information

MDPI

Article

A Systematic Exploration of Deep Neural Networks for EDA-Based Emotion Recognition

Dian Yu * **and Shouqian Sun**

College of Computer Science and Technology, Zhejiang University, Hangzhou 310037, China; ssq@zju.edu.cn
* Correspondence: yudian329@zju.edu.cn

Received: 25 March 2020; Accepted: 14 April 2020; Published: 15 April 2020

Abstract: Subject-independent emotion recognition based on physiological signals has become a research hotspot. Previous research has proved that electrodermal activity (EDA) signals are an effective data resource for emotion recognition. Benefiting from their great representation ability, an increasing number of deep neural networks have been applied for emotion recognition, and they can be classified as a Convolutional Neural Network (CNN), a Recurrent Neural Network (RNN), or a combination of these (CNN+RNN). However, there has been no systematic research on the predictive power and configurations of different deep neural networks in this task. In this work, we systematically explore the configurations and performances of three adapted deep neural networks: ResNet, LSTM, and hybrid ResNet-LSTM. Our experiments use the subject-independent method to evaluate the three-class classification on the MAHNOB dataset. The results prove that the CNN model (ResNet) reaches a better accuracy and F1 score than the RNN model (LSTM) and the CNN+RNN model (hybrid ResNet-LSTM). Extensive comparisons also reveal that our three deep neural networks with EDA data outperform previous models with handcraft features on emotion recognition, which proves the great potential of the end-to-end DNN method.

Keywords: emotion recognition; electrodermal activity; deep neural network

1. Introduction

Robust information about the emotional state of a user is key to providing an empathetic experience during human–machine interaction (HCI) [1]. To make the interaction go well, it is important to ensure that the computers can understand the feelings of users through the interaction process [2]. In recent decades, emotion recognition has become a significant field in HCI, and it has been applied in a wide range of areas such as usability testing, development process improvement, enhanced website customization, and video games [3].

In the study of emotion recognition, different data resources have been applied, including facial and body expressions, eye gaze, audio signals, physiological signals (ECG, EEG, and EDA/GSR), respiration amplitude, and skin temperature [4]. Among these data resources, physiological signals have been paid more attention in studies recently, as they can reflect the emotional states objectively, while expressions and body motions can be influenced by subjective behavior and therefore misleading. In the research on emotion recognition based on physiological signals, studies have come up with a subject-dependent method [5]. In this field, subject-dependent means that the source of data comes from the same person. After decades of research, the subject-dependent method has achieved an accuracy of more than 90% [6]. However, the subject-dependent model is not satisfying on robustness and universality, as the performance shows instability when moving from experiment participants to the general population [7]. Hence, some researchers have started to focus on the subject-independent method, where the data is acquired from multiple persons. Compared to the subject-dependent method, the great generality determines that the subject-independent approaches perform better on different subjects. Until now, researchers still haven't achieved a satisfying recognition accuracy [8–11]. Circumventing this problem, we make great attempts on subject-independent emotion recognition in this work.

Many studies in the past years have focused on physiological signals, such as EEG [12,13], ECG [14,15], and EDA [9,10]. Compared with other physiological signals, EDA can be measured on the skin surfaces of hands and wrists in a non-invasive way. Benefiting from this easy and efficient acquisition method, the EDA-based emotion recognition algorithm has a broad application prospect in the development of sensors, Internet of Things (IoT), and intelligent wearable devices. Moreover, EDA is controlled by the autonomic nervous system, which corresponds to the arousal state of people [16]. On the other aspect, EDA has fewer channels and data, compared to EEG signals. Therefore, making full use of the limited EDA data is a great challenge in the field of EDA-based emotion recognition.

The methods of physiological-signal-based emotion recognition can be classified into two types based on how features are extracted: Hand-crafted feature selection and auto feature extraction. In the first method, hand-crafted features can be extracted in the time domain, the frequency domain, the time-frequency domain, etc. [17]. After that, the hand-crafted features are fed into classifiers such as KNN [18] and SVM [19]. However, the formula of feature extraction is established manually, which means that it cannot extract other unknown important features. The method based on auto feature extraction can solve the defect of hand-crafted feature selection. It utilize deep learning networks, which can extract implicit and complex features automatically. In the field of emotion recognition based on physiological signals, an increasing amount of advanced research uses the auto feature extraction method rather than hand-crafted feature selection for its advantages mentioned above. Hence, we choose the auto feature extraction method utilizing deep learning model.

According to the ability of utilizing sequential messages, deep neural networks can be divided into Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). As EDA is a sequence in the time domain essentially, we intuitively conduct an RNN to mine the sequential relationships between different periods of EDA signals. Meanwhile, the CNN has achieved great performance in sequential classification tasks, such as video classification [20,21] and audio classification [22,23]. Therefore, it is necessary to compare CNN and RNN models as regards EDA-based emotion recognition. Considering the descriptive powers of CNNs and the ability to capture sequential features of RNNs, researchers have also combined the advantages of CNNs and RNNs to propose hybrid CNN+RNN networks, the effectiveness of which has been demonstrated in language identification [24], video-based emotion recognition [25], etc. To the best of the authors' knowledge, no previous study has tackled a systematic comparison of the three above-mentioned networks. Hence, we seek to compare the performance of CNN, RNN, and CNN+RNN on EDA-based emotion recognition in this paper.

To compare CNN, RNN, and CNN+RNN, we need to choose a typical structure for each of them. There are many CNN architectures, such as VGG [26], GoogLeNet [27], and ResNet [28]. Benefiting from the advantage of solving the degradation problem, ResNet has been broadly used in various tasks such as object detection [29] and image classification [30]. Thus, we choose ResNet as the backbone of our CNN model. To fit with the EDA signal, we adapt the original ResNet to replace the 2D convolution operation with 1D convolution operation. For RNN, we choose the most popular LSTM network and adapt it to fit our task. For consistency of the comparison, the CNN+RNN network is composed of ResNet and LSTM as well. To make full use of the representation ability of DNN, we apply CvxEDA [31] to decompose the dynamic and static changes from original EDA inputs (more details in Section 3.1).

In this work, we systematically compare three typical deep neural networks for EDA-based emotion recognition. To make full use of limited 1-channel EDA signals, we apply a novel EDA analysis method—the Convex Optimization-Based EDA method (CvxEDA) [31]—to decompose EDA into phasic and tonic components. After that, the three-channel EDA signals (the origin signal parallels with tonic and phasic signals) are, respectively, fed into ResNet, LSTM, and hybrid ResNet-LSTM for emotion classification. To fit with our task, the dimension of the convolutional operation in the original ResNet is changed from 2D to 1D, which can directly process the sequential EDA signal. We evaluate our models in the commonly used open-source dataset MAHNOB-HCI for three-class classification. The extensive ablation experiments are used to systematically compare the performance between the different structures of three types of deep neural networks, the results of which also demonstrate the superiority of the deep neural network by comparison with previous methods in the MAHNOB-HCI dataset.

The rest of this article is structured as follows: Section 2 reviews the related works, including the theories of emotion models and EDA analysis. Section 3 presents our proposed method. Section 4 discusses the results, and Section 5 presents the conclusions.

2. Related Work

2.1. Theories of Emotion Models

Emotion is a complex concept and has many different descriptions when emphasizing different aspects such as feelings of arousal and/or hedonic value, appraisal and/or labeling processes, external emotion generating stimuli, and the relationship between emotion and motivation [32]. One definition [32] written by Theodore D. Kemper in 1978 is that "emotion is a relatively short-term evaluative response essentially positive or negative in nature involving distinct somatic (and often cognitive) components."

For further study, researchers built emotion models to evaluate people's emotions in qualitative and quantitative ways. The main emotion models can be classified into discrete, dimensional, dynamical, and appraisal models. Discrete models consider emotions to be discrete and mixable. The color wheel model proposed by Plutchik [33] is a typical discrete model (see Figure 1). In this model, emotions are just like colors. There are eight basic emotions (joy, trust, fear, surprise, sadness, disgust, anger, and anticipation) in a ring order, which is analogous to the primary colors. Other emotions can be created by mixing two basic emotions. For example, the mixture of "joy" and "trust" creates a new emotion "love", just like the color orange can be acquired from mixing red and yellow.

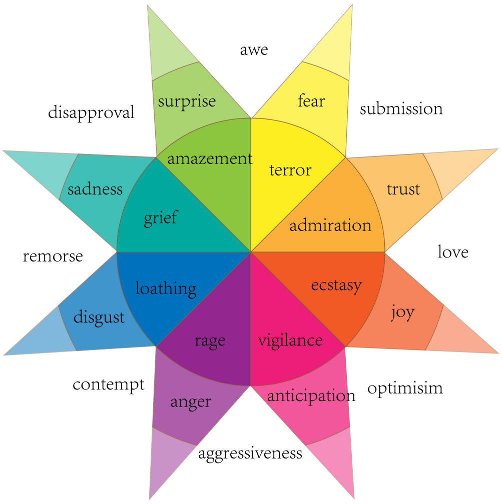

Figure 1. Color wheel model proposed by Plutchik.

Dimensional models are very different from discrete models. In the theory of dimensional models, emotions such as joy, trust, and surprise all have strong intrinsic connections within each other, rather than being isolated, and have some common features that can be measured quantitatively. Based on this theory, dimensional models appear as coordinate systems with two dimensions or multiple dimensions; each axis (or dimension) represents a feature, and different emotions are at different positions of the coordinate system. Figure 2 shows the valence and arousal model proposed by

Russel [34], which is the most representative dimensional model and is most used in the study of emotion recognition. Russel's model is quantized by the discrete values in the 2D coordinate of valence and arousal. In this model, the valence axis represents the level of positive or negative, where high valence means pleasant and low valence means unpleasant, and the arousal axis represents the intensity of the emotion, where high arousal means activation and low arousal means deactivation. For further explanation, emotions such as stressed and nervous show high valence and low arousal, while excited shows both high valence and high arousal; depressed shows both low valence and low arousal, while relaxed shows high valence and low arousal.

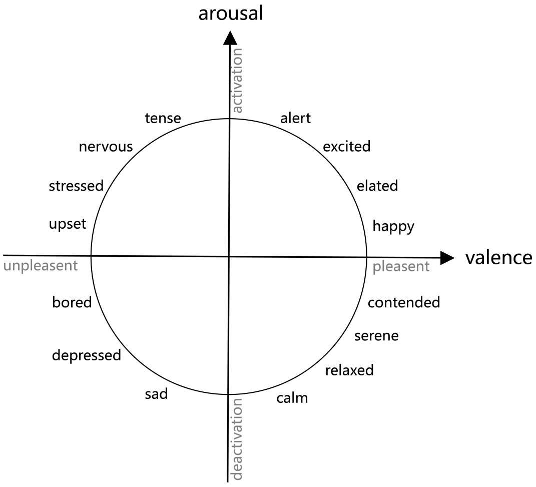

Figure 2. Valence and arousal model proposed by Russel.

The valence and arousal model established a quantifiable evaluation index for emotions, and our work is based on this model.

2.2. EDA Analysis

EDA is a general term for the autonomic changes in the electrical properties of the skin [35]. Electrodermal activities are formed and influenced by many complex factors. According to neurophysiology and psychophysiology research, EDA reflects the sweat gland activities on the skin surface, which is innervated by the autonomic nervous system (mainly the sudomotor nerves) [35]. The underlying reason for the connection between EDA and sweat gland activities is that sweat on the skin surface can change the electrical conductivity of the skin. Furthermore, the sweat gland activities is influenced by people's emotional states and changes. One simple example in our daily experience is that people sweat when they are scared and nervous. These findings prove that human emotional states can indeed be speculated from EDA signals, and the theoretical foundation for the study of emotion recognition can therefore be conducted based on EDA.

Generally researchers use skin conductance (SC), a representative manifestation of EDA, as the parameter measured in experiments. SC activity is composed of two parts: Tonic and phasic activity. The tonic signal is a slowly changing baseline that is caused by the drifting skin conductance level (SCL) and other unconscious activities. The phasic signal, also known as skin conductance response (SCR), is a quick response caused by external stimuli emotion (e.g., changes in emotional states). The EDA data that researchers collect from the experiment equipment are the origin SC signals, which need to be decomposed into tonic and phasic signals.

For the process of EDA, there are model-based methods, including discrete deconvolution analysis (DDA), continuous decomposition analysis (CDA), and convex optimization-based electrodermal activity (CvxEDA). DDA uses nonnegative deconvolution to decompose SC data into discrete compact responses [36], and CDA outperforms DDA by establishing a continuous measure that reflects the origin signal more closely [37]. CvxEDA is a more advanced method brought up by A. Greco et al.

recently in 2016 [31]. As it was shown by Greco et al. that CvxEDA has a stronger correlation and discriminant ability than CDA [35], we choose cvxEDA to decompose the EDA signal in this work.

3. Materials and Methods

In this work, we attempt to use three typical deep neural networks: CNN (ResNet), RNN (LSTM), and CNN+RNN (ResNet-LSTM) to promote the recognition baseline on the MAHNOB-HCI dataset and systematically compare the performance between them. Before being fed into the DNN models, the origin signal was decomposed into tonic and phasic signals using the CvxEDA method. We then divided the signals equally into three networks, respectively, for training. The DNN models output the prediction probability of three classes with a nonlinear function method and choose the class with the maximum probability as the final result. Compared with the corresponding ground truth labels, we obtained the performance of models with evaluation metrics.

3.1. CvxEDA

The convex optimization-based electrodermal activity (CvxEDA) [31] is used to decompose the origin EDA signal into tonic and phasic signals. The model of CvxEDA can be written as

$$y = r + t + \epsilon, \tag{1}$$

where y represents the origin EDA signal, r represents the phasic signal, t represents the tonic signal, and ϵ represents the noise produced by measurement and modeling errors.

The tonic signal t is composed of B-spline functions as well as the offset and linear trend term

$$t = Bl + Cd, \tag{2}$$

where B-spline functions make up the columns of the matrix B, l represents the spline coefficients, C represents a Nx2 matrix ($C_{i,1} = 1$, $C_{i,2} = i/N$), and d is a vector that represents the offset and linear trend.

The phasic signal can be modeled by the Bateman function

$$h(\tau) = (e^{-\frac{\tau}{\tau_0}} - e^{-\frac{\tau}{\tau_1}})u(\tau), \tag{3}$$

where τ_0 and τ_1 represent the slow and fast time constants, respectively, and $u(\tau)$ represents the step function. We transform Equation (3) through the Laplace transform and replace s as

$$s = \frac{2}{\delta}\frac{z-1}{z+1}, \tag{4}$$

where δ is the sampling interval. Following the ARMA model, we can finally obtain the phasic expression of discrete-time approximation as

$$r = MA^{-1}p, \tag{5}$$

where M ($M_{i,i} = M_{i,i-2} = 1$, $M_{i,i-1} = 2, 3 \leq i \leq N$) and A ($A_{i,i} = \psi$, $A_{i,i-1} = \theta$, $M_{i,i-2} = \xi, 3 \leq i \leq N$, ψ, θ and ξ are constants calculated from τ_0, τ_1 and δ) represent the tridiagonal matrix, respectively, and p is the activity of the autonomic nervous system.

According to Equations (2) and (5), Equation (1) can be written as

$$y = MA^{-1}p + Bl + Cd + \epsilon. \tag{6}$$

Using the maximum a posteriori estimation, the parameters p, l, and d can be represented as

$$[p, l, d] = \underset{p,l,d}{\operatorname{argmax}}\, P(p, l, d|y). \tag{7}$$

According to Bayes' theorem, we have

$$P(p, l, d|y) \propto P(y|p, l, d)P(p)P(l)P(d), \tag{8}$$

where p can be written as a Poisson distribution [38], l as a normal distribution, and $P(p, l, d|y)$ as an error model, while $P(d)$ is discarded from further computations. Therefore, we have the transformed equation as

$$\ln P(p, l, d|y) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(MA^{-1}p + Bl + Cd - y)_i^2 - \frac{1}{\lambda\delta}\sum_{i=1}^{N}(p)_i - \frac{1}{2\sigma_l^2}\sum_{i=1}^{Q} l_i^2 + \text{const.} \tag{9}$$

Finally, Equation (9) can be written into an optimization problem:

$$\text{minimize } \frac{1}{2}\|MA^{-1}p + Bl + Cd - y\|_2^2 + \alpha\|p\|_1 + \frac{\gamma}{2}\|l\|_2^2, \text{ subj. to } p \geq 0. \tag{10}$$

The optimization problem of Equation (10) can be easily solved by many existing methods. More details can be seen in [31].

3.2. ResNet

The ResNet [28] was brought up by K He et al. In 2015. ResNet is an excellent deep convolutional neural network, broadly used in image recognition and image feature extraction. ResNet mainly solves the degradation problem in convolutional neural networks with deeper layers (such as VGG) by presenting a residual learning block. As shown in Figure 3, the residual learning block has an "identity shortcut connection" to skip some layers and feed the data straight into the next layer.

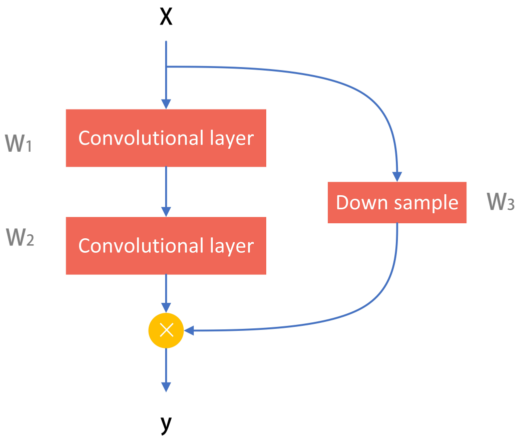

Figure 3. Residual learning block.

The output of the block can be represented as

$$y = W_2((W_1(x))) + W_3(x), \tag{11}$$

where x is the input, W_1 and W_2 are the convolutional layers, and W_3 is the downsampling operation.

Our model is adapted from the backbone of ResNet and we substitute a one-dimensional convolution for a two-dimensional convolution in ResNet in the area of emotion recognition based on

EDA. ResNet and many other deep convolutional neural networks are brought up in the image domain with two-dimensional convolution operation for image feature extraction. If applying the original 2D ResNet, the 1D input signal should be rearranged to a 2D matrix, like in [11]. However, the 2D convolutional kernel will disturb the sequential distribution to extract features. As the physiological signals such as EDA and EEG are one-dimensional sequences, we replaced the two-dimensional convolution into a one-dimensional convolution to process input signal directly and conveniently. We use the one-dimensional ResNet to extract features and feed the feature vector into a regression layer to obtain the prediction of three classes. The structure of our ResNet is shown as Table 1.

Table 1. The structure of our one-dimensional ResNet. Each convolution layer in this network is followed with a batch normalization operation.

Layer Name	Input Channel	Output Channel	Kernel Size	Stride	Padding	ReLU
Conv_original	3	64	7 × 1	2	3	Y
MaxPool1	64	64	3 × 1	2	1	N
ConvGroup1_X	64 64 64	64 64 64	3 × 1 1 × 1 3 × 1	2 2 1	1 0 1	Y N Y
ConvGroup2_X	64 64 128	128 128 128	3 × 1 1 × 1 3 × 1	2 2 1	1 0 0	Y N Y
ConvGroup3_X	128 128 256	256 256 256	3 × 1 1 × 1 3 × 1	2 2 1	1 0 1	Y N Y
ConvGroup4_X	256 256 512	512 512 512	3 × 1 1 × 1 3 × 1	2 2 1	1 0 1	Y N Y
ConvDown	512	512	1 × 1	3	0	N
MaxPool2	512	512	14 × 1	1	0	N
Regression	Linear 256 × 1 Linear 3 × 1 SoftMax					Y Y N

3.3. LSTM Network

One of the special RNN structures is the LSTM. It has a more complex structure to solve the problem of gradient vanishing in a conditional RNN. It is capable for capturing information that has connections between long distances and is especially suitable for long-sequence feature extraction. Figure 4 shows the structure of LSTM. It has a repeating module for recurrence, as do all other RNNs, but is much more complex.

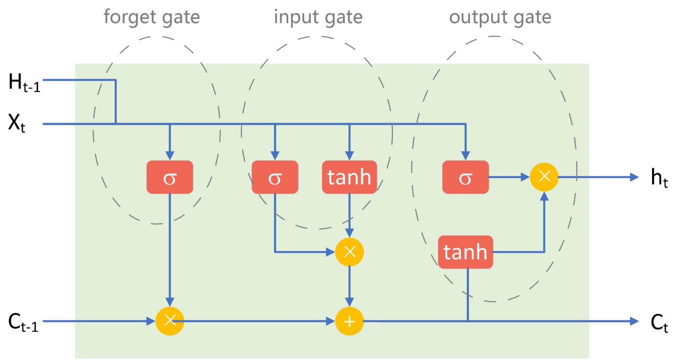

Figure 4. Structure of the repeating LSTM module.

Each module contains a forget gate, an input gate, an output gate, and a cell state (c_t). The cell state is the core of LSTM. It runs like a belt through the repeating modules: It receives messages from the previous cell state, adjust the messages, and then delivers them to the next cell state. The gates are used to control the delivery process and filter the messages in the network by using a sigmoid function. The forget gate decides how much information that the cell state receives from the previous one should be maintained. The maintained information can be written as

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{(t-1)} + b_{hf}), \tag{12}$$

where σ represents the sigmoid function, W_{if} and W_{hf} represent the weight coefficients, b_{if} and b_{hf} represent the biases, x_t represents the input, and h_{t-1} is both the state of the hidden layer and the output of the previous module. The input gate screens the information from the input and the previous output, and generates new information together with a tanh function. This process can be described as

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{(t-1)} + b_{hi}) \tag{13}$$

$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{(t-1)} + b_{hg}), \tag{14}$$

where i_t represents the input, g_t represents the result of the tanh function, and W and b, respectively, represent the weight coefficients and biases. Information from the forget gate and the input gate constitute a new cell state:

$$c_t = f_t c_{(t-1)} + i_t g_t, \tag{15}$$

where c_t represents the cell state. The output gate selects the information from the input and the previous output as well, and together with a tanh function decides the new output:

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{(t-1)} + b_{ho}) \tag{16}$$

$$h_t = o_t * \tanh(c_t), \tag{17}$$

where o_t represents the output gate, and h_t is the output and is sent to the next module as the hidden state.

The structure of our LSTM model is shown in Figure 5. The input EDA signal is fed into each LSTM module point by point, each point containing three channels (phasic, tonic, and origin). Each LSTM module outputs an eigenvector (also called the state of the hidden layer mentioned above) to the next. At the end of the final module, there is a regression layer that contains the stacked linear functions, ReLU functions, and the SoftMax function to predict the classification.

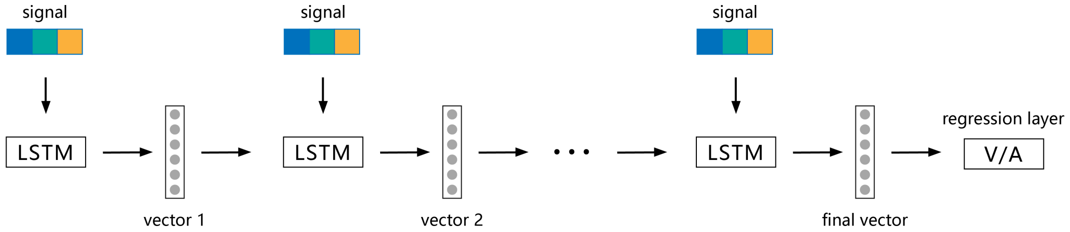

Figure 5. The LSTM model. The signal is fed into each LSTM module point by point, and each point contains three channels (phasic, tonic, and origin).

3.4. ResNet-LSTM

The ResNet-LSTM network combines both the descriptive power of ResNet and the sequential features capturing the ability of LSTM. The main structure of ResNet and LSTM in this hybrid network are essentially the same as the single ResNet and LSTM mentioned above. As illustrated in Figure 6, in this ResNet-LSTM network, the EDA signal is fed into ResNet as a three-channel input (phasic, tonic, and origin) first. The ResNet structure outputs a one-dimensional eigenvector and feeds it into the LSTM structure. Similar to the single LSTM, there is a regression layer containing a linear function at the end of the final LSTM module for three-class classification.

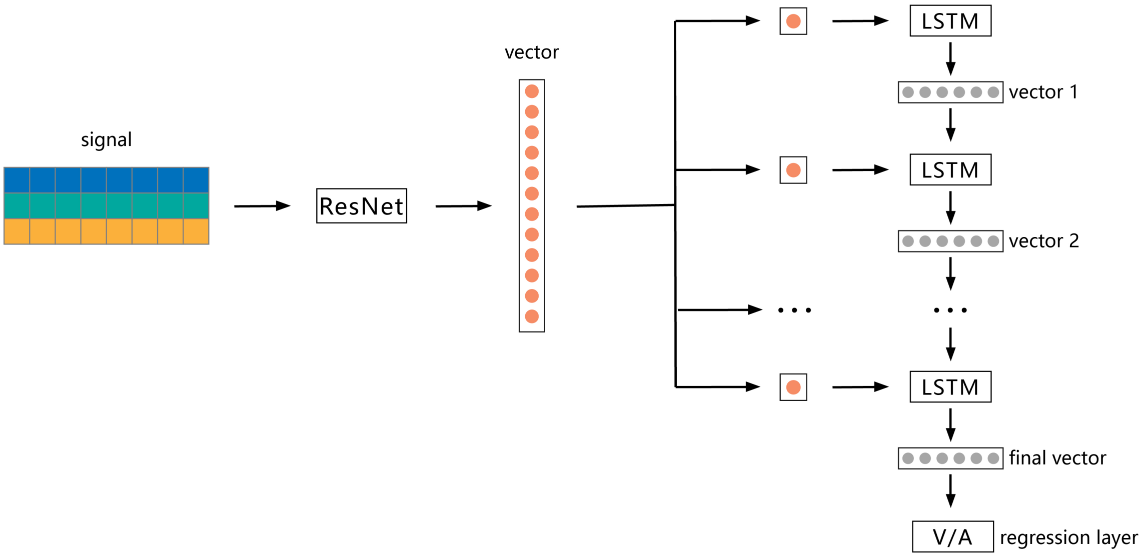

Figure 6. ResNet-LSTM model. The signal is first fed into ResNet as a three-channel input (phasic, tonic, and origin).

4. Implementation Details

In this work, we use a public benchmark dataset, MAHNOB-HCI, of EDA signals. The emotional states of MAHNOB-HCI are labeled with nine emotion keywords in valence and arousal dimensions. Following the baseline of the MAHNOB-HCI dataset [4], we relabeled nine emotional states to three classes, conducted classification with our models, and evaluated the performance with an average accuracy and F1 score. The details are described as follows.

4.1. Dataset

The MAHNOB-HCI [4] is a popular emotion dataset for affective computing. The 32-channel EEG and multiple physiological signals including EDA, ECG, RSP, and TMP signals were recorded from 30 participants in response to external stimulus (video and imagery) [4]. MAHNOB-HCI includes two experiments: An emotion recognition (also called emotion elicitation) experiment and an implicit tagging experiment. The data we used in this work are from the emotion elicitation experiment, in which 27 participants were asked to watch 20 emotional videos. To be more specific, in each sample, a neutral clip was shown first to relax the participant's emotion, and the emotional video was then played, after which the participant filled in the form. EDA (GSR) and other physiological signals were recorded 30 s before and after the emotional video. After watching each video, participants conducted a self-assessment with nine emotion keywords.

Referring to the previous baseline in [9,10], we utilized the valid samples downloaded from the dataset server in the "Selection of Emotion Elicitation" item. We followed the annotation strategy of [4,9,10] to relabel the nine annotations for three-class classification as shown in Table 2.

Table 2. Three emotional classes of MAHNOB-HCI on valence and arousal.

V/A	Label	Emotion Keywords
valence	1	fear, anger, anxiety, sadness, disgust
	2	neutral, surprise
	3	joy, amusement
arousal	1	sadness, disgust, neutral
	2	joy, amusement
	3	surprise, fear, anger, anxiety

4.2. Evaluation Metrics

Accuracy and $F1$ score are used as evaluation metrics in this work. In this subsection, we briefly introduce their definitions as follows.

Accuracy is calculated as

$$accuracy = \frac{N_{correct}}{N_{total}}, \tag{18}$$

where $N_{correct}$ is the number of samples classified correctly, and N_{total} is the number of total samples. In this work, accuracies are all finally calculated in averages, which are subject to 10-fold cross validation.

$F1$ score is based on precision and recall:

$$precision = \frac{TP}{TP + FP} \tag{19}$$

$$recall = \frac{TP}{TP + FN}, \tag{20}$$

where TP, FP, and FN represents true positive (predicted as positive and actually active), false positive (predicted as positive but actually negative), and false negative (predicted as negative but actually active).

As this is a three-class problem. there are $F1$ scores for each class and an average $F1$ score for the overall classifier:

$$F1score = \frac{2 \times precision \times recall}{precision + recall}, \tag{21}$$

and the overall average $F1$ score can be presented as

$$F1 = \frac{F1_1 + F1_2 + F1_3}{3}, \tag{22}$$

where $F1_1$, $F1_2$, and $F1_3$, respectively, represent the $F1$ score of each class.

5. Experiment

5.1. Training Setup

As described in the previous sections, we built three models based on ResNet, LSTM, and hybrid ResNet-LSTM network, respectively. To figure out the best configuration of each model, three ablation experiments about the structure settings were conducted. After establishing the best architecture of each DNN model, we conducted an analysis comparing the three DNN models to discover which model performs best on EDA-based emotion recognition. Finally, we compared the results of the DNN models in this work with the existing studies based on the MAHNOB-HCI dataset.

In our experiments, emotions are classified based on the three-class strategy mentioned in Table 2, and results are evaluated based on the metrics in Section 4.2. It should be emphasized that accuracies shown in this work were all subject to 10-fold cross validation. For consistency of the comparison, each model was trained for 25 epochs with the Stochastic Gradient Descent (SGD) optimizer. The initial learning rate was set to 0.001, which was decreased by multiplying it by 0.1 at every five epochs. The detailed experimental results and analyses of our experiments will be described as follows.

5.2. Configuration of ResNet

For ResNet, the configuration is addressed on the number of stacked residual learning blocks (Res-blocks) in one convolutional group, seen in Table 1. More residual blocks means a deeper network. The depth of the DNN model is one of the key factors that influence the performance of the network. Previous research has shown that a deeper network generally has a better performance in image classification [28]. However, the relationship between the depth of ResNet and its performance has not been studied in the field of EDA-based emotion recognition in the MAHNOB-HCI dataset. Moreover, the deeper model is followed with a higher computational expense and a slower processing speed. We seek to balance recognition performance with computing expense. Therefore, we conducted an ablation experiment to consider three different values (1, 2, and 3) for the number of stacked Res-blocks and compared the original 2D ResNet with our adapted 1D ResNet.

Table 3 shows the configuration sets of ResNet in this experiment, and Table 4 shows the experiment results. From the results, we can see that the adapted 1D ResNet is superior than the original 2D ResNet. Moreover, with the increase of the stacked Res-block number, the accuracies do not significantly improve. This reveals that the representation ability of the simplest ResNet is good enough to distinguish the EDA signals from different emotional states in the MAHNOB-HCI dataset. Considering that the deeper network will bring a greater cost of GPU memory and time, we choose the architecture with one Res-block in one convolutional group (Configuration ID 1) as the best architecture for ResNet.

Table 3. ResNet configuration sets.

Configuration ID	1	2	3	4	5	6
number of stacked Res-blocks	1	2	3	1	2	3
convolutional dimension	1D	1D	1D	2D	2D	2D

Table 4. The performances of different configurations of ResNet for valence and arousal. The average accuracies (Accu%) and F1 score (%) are subject to 10-fold cross validation.

Configuration ID		1	2	3	4	5	6
valence	**Accu**	86.73 ± 3.41	86.96 ± 4.53	87.14 ± 6.37	84.96 ± 4.12	84.14 ± 5.02	85.80 ± 3.31
	F1 score	85.72 ± 2.93	85.87 ± 3.75	86.02 ± 4.93	83.42 ± 3.01	82.99 ± 2.87	83.80 ± 2.15
arousal	**Accu**	86.92 ± 4.22	86.68 ± 3.67	87.07 ± 5.73	86.22 ± 4.35	86.53 ± 4.29	86.42 ± 4.03
	F1 score	85.96 ± 3.61	85.42 ± 4.13	85.83 ± 5.18	84.87 ± 3.05	85.02 ± 3.73	85.87 ± 4.25

5.3. Configuration of LSTM

The hidden cell dimension and the layer dimension are two adjustable configurations of LSTM. For LSTM in physiological signal analysis [39], common settings of the hidden cell dimension are 128 and 256, and layer dimensions can be deepened from 1 to 2 or 3. To determine the best LSTM structure, we designed the ablation experiment from a combination of the numbers of hidden cell and layer dimensions. Table 5 shows the configuration sets of LSTM in this experiment, and Table 6 shows the experiment results.

Table 5. LSTM configuration sets.

Configuration ID	1	2	3	4	5	6
hidden cell dimension	128	128	128	256	256	256
layer dimension	1	2	3	1	2	3

Table 6. The performances of different configurations of LSTM for valence and arousal. The average accuracies (Accu%) and F1 score (%) are subject to 10-fold cross validation.

Configuration ID		**1**	**2**	**3**	**4**	**5**	**6**
valence	**Accu**	76.42 ± 4.61	82.12 ± 2.36	80.45 ± 2.51	76.61 ± 3.93	82.65 ± 3.35	81.48 ± 3.74
	F1 score	74.59 ± 4.12	81.46 ± 2.67	78.71 ± 4.92	75.03 ± 4.95	80.84 ± 2.35	79.39 ± 4.21
arousal	**Accu**	66.85 ± 7.81	75.58 ± 5.68	74.32 ± 5.90	67.27 ± 7.36	75.61 ± 3.82	74.80 ± 4.55
	F1 score	65.83 ± 6.96	74.71 ± 3.47	71.86 ± 4.33	65.94 ± 5.74	74.44 ± 5.00	72.05 ± 6.31

Comparing the performances between the different configurations, we can see that LSTM with one layer dimension (Configuration ID 1 and 4) performs poorly, which reveals that one layer is too shallow for mining useful features from complex EDA signals; LSTM with three layer dimensions (Configuration ID 3 and 6) is not as good as LSTM with two layer dimensions (Configuration ID 2 and 5), which means that deeper LSTM leads to instability and increases training difficulty; LSTM with two layer dimensions (Configuration ID 2 and 5) balance representation ability and training difficulty to obtain the best performance. Configuration ID 5 achieves the highest accuracies in terms of both valence and arousal, but is very similar to Configuration ID 2. As mentioned in Section 5.2, we choose the simpler LSTM (Configuration ID 2) configured with 128 hidden cells and two layers as the best architecture for LSTM.

5.4. Configuration of Hybrid ResNet-LSTM

The previous two ablation experiments indicate that ResNet performs well with low stacked Res-block numbers, and LSTM shows better performances with 128 or 256 hidden cell dimensions and two layer dimensions. To avoid unnecessary repetition, we combined the better configurations of the single ResNet and LSTM to obtain four sets of configurations for the Hybrid ResNet-LSTM (see configuration sets in Table 7 and results in Table 8).

Table 7. Hybrid ResNet-LSTM configuration sets.

Configuration ID	**1**	**2**	**3**	**4**
number of stacked Res-block	1	1	2	2
hidden cell dimension	128	256	128	256
layer dimension	2	2	2	2

Table 8. The performances of different configurations of Hybrid ResNet-LSTM for valence and arousal. Accuracies (%) and F1 score (%) are subject to 10-fold cross validation.

Configuration ID		**1**	**2**	**3**	**4**
valence	**average accuracy**	82.53 ± 5.61	82.12 ± 4.21	81.97 ± 7.63	82.74 ± 5.37
	F1 score	80.82 ± 1.87	80.53 ± 1.12	80.19 ± 2.29	81.03 ± 2.39
arousal	**average accuracy**	81.34 ± 4.19	81.47 ± 7.87	80.66 ± 6.84	81.52 ± 4.50
	F1 score	80.70 ± 1.53	81.03 ± 3.67	80.26 ± 3.70	80.87 ± 4.03

Corresponding to the results in Sections 5.2 and 5.3, the testing accuracies and F1 scores of four configurations are nearly closed. Therefore, the simplest structure (Configuration ID 1) is chosen as the best architecture for hybrid ResNet-LSTM. The best architecture consists of the ResNet with one Res-block and the LSTM with two-layer and 128 hidden cells.

5.5. Comparison of ResNet, LSTM, and Hybrid ResNet-LSTM

After establishing the best architectures of the three deep neural networks, we can do some analysis on their performances. We directly utilize the respective experiment results of the three best architectures from Sections 5.2–5.4 to produce Table 9.

Table 9. The comparison of ResNet, LSTM, and Hybrid ResNet-LSTM model for valence and arousal. Accuracies (%) and F1 score (%) are subject to 10-fold cross validation. The best results are highlighted in bold.

V/A	Model	Average Accuracy	F1 Score
valence	**ResNet**	**86.73 ± 3.41**	**85.71 ± 2.93**
	LSTM	82.12 ± 2.36	81.46 ± 2.67
	ResNet-LSTM	82.53 ± 5.61	80.82 ± 1.87
arousal	**ResNet**	**86.92 ± 4.22**	**85.96 ± 3.61**
	LSTM	75.58 ± 5.68	74.71 ± 3.47
	ResNet-LSTM	81.34 ± 4.19	80.70 ± 1.53

Obviously, ResNet achieves the best performance among the three DNN models, which means that the CNN framework has a greater ability to mine dynamic and static features from decomposed EDA data compared to the RNN and hybrid CNN+RNN. Considering that our task is to predict a global emotional state with a sequential EDA input, the reason LSTM network achieves the poorest results is that it pays a great amount of attention to conducting sequential processing over time and ignores some global information. Moreover, as described in Section 5.3, the training difficulty of LSTM is also a strong factor. While the hybrid ResNet-LSTM combines the advantages of a CNN and RNN, the complicated hybrid architecture further promotes the training difficulty, which limits the model performance.

5.6. Comparison with Baselines in MAHNOB-HCI Dataset

We conduct some comparisons between the three DNN models in this work and the existing approaches for emotion recognition of three-class classification using physiological signals of the MAHNOB-HCI dataset to validate the effectiveness of the DNN-based methods. Considering that there are very few methods only using EDA signals in the MAHNOB-HCI database, we also involve approaches using other signals. Ferdinando et al. [9] utilize a KNN classifier with handcrafted features to complement baseline accuracies for the MAHNOB-HCI database. They further [10] promote emotion recognition using fused physiological features. Moreover, Liu et al. [40] apply the LSTM with the combination of EEG signals and external videos features. We compare the three DNN models (described in Section 5.5) with them, which can be seen in Table 10. For the three-class prediction task in MAHNOB-HCI, almost all of the three DNN models in this work outperform the existing methods in terms of the valence and arousal dimensions. Moreover, the ResNet model in this work improves the recognition accuracy significantly and performs much better than the previous methods, which validates the superiority of the deep convolutional neural network in this task.

Table 10. Comparison with the state-of-the-art three-class classification tasks in the MAHNOB-HCI dataset. Accuracies (%) and F1 score (%) are subject to 10-fold cross validation. The best results are highlighted in bold.

V/A	Author	Model	Signal	Average Accuracy	F1 Score
valence	Ferdinando et al. [41]	KNN	ECG+HRV	68.60 ± 4.40	-
	liu et al. [40]	LSTM	EEG+video	72.06	73.00
	Ferdinando et al. [9]	KNN	EDA	74.60 ± 3.80	-
	Ferdinando et al. [10]	KNN	EDA+HRV	79.60 ± 3.70	-
	Our	LSTM	EDA	82.12 ± 2.36	80.82 ± 2.67
	Our	ResNet-LSTM	EDA	82.53 ± 5.61	80.82 ± 1.87
	Our	**ResNet**	**EDA**	**86.73 ± 3.41**	**85.71 ± 2.93**
arousal	Ferdinando et al. [41]	KNN	ECG+HRV	70.70 ± 4.30	-
	liu et al. [40]	LSTM	EEG+video	74.12	72.30
	Our	LSTM	EDA	75.58 ± 5.68	74.71 ± 3.47
	Ferdinando et al. [9]	KNN	EDA	77.30 ± 3.60	-
	Ferdinando et al. [10]	KNN	EDA+HRV	77.70 ± 3.90	-
	Our	ResNet-LSTM	EDA	81.34 ± 4.19	80.70 ± 1.53
	Our	**ResNet**	**EDA**	**86.92 ± 4.22**	**85.96 ± 3.61**

6. Conclusions

In this work, we investigate three typical deep neural networks—ResNet, LSTM, and hybrid ResNet-LSTM—with respect to an EDA-based three-class emotion recognition task and systematically conduct ablation experiments to compare the different structures and their advantages. The comparison between the three DNN networks and the existing methods shows that the CNN-based ResNet in this work has the best performance: It has the best average accuracy (86.73%) and F1 score (85.71%) for valance, and the best average accuracy (86.92%) and F1 score (85.95%) for arousal in the MAHNOB-HCI dataset, which validates that the CNN framework is superior in EDA-based emotion recognition. The great performance of ResNet can make the following contributions: (1) The great representation ability of the end-to-end deep learning network can directly extract useful features from EDA signals; (2) the novel CvxEDA decomposition augments the one-channel EDA data to obtain phasic and tonic components, and the phasic and tonic signals can, respectively, reveal dynamic and static emotion changes. For further research, we will explore more novel CNN architectures to enhance the accuracy and generalization of models for EDA-based emotion recognition.

Author Contributions: Conceptualization: D.Y.; methodology: D.Y.; software: D.Y.; validation: D.Y.; formal analysis: D.Y.; investigation: D.Y.; data curation: D.Y.; writing–original draft preparation: D.Y.; writing–review and editing: D.Y. and S.S.; visualization: D.Y.; funding acquisition: S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (2017YFD0700102).

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. Picard, R.W.; Vyzas, E.; Healey, J. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 1175–1191. [CrossRef]
2. Fragopanagos, N.; Taylor, J.G. Emotion recognition in human–computer interaction. *Neural Netw.* **2005**, *18*, 389–405. [CrossRef] [PubMed]
3. Kołakowska, A.; Landowska, A.; Szwoch, M.; Szwoch, W.; Wrobel, M.R. Emotion recognition and its applications. In *Human-Computer Systems Interaction: Backgrounds and Applications 3*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 51–62.
4. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput.* **2011**, *3*, 42–55. [CrossRef]
5. Jerritta, S.; Murugappan, M.; Nagarajan, R.; Wan, K. Physiological signals based human emotion recognition: A review. In Proceedings of the 2011 IEEE 7th International Colloquium on Signal Processing and its Applications, Penang, Malaysia, 4–6 March 2011; pp. 410–415.
6. Bota, P.J.; Wang, C.; Fred, A.L.; Da Silva, H.P. A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals. *IEEE Access* **2019**, *7*, 140990–141020. [CrossRef]
7. Alzoubi, O.; D'Mello, S.K.; Calvo, R.A. Detecting Naturalistic Expressions of Nonbasic Affect Using Physiological Signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 298–310. [CrossRef]
8. Kim, K.H.; Bang, S.W.; Kim, S.R. Emotion recognition system using short-term monitoring of physiological signals. *Med. Biol. Eng. Comput.* **2004**. [CrossRef] [PubMed]
9. Ferdinando, H.; Alasaarela, E. Emotion Recognition using cvxEDA-Based Features. *J. Telecommun. Electron. Comput. Eng. (JTEC)* **2018**, *10*, 19–23.
10. Ferdinando, H.; Alasaarela, E. Enhancement of Emotion Recogniton using Feature Fusion and the Neighborhood Components Analysis. In Proceedings of the ICPRAM, Funchal, Portugal, 16–18 January 2018; pp. 463–469.
11. Machot, F.A.; Elmachot, A.; Ali, M.; Machot, E.A.; Kyamakya, K. A Deep-Learning Model for Subject-Independent Human Emotion Recognition Using Electrodermal Activity Sensors. *Sensors* **2019**, *19*, 1659. [CrossRef]

12. Petrantonakis, P.C.; Hadjileontiadis, L.J. Emotion recognition from EEG using higher order crossings. *IEEE Trans. Inf. Technol. Biomed.* **2009**, *14*, 186–197. [CrossRef]
13. Jirayucharoensak, S.; Pan-Ngum, S.; Israsena, P. EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *Sci. World J.* **2014**, *2014*, 627892. [CrossRef]
14. Wan-Hui, W.; Yu-Hui, Q.; Guang-Yuan, L. Electrocardiography recording, feature extraction and classification for emotion recognition. In Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, USA, 31 March–2 April 2009; Volume 4, pp. 168–172.
15. Ma, C.-w.; Liu, G.-y. Feature extraction, feature selection and classification from electrocardiography to emotions. In Proceedings of the 2009 International Conference on Computational Intelligence and Natural Computing, Wuhan, China, 6–7 June 2009; Volume 1, pp. 190–193.
16. Venables, P.H.; Christie, M.J. Electrodermal activity. *Tech. Psychophysiol.* **1980**, *54*.
17. Shukla, J.; Barreda-Angeles, M.; Oliver, J.; Nandi, G.; Puig, D. Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity. *IEEE Trans. Affect. Comput.* **2019**. [CrossRef]
18. Greco, A.; Valenza, G.; Citi, L.; Scilingo, E.P. Arousal and valence recognition of affective sounds based on electrodermal activity. *IEEE Sensors J.* **2016**, *17*, 716–725. [CrossRef]
19. Placidi, G.; Di Giamberardino, P.; Petracca, A.; Spezialetti, M.; Iacoviello, D. Classification of Emotional Signals from the DEAP dataset. In Proceedings of the International Congress on Neurotechnology, Electronics and Informatics, SCITEPRESS, Porto, Portugal, 7–8 November 2016; Volume 2, pp. 15–21.
20. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
21. Zha, S.; Luisier, F.; Andrews, W.; Srivastava, N.; Salakhutdinov, R. Exploiting image-trained CNN architectures for unconstrained video classification. *arXiv* **2015**, arXiv:1503.04144.
22. Hershey, S.; Chaudhuri, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B.; et al. CNN architectures for large-scale audio classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135.
23. Lee, J.; Kim, T.; Park, J.; Nam, J. Raw waveform-based audio classification using sample-level CNN architectures. *arXiv* **2017**, arXiv:1712.00866.
24. Bartz, C.; Herold, T.; Yang, H.; Meinel, C. Language identification using deep convolutional recurrent neural networks. In Proceedings of the International Conference on Neural Information Processing, Guangzhou, China, 14–18 November 2017; pp. 880–889.
25. Fan, Y.; Lu, X.; Li, D.; Liu, Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 445–450.
26. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
27. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
29. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: New York, NY, USA, 2015; pp. 91–99.
30. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: New York, NY, USA, 2012; pp. 1097–1105.
31. Greco, A.; Valenza, G.; Lanata, A.; Scilingo, E.P.; Citi, L. cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Trans. Biomed. Eng.* **2015**, *63*, 797–804. [CrossRef]
32. Kleinginna, P.R.; Kleinginna, A.M. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motiv. Emot.* **1981**, *5*, 345–379. [CrossRef]

33. Plutchik, R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Sci.* **2001**, *89*, 344–350. [CrossRef]
34. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [CrossRef]
35. Greco, A.; Valenza, G.; Scilingo, E.P. *Advances in Electrodermal Activity Processing with Applications for Mental Health*; Springer: Berlin/Heidelberg, Germany, 2016.
36. Benedek, M.; Kaernbach, C. Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology* **2010**, *47*, 647–658. [CrossRef] [PubMed]
37. Benedek, M.; Kaernbach, C. A continuous measure of phasic electrodermal activity. *J. Neurosci. Methods* **2010**, *190*, 80–91. [CrossRef] [PubMed]
38. Vogelstein, J.T.; Packer, A.M.; Machado, T.A.; Sippy, T.; Babadi, B.; Yuste, R.; Paninski, L. Fast nonnegative deconvolution for spike train inference from population calcium imaging. *J. Neurophysiol.* **2010**, *104*, 3691–3704. [CrossRef] [PubMed]
39. Spampinato, C.; Palazzo, S.; Kavasidis, I.; Giordano, D.; Souly, N.; Shah, M. Deep learning human mind for automated visual classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6809–6817.
40. Liu, J.; Su, Y.; Liu, Y. Multi-modal emotion recognition with temporal-band attention based on lstm-rnn. In Proceedings of the Pacific Rim Conference on Multimedia, Harbin, China, 28–29 September 2017; pp. 194–204.
41. Ferdinando, H.; Seppänen, T.; Alasaarela, E. Emotion Recognition Using Neighborhood Components Analysis and ECG/HRV-Based Features. In Proceedings of the International Conference on Pattern Recognition Applications and Methods, Porto, Portugal, 24–26 February 2017; pp. 99–113.