272 (6) | 0.6474 (6) | 5.8 (6) |
| SA-IIS[33] | 0.2036 (4) | 0.8969 (4) | 2.2107 (4) | 0.2701 (3) | 0.7912 (3) | 0.6793 (4) | 3.7 (4) |
| SA-CPNN[32] | 0.2005 (3) | 0.8866 (3) | 2.2025 (3) | 0.2483 (2) | 0.8009 (2) | 0.6826 (3) | 2.7 (2) |
| Ours | **0.1910**$^{***}$ **(1)** | **0.8392**$^{***}$ **(1)** | **2.0319**$^{***}$ **(1)** | **0.2322**$^{***}$ **(1)** | **0.8139**$^{***}$ **(1)** | **0.7035**$^{***}$ **(1)** | **1.0 (1)** |

*The difference in distribution prediction between our method and the best compared method are highlighted with stars in each measurement ($^{***}p < .001$, $^{**}p < .01$,$^{*}p < .05$).*

response (SCR) and arousal, and emotions are discriminated by heart rate responses. Our results in Section 5.2.2 showed that the selection ratios of GSR (56.4%) and ECG (64.1%) were higher than those of HR (42.9%) and SKT (0%), and the average ranks of GSR and ECG features were smaller. Our results that the feature selection step selected features mostly from GSR and ECG is consistency with the experimental results in [52] and [17].

On the other hand, no SKT features were selected according to our experimental results (the ranks of SKT features were all greater than 50). However, it was reported in [53] that skin temperature difference can be used to distinguish emotions, and the change associated with anger was significantly different from that of all other emotions. Our results did not show the same observation. The reason may possibly be that the work [53] elicited target emotions from two tasks (which were directed facial action and relived emotions), while we used film clips.

*Limitation.* First, the cross-subject validation accuracy shows in Table 7 are worse than the intra-subject validation, indicating that cross-subject evaluation is a much more difficult task. Further efforts are needed to improve our performance on intra-subject validation. Second, our method is designed especially for emotion distribution learning, and when applying to emotion classification task, the accuracy is not high (ref. Table 8). So another direction in the future work is to improve the performance on the emotion classification problem.

TABLE 7
Results of Our Method Using Different Numbers of Selected Features

| Measure | Feature Number | | |
|---|---|---|---|
| | 40 | 50 | 60 |
| Chebyshev ↓ | 0.1857 (1) | 0.1857 (1) | 0.1867 (3) |
| Clark ↓ | 0.8491 (1) | 0.8495 (2) | 0.8495 (2) |
| Canberra ↓ | 2.0239 (2) | 2.0154 (1) | 2.0290 (3) |
| Kullback-Leibler ↓ | 0.2448 (2) | 0.2426 (1) | 0.2458 (3) |
| Cosine ↑ | 0.8060 (2) | 0.8065 (1) | 0.8057 (3) |
| Intersection ↑ | 0.7045 (2) | 0.7049 (1) | 0.7037 (3) |
| Average Rank | 1.7 (2) | **1.2 (1)** | 2.8 (3) |

*Numbers in parentheses indicate corresponding ranks.*

TABLE 8
Accuracy of of Each Algorithm on Emotion Classification Task

| Measure | SVM | RF | kNN | DBN | Ours$_{de}$ | Ours$_{in}$ |
|---|---|---|---|---|---|---|
| Acc(%) | 16.89 | 19.04 | 17.47 | 17.68 | **19.35** | **19.01** |

*Subscript $de$ denotes subject-dependent evaluation and subscript $in$ denotes subject-independent evaluation.*

## 6 Conclusion

In this paper, we studied the emotion distribution learning (EDL) task based on peripheral physiology signals and established a standardised dataset for this problem. To solve this EDL task, we proposed a system which consisted of four modules: data preprocessing, feature extraction, feature selection and emotion distribution prediction using a CNN-based deep learning model. Experiments were conducted on our constructed dataset to verify the effectiveness of our system, including the effectiveness of the feature selection strategy and the EDL model. In details, the feature selection strategy could select significant features and improve the performance

TABLE 9
Common Selected Features Shared in all Folds of Subject-Independent Evaluation

| Channel | Features |
|---|---|
| GSR (20) | standard deviation, variance, skewness, median, average, proportion of negative samples in the derivative versus negative samples in all samples, number of local minima in GSR signal, average rising time, 3rd order moment, second degree difference, max psd on 0-2.4Hz, min psd on 0-2.4Hz, sum of psd on 0-2.4Hz, variance of psd on 0-2.4Hz, mean of latency, mean of amplitude, mean of EDAatApex, first derivative of rise time, first derivative of latency, first derivative of SCRWidth |
| SKT (0) | none |
| ECG (9) | median, maximum of first degree difference of NNI, RMSSD, NN50, pNN50, relative power of VLF, relative power of LF, relative power of HF, sd1 of Poincare' plot |
| HR (0) | none |

TABLE 10
Results of the Two Random Tests

| Algorithm | Chebyshev ↓ | Clark ↓ | Canberra ↓ | Kullback-Leibler ↓ | Cosine ↑ | Intersection ↑ |
|---|---|---|---|---|---|---|
| $R1_{de}$ | 0.2234 | 0.9136 | 2.1777 | 0.4864 | 0.6704 | 0.6300 |
| $R2_{de}$ | 0.2590 | 0.9259 | 2.1936 | 0.2996 | 0.7438 | 0.6644 |
| $Ours_{de}$ | **0.1857**$^{***}$ | **0.8495**$^{***}$ | **2.0154**$^{***}$ | **0.2426**$^{***}$ | **0.8065**$^{***}$ | **0.7049**$^{***}$ |
| $R1_{in}$ | 0.2288 | 0.9199 | 2.1902 | 0.3964 | 0.6509 | 0.6257 |
| $R2_{in}$ | 0.2648 | 0.9118 | 2.2054 | 0.3113 | 0.7400 | 0.6609 |
| $Ours_{in}$ | **0.1910**$^{***}$ | **0.8392**$^{***}$ | **2.0319**$^{***}$ | **0.2322**$^{***}$ | **0.8139**$^{***}$ | **0.7035**$^{***}$ |

*Subscript $de$ denotes subject-dependent evaluation and subscript $in$ denotes subject-independent evaluation. The difference in distribution prediction between two random tests and our methods are highlighted with stars in each measurement ($^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$).*

of algorithms, and both quantitative and qualitative results of our EDL model showed the superiority of our method than six baseline methods.

## Acknowledgments

Yezhi Shu, Pei Yang and Niqi Liu have contributed equally to this work.

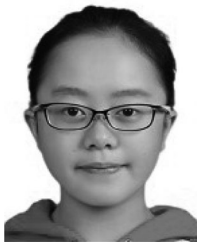
## References

[1] P. T. Young, *Motivation and Emotion: A Survey of The Determinants of Human and Animal Activity*. Hoboken, NJ, USA: Wiley, 1961.
[2] C. Mühl, B. Allison, A. Nijholt, and G. Chanel, "A survey of affective brain computer interfaces: Principles, state-of-the-art, and challenges," *Brain-Comput. Interfaces*, vol. 1, no. 2, pp. 66–84, 2014.
[3] P. Ekman, "The argument and evidence about universals in facial expressions," in *Handbook of Social Psychophysiology*, Hoboken, NJ, USA: Wiley, 1989, pp. 143–164.
[4] R. W. Levenson, P. Ekman, K. Heider, and W. V. Friesen, "Emotion and autonomic nervous system activity in the minangkabau of west sumatra," *J. Pers. Soc. Psychol.*, vol. 62, no. 6, 1992, Art. no. 972.
[5] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 374–393, Third Quarter 2019.
[6] E. L. Van den Broek, "Ubiquitous emotion-aware computing," *Pers. Ubiquitous Comput.*, vol. 17, no. 1, pp. 53–67, 2013.
[7] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Curr. Psychol.*, vol. 14, no. 4, pp. 261–292, 1996.
[8] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Develop. Psychopathol.*, vol. 17, no. 3, 2005, Art. no. 715.
[9] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *in Proc. Nat. Acad. Sci. USA*, vol. 114, no. 38, pp. E7900–E7909, 2017. [Online]. Available: https://www.pnas.org/content/114/38/E7900
[10] S.-T. Kousta, D. P. Vinson, and G. Vigliocco, "Emotion words, regardless of polarity, have a processing advantage over neutral words," *Cognition*, vol. 112, no. 3, pp. 473–481, 2009.
[11] P. J. Lang, "Emotion and motivation: Toward consensus definitions and a common research purpose," *Emotion Rev.*, vol. 2, no. 3, pp. 229–233, 2010.
[12] C. A. Frantzidis *et al.*, "On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 309–318, Mar. 2010.
[13] D. L. Strait, N. Kraus, E. Skoe, and R. Ashleya, "Musical experience promotes subcortical efficiency in processing emotional vocal sounds," *Neurosciences Music III: Disord. Plast.*, vol. 1169, 2009, Art. no. 209.
[14] M. M. Plichta *et al.*, "Auditory cortex activation is modulated by emotion: A functional near-infrared spectroscopy (fNIRS) study," *Neuroimage*, vol. 55, no. 3, pp. 1200–1207, 2011.
[15] A. J. Blood, R. J. Zatorre, P. Bermudez, and A. C. Evans, "Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions," *Nature Neurosci.*, vol. 2, no. 4, pp. 382–387, 1999.
[16] S. Koelsch, "Towards a neural basis of music-evoked emotions," *Trends Cogn. Sci.*, vol. 14, no. 3, pp. 131–137, 2010.
[17] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, Dec. 2008.
[18] P. G. Hunter, E. G. Schellenberg, and U. Schimmack, "Mixed affective responses to music with conflicting cues," *Cogn. Emotion*, vol. 22, no. 2, pp. 327–352, 2008.
[19] P. G. Hunter, E. G. Schellenberg, and Schimmack, Ulrich, "Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions," *Psychol. Aesthetics, Creativity, Arts*, vol. 4, no. 1, 2010, Art. no. 47.
[20] Y.-J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-time movie-induced discrete emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 550–562, Fourth Quarter 2018.
[21] M. K. Uhrig *et al.*, "Emotion elicitation: A comparison of pictures and films," *Front. Psychol.*, vol. 7, 2016, Art. no. 180.
[22] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cogn. Emotion*, vol. 9, no. 1, pp. 87–108, 1995.
[23] E. Ruiz-Padial and A. J. Ibáñez-Molina, "Fractal dimension of EEG signals and heart dynamics in discrete emotional states," *Biol. Psychol.*, vol. 137, pp. 42–48, 2018.
[24] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Ment. Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.
[25] H. Tang, W. Liu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep neural networks," in *Proc. Int. Conf. Neural Inf. Process.*, 2017, pp. 811–819.
[26] X. Du *et al.*, "An efficient LSTM network for emotion recognition from multichannel EEG signals," *IEEE Trans. Affective Comput.*, to be published, doi:10.1109/TAFFC.2020.3013711
[27] J. N. Bailenson *et al.*, "Real-time classification of evoked emotions using facial feature tracking and physiological responses," *Int. J. Hum.-Comput. Stud.*, vol. 66, no. 5, pp. 303–317, 2008.
[28] P. A. Kragel and K. S. LaBar , "Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions," *Emotion*, vol. 13, no. 4, 2013, Art. no. 681.
[29] Z. Zhang *et al.*, "Multimodal spontaneous emotion corpus for human behavior analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3438–3446.
[30] D.-A. Phan, H. Shindo, and Y. Matsumoto, "Multiple emotions detection in conversation transcripts," in *Proc. 30th Pacific Asia Conf. Lang. Inf. Comput.: Oral Papers*, 2016, pp. 85–94.
[31] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, Jun. 2017.
[32] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.
[33] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, Jul. 2016.
[34] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 3266–3272.
[35] J. Yang, M. Sun, and X. Sun, "Learning visual sentiment distributions via augmented conditional probability neural network," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 224–230.
[36] Y. Zhang, J. Fu, D. She, Y. Zhang, S. Wang, and J. Yang, "Text emotion distribution learning via multi-task convolutional neural network," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 4595–4601.

[37] Y. Ge, G. Zhao, Y. Zhang, R. J. Houston, and J. Song, "A standardised database of chinese emotional film clips," *Cogn. Emotion*, vol. 33, no. 5, pp. 976–990, 2019.
[38] G. Zhao, Y. Zhang, G. Zhang, D. Zhang, and Y.-J. Liu, "Multi-target positive emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, to be published, doi:10.1109/TAFFC.2020.3043135
[39] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective audio-visual words and latent topic driving model for realizing movie affective scene classification," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 523–535, Oct. 2010.
[40] C.-X. Ma, J.-C. Song, Q. Zhu, K. Maher, Z.-Y. Huang, and H.-A. Wang, "EmotionMap: Visual analysis of video emotional content on a map," *J. Comput. Sci. Technol.*, vol. 35, pp. 576–591, 2020.
[41] M. A. Tinati and B. Mozaffary, "A wavelet packets approach to electrocardiograph baseline drift cancellation," *Int. J. Biomed. Imag.*, vol. 2006, 2006, Art. no. 097157.
[42] G. M. Friesen, T. C. Jannett, M. A. Jadallah, S. L. Yates, S. R. Quint, and H. T. Nagle, "A comparison of the noise sensitivity of nine QRS detection algorithms," *IEEE Trans. Biomed. Eng.*, vol. 37, no. 1, pp. 85–98, Jan. 1990.
[43] M. Ahlstrom and W. Tompkins, "Digital filters for real-time ECG signal processing using microprocessors," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 9, pp. 708–713, Sep. 1985.
[44] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, First Quarter 2012.
[45] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 98–107, Jan. 2018.
[46] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Third Quarter 2014.
[47] P. Somol, J. Novovicová, and P. Pudil, "Efficient feature subset selection and subset size optimization," *Pattern Recognit. Recent Adv.*, vol. 56, pp. 75–97, 2010.
[48] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 389–422, 2002.
[49] H.-T. Lin, C.-J. Lin, and R. C. Weng, "A note on platt's probabilistic outputs for support vector machines," *Mach. Learn.*, vol. 68, no. 3, pp. 267–276, 2007.
[50] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, no. Aug, pp. 975–1005, 2004.
[51] D. Novak, M. Mihelj, and M. Munih, "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing," *Interacting Comput.*, vol. 24, no. 3, pp. 154–172, 2012.
[52] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm, "Looking at pictures: Affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, no. 3, pp. 261–273, 1993.
[53] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, 1983.

**Yezhi Shu** received the BEng degree from Shandong University, China, in 2019. She is currently working toward the PhD degree with the Department of Computer Science and Technology, Tsinghua University, China. Her research interests include affective computing, computer vision, deep learning algorithms and applications.

**Pei Yang** received the bachelor's degree from Inner Mongolia University, China, in 2009, and the master's degree from the Minzu University of China, China, in 2016. He is currently working toward the PhD degree with the Department of Computer Science and Technology, Tsinghua University, China. His research interests include emotion recognition and machine learning.



**Niqi Liu** is currently working toward the year-four undergraduate degree with the Department of Computer Science and Technology, Tsinghua University, China. Her research interests include emotion recognition and affective computing.

**Shu Zhang** received the master's degree from the Institute of Psychology, Chinese Academy of Sciences, China, in 2020. She is currently a research associate with the Department of Computer Science and Technology, Tsinghua University, China. Her research interesting include emotional computing and cognitive analysis.

**Guozhen Zhao** received the BS degree in industrial engineering, from Tianjin University, China, in 2007, and the MS and PhD degrees in industrial and systems engineering from the State University of New York, Buffalo, in 2009 and 2011, respectively. He is currently an associate professor with the Institute of Psychology, Chinese Academy of Sciences, China. His current research interest is the mathematical modeling of human cognition and performance, emotion recognition, emotional interaction and augmentation.

**Yong-Jin Liu** (Senior Member, IEEE) received the BEng degree in mechano-electronic engineering from Tianjin University, China, in 1998, and the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, in 2003. He is currently a tenured full professor with the Department of Computer Science and Technology, Tsinghua University, China. His research interests include pattern analysis, human-computer interaction and affective computing.