Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 3608115, 11 pages
https://doi.org/10.1155/2023/3608115

Research Article

Semisupervised Deep Features of Time-Frequency Maps for Multimodal Emotion Recognition

Behrooz Zali-Vargahan,[1] Asghar Charmin,[1] Hashem Kalbkhani,[2] and Saeed Barghandan[1]

[1]Department of Electrical Engineering, Ahar Branch, Islamic Azad University, Ahar, Iran
[2]Faculty of Electrical Engineering, Urmia University of Technology, Urmia, Iran

Correspondence should be addressed to Hashem Kalbkhani; h.kalbkhani@uut.ac.ir

Received 3 August 2023; Revised 15 September 2023; Accepted 26 October 2023; Published 13 November 2023

Academic Editor: Gennaro Vessio

Traditional approaches for emotion recognition utilize unimodal physiological signals. The effectiveness of such systems is affected by some limitations. To overcome them, this paper proposes a new method based on time-frequency maps that extract the features from multimodal biological signals. At first, the fusion of electroencephalogram (EEG) and peripheral physiological signal (PPS) is performed, and then, the two-dimensional discrete orthonormal Stockwell transform (2D-DOST) of the multimodal signal matrix is calculated to obtain time-frequency maps. A convolutional neural network (CNN) is then utilized to extract the local deep features from the absolute output of the 2D-DOST. Since there are uninformative deep features, the semisupervised dimension reduction scheme reduces them by balancing the generalization and discrimination. Finally, the classifier recognizes the emotion. The Bayesian optimizer finds the proper SSDR and classifier parameter values to maximize the recognition accuracy. The performance of the proposed method is evaluated on the DEAP dataset considering the two- and four-class scenarios through extensive simulations. This dataset consists of electroencephalograph (EEG) signals in 32 channels and peripheral physiological signals (PPSs) in eight channels from 32 subjects. The proposed method reaches the accuracy of 0.953 and 0.928 for two- and four-class scenarios, respectively. The results indicate the efficiency of the multimodal signals for detecting emotions compared to that of unimodal signals. Also, the results indicate that the proposed method outperforms the recently introduced ones.

1. Introduction

Emotion recognition is widely used in healthcare, teaching, human-computer interaction, and other fields. Since the physiological signals can reflect the real emotional state of an individual, they are widely used for emotion recognition. Single modality approaches extract the series of features from some channels. This approach cannot make full use of the relevant information among channels. Multimodal emotion recognition is an emerging interdisciplinary field of research in affective computing and sentiment analysis. It aims at exploiting the information carried by signals of different natures to make emotion recognition systems more accurate. This is achieved by employing a powerful multimodal fusion method [1].

This paper proposes an emotion recognition scheme based on multimodal signals consisting of electroencephalograph (EEG) and peripheral physiological signals (PPSs). The proposed method utilizes the two-dimensional discrete orthonormal Stockwell transform (2D-DOST) to consider the intramodal and cross-modal correlation among the multimodal signals, including EEG and PPS signals, and the relations between the samples of each signal. Then, a convolutional neural network (CNN) is considered to extract the local deep features among the output of the 2D-DOST. Since there are several redundant features in the set of deep features, semisupervised dimension reduction (SSDR) is used and a classifier recognizes the emotion. The feature reduction and classification performance depend on some parameters obtained by the Bayesian optimization approach to maximize accuracy. We considered the binary and four-class scenarios on the database for emotion analysis using the physiological signals (DEAP) dataset to assess the performance of the proposed method. The results demonstrate that the proposed method outperforms the recently introduced methods. Hence, the contributions of this paper are as follows:

(i) Proposing a new method for multimodal emotion recognition using EEG and PPS

(ii) Using the 2D-DOST to analyze the intramodal and cross-modal correlations

(iii) Extracting deep features by the CNN and then reducing the number of deep features by a semi-supervised method

(iv) Joint optimization of the parameters of SSDR and classifier

(v) Performing extensive simulations to indicate the performance of the proposed method.

Following this introduction, Section 2 presents the related works on multimodal emotion recognition. Section 3 describes the dataset and a detailed description of the proposed method. Section 4 contains the results and discussion, and Section 5 concludes the paper.

2. Related Works

The EEG is the most used physiological signal in single-modal emotion recognition systems [2–6]. EEG and other physiological signals, such as PPS, are usually used for emotion recognition in multimodal systems. The hierarchical fusion based on the CNN was proposed in [7] to extract the potential information multimodal signals, including the EEG and the PPS, and feature-level fusion was performed to merge the deep and statistical features. The binary classification scenarios based on valence and arousal dimensions were considered in the DEAP and MAHNOB-HCI datasets. The method presented in [8] combines the EEG and PPS with eye movement signals, and the joint oscillation structure of multichannel signals was analyzed by the multivariate synchrosqueezing transform (MSST). After that, a deep CNN extracts the local features from the MSST. Binary scenarios were evaluated based on the dimensions of arousal and valence on DEAP and MAHNOB-HCI datasets. An ensemble CNN was utilized in [9] to analyze the correlation between EEG and PPS signals from the DEAP dataset to develop multimodal emotion recognition. The multistage multimodal dynamical fusion network was proposed in [10] to analyze the unimodal, bimodal, and trimodal intercorrelations. It was shown that multistage fusion performs better than single-stage fusion on the DEAP dataset. The multiple-fusion-layer-based ensemble classifier of stacked autoencoder was proposed in [11] to recognize the emotions from the DEAP dataset. PPSs such as galvanic skin response (GSR), respiration patterns, and blood volume pressure were utilized in [12]. This method combines some continuous wavelet transforms (CWTs) and classifies them using a CNN. The four-class scenario on the DEAP dataset was considered for performance evaluation. The EEG, pulse, skin temperature, and blood pressure are recorded by the wearable sensor nodes in [13], and the fuzzy support vector machine (SVM) performs the emotion recognition.

Audio- and video-based signals are used separately or combined with the physiological signals for multimodal emotion recognition [14–16]. The EEG and facial expressions were used in [17] for multimodal emotion recognition. The combination of the CNN and the attention mechanism extracts the essential features from facial expressions, and a CNN extracts the spatial features from EEG signals. The features of different modalities are merged at the feature level. Binary scenarios on DEAP and MAHNOB-HCI datasets were considered for performance evaluation. Another method based on EEG signals and facial expressions was presented in [18]. The authors in [19] used facial expressions, GSR, and EEG signals with a hybrid fusion strategy. They considered the three emotions on the LUMED-2 dataset and four classes on the DEAP dataset. In [20], the 3D-CNN extracts the spatiotemporal features from the EEG signals and the video. A hybrid multimodal data fusion method was presented in [21] to fuse the audio and video signals from the DEAP dataset using a latent space linear map. The principal component analysis (PCA) and CNN were used in fusion and feature extraction from EEG and audio signals in [22] and then the grey wolf optimization algorithm was employed for selecting combined features. The heart rate can be detected from the photoplethysmography (PPG) signal; hence, some research used PPG. A method based on PPG and GSR signals was proposed in [23], which uses the 1D-CNN autoencoder model and lightweight model obtained using knowledge distillation. The performance of the model is evaluated on DEAP and MERTI-Apps datasets. The heart rate was extracted from PPG signals in [24], and then the combination of the 1D-CNN and long short-term memory (LSTM) was adopted for classification on MAHNOB-HCI. The features in time and frequency domains were extracted from PPG and GSR signals in [25] for emotion recognition. It was shown that feature selection with random forest recursive feature elimination and classification by the SVM yields the highest accuracy. Table 1 summarizes the recently introduced research on multimodal emotion recognition from biological signals. It is observed that DEAP is the most used dataset. Also, most works focusing on time-domain and time-frequency analyses were adopted in [8, 12]. The feature concatenation was considered after feature extraction from each modal, and cross-modal correlation was not considered in the feature extraction process.

3. Proposed Method

3.1. Dataset. To evaluate the performance of the proposed method, we consider the DEAP dataset [26]. Researchers at the Queen Mary University of London developed this publicly available dataset to analyze the emotions of 32 subjects on a scale of one to nine for valence and arousal. The 40 videos with the duration of 63 seconds were selected as trigger stimuli during the experiments. This dataset contains

Table 1: Summary of multimodal emotion recognition from biological signals.

References	Modality	Dataset	Domain analysis	Fusion/classification
[7]	EEG PPS	DEAP MAHNOB-HCI	Time domain	After feature extraction from each modal/decision tree
[8]	EEG PPS Video	DEAP MAHNOB-HCI	Time-frequency domain (MSST) for EEG Time domain for PPS Time and frequency domains for video	After feature extraction from each modal/CNN
[9]	EEG PPS	DEAP	Time domain	Before feature extraction/ensemble CNN
[10]	EEG PPS	DEAP	Time domain	After feature extraction from each modal/MLP
[11]	EEG PPS	DEAP	Time and frequency domains for different modalities	After feature extraction from each modal/CNN
[12]	PPS	DEAP	Time-frequency domain	After computing the CWT of each modal/CNN
[17]	EEG Video	DEAP MAHNOB-HCI	Time domain	After feature extraction from each modal by CNN/MLP
[19]	EEG GSR Video	DEAP LUMED-2	Time domain	After feature extraction from each modal by CNN/Decision tree
[20]	EEG Video	DEAP	Time domain	After feature extraction from each modal by 3D-CNN/MLP
[22]	EEG Audio	Private	Frequency domain	After feature extraction from each modal/CNN

EEG and PPS signals. EEG signals were recorded using 48 electrodes. PPSs are horizontal electrooculography (hEOG), vertical EOG (vEOG), zygomaticus major electromyography (zEMG), trapezius EMG (tEMG), galvanic skin response (GSR), respiration belt, plethysmograph, and temperature. All signals were downsampled to 128 Hz. EEG and PPS signals were passed through bandpass and lowpass filters, respectively. The middle 30 seconds of the 63 seconds of recorded data were considered for further processing. Since it was generally adopted that each subject reaches a stable in the middle of the video, the selected part of the signals was partitioned into segments with a duration of three seconds so that consecutive segments have a 50% overlap with each other. Therefore, there are 40 trials for each subject, each trial with 19 segments with 384 samples.

This paper considers two scenarios based on valence and arousal for rating the emotional signals. The binary scenario classifies the multimodal signals based on the valence rating into positive and negative emotions, as shown in Figure 1(a). Conversely, the four-class scenario considers the 2D valence-arousal model for classifying emotions into one of the following categories: sad, calm, happy, and angry, as shown in Figure 1(b).

3.2. Proposed Method. Here, the proposed method for multimodal emotion recognition from EEG and PPS signals is explained in detail. The general framework of the proposed method is shown in Figure 2, which consists of the following four main steps: data fusion, feature extraction, feature reduction, and classification.

3.2.1. Fusion. Previous works based on multimodal signals usually extract the features from different modalities separately and then merge the extracted features. In this manner, the cross-modal correlation is not considered. Also, there are many redundant features. To overcome this drawback, we propose to merge the multimodal signals before any feature extraction process. Let $\mathbf{X}_{\mathrm{EEG}}$ and $\mathbf{X}_{\mathrm{PPS}}$ denote the matrices with the size of 32×384 and 8×384, respectively. After fusion, there is the matrix $\mathbf{X}_m$ with the size of 40×384, which is considered for further processing.

3.2.2. Feature Extraction. The Stockwell transform was introduced to overcome the drawbacks of short-time Fourier transform (STFT) and wavelet transform while benefiting their advantages and characteristics [27]; however, there are some differences. STFT uses a fixed window size for signal analysis, resulting in a tradeoff between time and frequency resolution. In contrast, the Stockwell transform uses the variable-length window; hence, different frequency components can be analyzed with different time resolutions, which is necessary for transient and stationary signals. Since the Stockwell transform uses the Gaussian window, it provides a localized time-frequency map (TFM). In contrast, the STFT spreads the spectral energy over multiple time-frequency bins due to the use of rectangular windows. This Stockwell transform characteristic accurately identifies signal components' time and frequency characteristics. The STFT suffers from smearing due to the rectangular analysis window. The Stockwell transform mitigates this issue using a window that smoothly tapers off. The Stockwell transform retains phase information, while STFT distorts the phase due to the windowing process [28, 29].

For a continuous-time signal $x(t)$, the continuous Stockwell transform, $S(\tau, f)$, is computed as follows [30]:

$$S(\tau, f) = \frac{|f|}{2\pi} \int_{-\infty}^{\infty} x(t) e^{(t-\tau)^2/2\sigma^2} e^{-j2\pi f t} \mathrm{d}t = A(\tau, f) e^{j\theta(\tau, f)}, \tag{1}$$

where $j = \sqrt{-1}$, t, and τ are the time variables, f denotes the frequency, and $\sigma = 1/|f|$ is the scale factor. Also, $A(\tau, f)$ and $e^{j\theta(\tau, f)}$ are the magnitude and phase of the Stockwell transform, respectively. The output of the Stockwell transform is a complex-valued matrix whose rows and columns are concerned with time and frequency, respectively.

For the discrete signal $x[k]$, $k = 0, 1, \ldots, N-1$, obtained from $x(t)$, by sampling, with the discrete Fourier transform (DFT) of $X[n]$, $n = 0, 1, \ldots, N-1$, the discrete Stockwell transform for $x[k]$, $S[k, n]$, for $n \neq 0$, can be calculated by replacing $\tau \longrightarrow k$ and $f \longrightarrow n/N$ as follows [30]:

$$S[k, n] = \sum_{m=0}^{N-1} X[m+n] e^{2\pi^2 m^2/n^2} e^{j2\pi mk/N} = A[k, n] e^{[-k, n/N]}. \tag{2}$$

For $n = 0$, the Stockwell transform equals the DC value of DFT as $S[k, 0] = 1/N \sum_{m=0}^{N-1} x[m]$. The 2D Stockwell transform for the 2D image $f(x, y)$ is computed as follows [30]:

$$S(u, v, f_u, f_v) = \frac{|f_u||f_v|}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) e^{(u-x)^2 (v-y)^2/2} e^{-j2\pi (f_u x + f_v y)} \mathrm{d}x \, \mathrm{d}y. \tag{3}$$

The shift parameters u and v control the centre position of Gaussian windows on different axes. Also, f_u and f_v ($f_u \neq 0$ and $f_v \neq 0$) denote the frequencies. There is considerable redundancy in the time-frequency matrix provided by the Stockwell transform. The DOST is proposed in [31, 32] to overcome this drawback. The DOST provides spatial frequency representation similar to the wavelet transform [32]. The 2D-DOST of an $N \times N$ image $f(m, n)$, with 2D Fourier transform $F(m, n)$, is defined as follows:

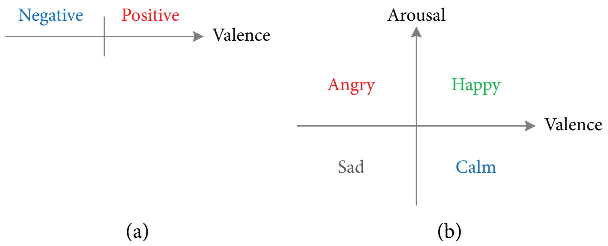

Figure 1: Different scenarios based on valence and arousal values. (a) Binary scenario. (b) Four-class scenario.

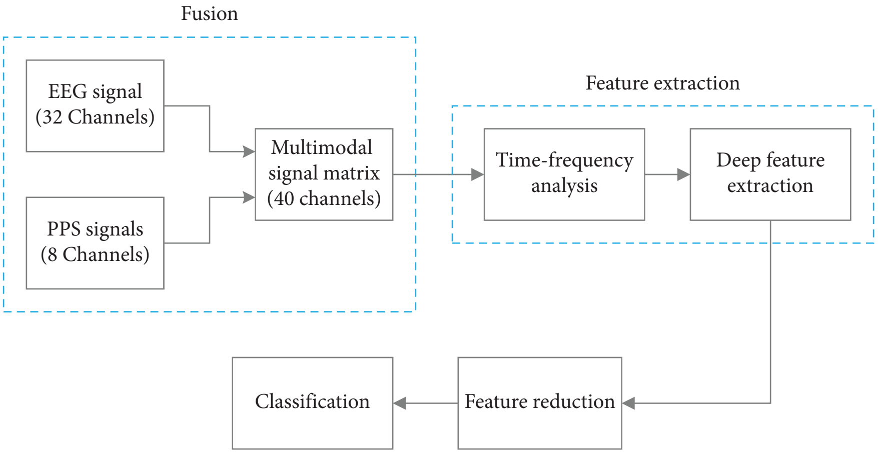

Figure 2: General framework of the proposed multimodal emotion recognition.

$$S(u,v,f_u,f_v)=\frac{1}{\sqrt{2^{p_x+p_y-2}}}\sum_{m=-2^{p_x-2}}^{2^{p_x-2}-1}\sum_{n=-2^{p_y-2}}^{n=-2^{p_y-2}-1}F\left(m+v_x,n+v_y\right)e^{j2\pi\left(mu/2^{p_x-1}+nv/2^{p_y-1}\right)}. \quad (4)$$

Here, $v_x = 2^{p_x-1}+2^{p_x-2}$ and $v_y = 2^{p_y-1}+2^{p_y-2}$ are the horizontal and vertical frequencies, respectively, and p_x and $p_y = 0, 1, \log(N-1)$. For this image, there are N^2 DOST points. The 2D-DOST gives information about the frequencies $S(u,v)$ in the bandwidth of $2^{p_x-1}\times 2^{p_y-1}$ frequencies [30].

As mentioned, the input of the 2D-DOST is an $N\times N$ image, and usually, N is a power of two for computational efficiency. Hence, each $\mathbf{X}_m$ with the size of 40×384 is partitioned into six partitions, resulting in $\mathbf{X}_m^{(i)}, i = 1,\ldots,6$, each with the size of 40×64. Finally, each $\mathbf{X}_m^{(i)}$ was resized to the size of 64×64. After that, the 2D-DOST is computed for each $\mathbf{X}_m^{(i)}$ to obtain $\mathbf{S}_m^{(i)}$. Finally, the time-frequency matrix, $\mathbf{S}_m$, of trial $\mathbf{X}_m$ is computed as follows:

$$\mathbf{S}_m=\frac{1}{6}\sum_{i=1}^{6}\mathbf{S}_m^{(i)}. \quad (5)$$

CNNs provide several benefits for analyzing TFMs. CNNs are particularly effective at capturing local patterns and features. TFMs contain localized structures; hence, CNNs can automatically learn and extract relevant local features from these maps. This enables the model to capture time-varying patterns and frequency-specific information. TFMs often exhibit hierarchical structures, where low-level features correspond to basic signal components and higher-level features capture more complex relationships and patterns. CNNs can learn these hierarchical representations by stacking multiple convolutional layers. This allows them to capture both low-level details, such as individual frequency components, and high-level features that represent more abstract signal characteristics. TFMs are susceptible to noise and variations introduced during signal acquisition or processing. CNNs have demonstrated robustness to noise and variations. By leveraging local receptive fields and pooling operations, CNNs can effectively suppress noise and capture invariant features in TFMs. This robustness enhances the model's ability to analyze the TFM in the presence of noise or variations [33–35].

The CNN extracts the multiscale localized spatial features from the input image using different layers, including image input, convolutional, batch normalization, rectified linear unit (ReLU), pooling, fully connected, and softmax. The convolutional layers generate high-level features by detecting local patterns such as lines and edges. The small-sized filters, or kernels, are employed for this purpose. The minibatch process normalizes the output of convolution layers to reduce the sensitivity to the initialization and increase the training speed. There is a nonlinear activation filter, called ReLU, after this layer, with the input-output relation function as $r_{\text{out}} = \max\{0, r_{\text{in}}\}$. There are many high-level features at the output of the ReLU layer with high correlation, and training such features requires more computational resources. Therefore, the pooling layer is employed to reduce the number of high-level features at the output of the ReLU layer. This layer generally performs the downsampling with functions such as average pooling, global maximum pooling, maximum pooling, and global average pooling, in which the max-pooling is the most frequently used. This function selects the maximum value in the pooling window. The output of the last pooling layer is given to the flatten layer that converts the feature maps from the matrix form to the vector one. The elements of this vector are the input of fully connected and softmax layers that act as the traditional multilayer perceptron.

Designing the new structures for the CNN and training them is time consuming and requires a huge number of labelled training samples. Transfer learning is utilized to solve this challenge. Generally, transfer learning is using the pretrained CNN for a new problem. To this end, only the number of neurons in the last dense layer is modified according to the number of classes of the new problem and the whole or some weights of the pretrained network are refined considering the training data of the new scenario. Also, the training samples are resized considering the size of the input image layer. After training, the features at the flatten layer's output are considered deep features and used for further processing.

3.3. Feature Reduction. Some high-level deep features obtained from the flatten layer may be highly correlated, increasing the redundancy in the feature vector given to the classifier. The redundant features increase the training complexity and probability of overfitting. Hence, they should be removed from the feature vector. The semisupervised methods combine the efficiencies of both supervised and unsupervised methods and balance the discrimination and generalization. This paper uses the semisupervised dimensionality reduction (SSDR) proposed in [33] for feature reduction.

Let n_t and n_0, respectively, denote the number of training samples and the number of deep extracted features. Accordingly, $\mathbf{s}_1, \ldots, \mathbf{s}_{n_t} \in \mathbb{R}^{n_0}$ are training feature vectors and $\mathbf{S}_1 = [\mathbf{s}_1, \ldots, \mathbf{s}_{n_t}]$. In this method, the n_M pairs of training samples belonging to the same class and n_C samples from different classes, respectively, construct the must-link constraints, M, and the cannot-link constraints, C. SSDR obtains the new feature vectors set $G = ?^{\mathrm{T}}?$, where $\mathbf{W}_1 = [\mathbf{w}_1, \ldots, \mathbf{w}_{n_r}]$, $?^{\mathrm{T}}\mathbf{W} = 1$, is the projection matrix, and the new features should preserve the structure of the original features. To this end, the objective function (?) is defined as follows:

$$J(W) = \frac{1}{n_t^2} \sum_{(\mathbf{s}_i, \mathbf{s}_j)} \left(\mathbf{W}^T \mathbf{s}_i - \mathbf{W}^T \mathbf{s}_j\right)^2 + \frac{\alpha}{2n_C} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in C} \left(\mathbf{W}^T \mathbf{s}_i - \mathbf{W}^T \mathbf{s}_j\right)^2 - \frac{\beta}{2n_M} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in M} \left(\mathbf{W}^T \mathbf{s}_i - \mathbf{W}^T \mathbf{s}_j\right)^2. \tag{6}$$

The parameters α and β balance the cannot- and must-link constraints. The concise form of the objective function can be expressed as follows:

$$J(W) = \frac{1}{2} \sum_{(\mathbf{x}_i, \mathbf{x}_j)} \left(\mathbf{W}^T \mathbf{s}_i - \mathbf{W}^T \mathbf{s}_j\right)^2 Y_{i,j} = \mathbf{W}^T \mathrm{S} \mathrm{L} \mathbf{S}^T \mathbf{W}, \tag{7}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{S}$ denotes the Laplacian matrix, and $\mathbf{D}$ is the diagonal matrix obtained as $\mathbf{D}_{ii} = \sum_j \mathbf{Y}_{i,j}$. The elements of matrix $\mathbf{Y}$ are obtained as follows:

$$Y_{i,j} = \begin{cases} \dfrac{1}{n_t^2 + n_C}, & \text{if } \left(\mathbf{s}_i, \mathbf{s}_j\right) \epsilon C, \\[2ex] \dfrac{1}{n_t^2 + n_M}, & \text{if } \left(\mathbf{s}_i, \mathbf{s}_j\right) \epsilon M, \\[2ex] \dfrac{1}{n_t^2}, & \text{otherwise.} \end{cases} \tag{8}$$

It is observed that the performance of SSDR depends on parameters α and β. Hence, Bayesian optimization is utilized to find their optimum value that maximizes the accuracy.

3.4. Classification. Here, several classifiers, including SVM, *k*NN, ANN, decision tree, and random forest, are considered separately to obtain the performance of the proposed method. The performance of these classifiers depends on their parameters. For the SVM, the kernel type and box constraint; for *k*NN, the number of neighbours, distance metric, and weighting scheme; for the decision tree, the maximum number of splits; and for the random forest, the minimum number of leaf sizes and number of predictors to sample should be optimized. A joint optimization based on Bayesian finds its optimum value, as shown in Figure 3. It should be mentioned that the structure of the ANN is chosen according to the dense layers in the corresponding CNN.

4. Results and Discussion

This section explains the simulations performed to assess the performance of the proposed method and the obtained results. The confusion matrix, accuracy (Acc), sensitivity (Sens), precision (Prec), kappa, and F_1 scores are calculated and reported. These metrics are calculated as follows:

$$\begin{aligned} \mathrm{Acc} &= \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}, \\ \mathrm{Sens} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ \mathrm{Prec} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \\ \mathrm{Kappa} &= \frac{\mathrm{Acc} - A_r}{1 - A_r} = \frac{\mathrm{Acc} - 1/N_c}{1 - 1/N_c}, \\ F_1 &= 2\frac{\mathrm{Prec} \times \mathrm{Sens}}{\mathrm{Prec} + \mathrm{Sens}}, \end{aligned} \tag{9}$$

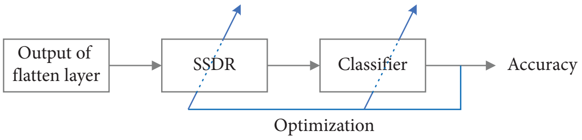

Figure 3: The procedure used to optimize the parameters of SSDR and classifier.

where the number of correctly classified and rejected multimodal signals is, respectively, denoted by true positive (TP) and true negative (TN). Conversely, the number of incorrectly identified and incorrectly rejected multimodal signals is given by false positive (FP) and false negative (FN), respectively. Also, $A_r = 1/N_c$ is the random accuracy, where N_c is the number of classes.

4.1. Simulation Setup. We adopt the cross-subject validation protocol to determine the train and test data. Hence, the proposed method is subject independent and considers the data of one subject for testing and the data of the remaining subjects train the model. This validation scheme repeats this procedure for all subjects as test data, and finally, the results are averaged. This paper considers some frequently used pretrained CNNs for deep feature extraction from the 2D-DOST content, including AlexNet, VGG19, ResNet18, Inception-v3, and EfficientNet-B0. Table 2 contains the parameters used in the tuning process of deep feature extractors.

4.2. Fusion Model. There are several ways to combine EEG and PPS signals to construct the matrix $\mathbf{X}_m$ such as EEG-PPS, $\mathbf{X}_m = [\mathbf{X}_{EEG}; \mathbf{X}_{PPS}]$, and PPS-EEG, $\mathbf{X}_m = [\mathbf{X}_{PPS}; \mathbf{X}_{EEG}]$. The other way is that channels of EEG and PPS signals are randomly located at the rows of the matrix $\mathbf{X}_m$. There are several placements for this purpose. We examined several placements, and the highest accuracy was reported. Also, the results of using only EEG and PPS signals are obtained. The results given in Table 3 depict that the EEG-PPS fusion yields the highest accuracy equal to 0.953 and 0.928 for two- and four-class scenarios, respectively. It is observed that random and PPS-EEG fusions have close accuracy, where the accuracy of the EEG-PPS scheme is slightly higher. This fusion scheme preserves the intramodal correlations among different channels and also considers the cross-modal correlations among the signals of different modalities. In contrast, a random manner cannot preserve the intramodal correlations among channels due to the random location of signals.

Also, comparing the results of only EEG and only PPS signals indicates that EEG signals are more informative than the PPSs; hence, their fusion reaches a higher accuracy than using only one. It should be noted that the maximum accuracy of both scenarios is obtained considering the deep features extracted by Inception-V3 CNN and SVM classifier. The structure of Inception-v3 [36] is given in Table 4. It should be noted that the output size of each module is the input size of the next one. The structure of inception modules is also given in Figure 4.

Tables 5 and 6 present the confusion matrix of the proposed method for two- and four-class scenarios, respectively. It is observed that the accuracy of the detection of negative emotions is slightly higher than positive ones in the two-class scenario. Notably, the minimum sensitivity is 94.7%, higher than the recently introduced works. The angry, happy, calm, and sad emotions are most accurate in the four-class scenario. Also, the values of kappa and F_1 scores indicate the efficiency of the proposed method.

4.3. Accuracy for Different Pairs of Classifiers and the CNN. Tables 7 and 8 present the accuracy and kappa score of the proposed method for different pairs of CNN and classifier to find the set of CNN and classifier that reaches the highest accuracy. Notably, each pair's reported accuracy is the maximum obtained by the optimization of SSDR and classifier parameters in the EEG-PPS fusion scheme. It is observed that in both scenarios, the combination of Inception-v3 and the SVM yields the highest accuracy. The ResNet18 and EfficientNet-B0 have a close performance that is lower than Inception-v3 and higher than AlexNet and VGG19. Also, the performance of VGG19 is better than AlexNet. For all CNNs, the SVM with Gaussian kernel reaches the highest accuracy, and after that, ANN has the highest accuracy in most cases.

Table 9 discusses the effect of feature reduction on the performance of the proposed method. We considered the proposed method without feature reduction, with unsupervised PCA, with supervised LDA, with the combination of PCA and LDA, with static SSDR, in which parameters are not optimized, and with optimized SSDR. It is observed that generally, using feature reduction increases the accuracy. Since LDA is supervised, it has higher accuracy than unsupervised PCA. However, the generalization of LDA is lower than PCA. To overcome this issue, a combination of them, PCA + LDA, can be used that reaches a higher accuracy than when used alone. The parameters of static SSR are set randomly, and it is observed that its performance is slightly lower than the hybrid PCA + LDA scheme.

4.4. Performance Comparison. Table 10 compares the performance of the recently introduced multimodal emotion recognition approaches. As observed, the EEG is the frequently used modality in multimodal emotion recognition systems. Most multimodal schemes considered the EEG and other biological signals such as EOG, PPS, GSR, and facial expressions. Also, the EEG and PPS signals are the most used. Generally, the EEG + PPS scheme reaches a higher accuracy than the other combinations of biological signals. It is observed that the proposed method has more accuracy than the recently introduced works.

Table 2: Parameters used for tuning the deep feature extractors.

Parameters	Values
Optimizer	Stochastic gradient descent with momentum (SGDM)
Loss function	Cross-entropy
Batch size	32
Momentum	0.80
Learning rate	10^{-4}
Number of epochs	60

Table 3: Accuracy of different fusion models.

Fusion models	Accuracy	
	Two class	Four class
Only EEG	0.923	0.894
Only PPS	0.875	0.847
EEG-PPS	**0.953**	**0.928**
PPS-EEG	0.942	0.921
Random	0.937	0.915

The bold values represent the highest accuracies.

Table 4: Structure of Inception-v3.

Types	Patch size/stride (or remarks)	Input size
Convolution	3 × 3/2	299 × 299 × 3
Convolution	3 × 3/1	149 × 149 × 32
Convolution padded	3 × 3/1	147 × 147 × 32
Maximum pooling	3 × 3/2	147 × 147 × 64
Convolution	3 × 3/1	73 × 73 × 64
Convolution	3 × 3/2	71 × 71 × 80
Convolution	3 × 3/1	35 × 35 × 192
3 × inception	As in Figure 3(a)	35 × 35 × 288
5 × inception	As in Figure 3(b)	17 × 17 × 768
2 × inception	As in Figure 3(c)	8 × 8 × 1280
Maximum pooling	8 × 8	8 × 8 × 2048
Linear	Logits (unnormalized log-probabilities)	8 × 8 × 2048
Softmax	Classifier	$8 \times 8 \times n_c$

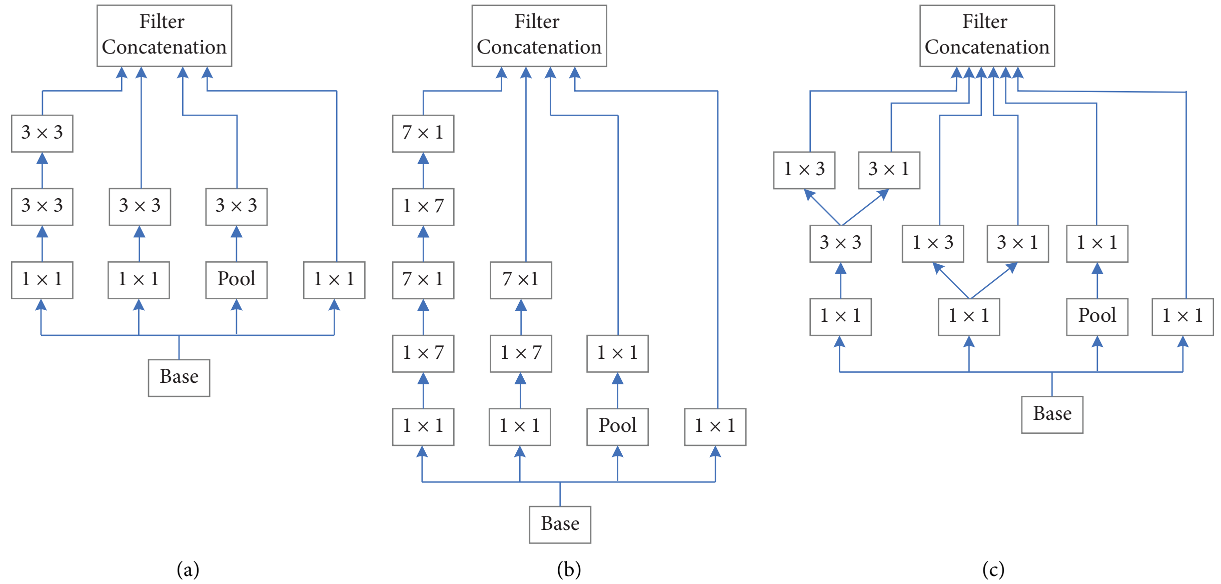

Figure 4: The structure of inception modules used in Inception-v3 CNN. (a) First inception module. (b) Second inception module. (c) Third inception module.

Table 5: Confusion matrix of the two-class scenario.

		Predicted emotion		Sens	Prec	Kappa	F_1 score
		Positive	Negative				
Actual emotion	Positive	0.947	0.053	0.947	0.958	0.894	0.953
	Negative	0.041	0.959	0.959	0.948	0.918	0.953

Table 6: Confusion matrix of the four-class scenario.

		Predicted emotion				Sens	Prec	Kappa	F_1 score
		Happy	Angry	Calm	Sad				
Actual emotion	Happy	0.933	0.021	0.028	0.018	0.933	0.931	0.862	0.932
	Angry	0.019	0.937	0.018	0.026	0.937	0.922	0.844	0.929
	Calm	0.032	0.02	0.923	0.025	0.923	0.928	0.857	0.926
	Sad	0.018	0.038	0.025	0.919	0.919	0.931	0.861	0.925

Table 7: Classification accuracy and kappa score of the two-class scenario for different CNNs and classifiers.

	CNN									
Classifiers	AlexNet		VGG19		ResNet18		Inception-v3		EfficientNet-B0	
	Acc	Kappa	Acc	Kappa	Acc	Kappa	Acc	Kappa	Acc	Kappa
SVM	0.891	0.782	0.903	0.806	0.918	0.836	**0.953**	**0.906**	0.915	0.830
ANN	0.889	0.778	0.895	0.790	0.909	0.818	0.915	0.830	0.911	0.822
*k*NN	0.879	0.758	0.892	0.784	0.914	0.828	0.912	0.824	0.908	0.816
Random forest	0.881	0.762	0.897	0.794	0.901	0.802	0.909	0.818	0.908	0.816
Decision tree	0.874	0.748	0.867	0.734	0.899	0.798	0.901	0.802	0.901	0.802

The bold values represent the highest accuracies.

Table 8: Classification accuracy and kappa score of the four-class scenario for different CNNs and classifiers.

	CNN									
Classifier	AlexNet		VGG19		ResNet18		Inception-V3		EfficientNet-b0	
	Acc	Kappa	Acc	Kappa	Acc	Kappa	Acc	Kappa	Acc	Kappa
SVM	0.876	0.752	0.881	0.762	0.901	0.816	**0.928**	**0.856**	0.904	0.808
ANN	0.875	0.750	0.886	0.774	0.885	0.802	0.901	0.801	0.898	0.796
*k*NN	0.864	0.728	0.866	0.732	0.875	0.782	0.891	0.783	0.884	0.768
Random forest	0.847	0.694	0.875	0.751	0.901	0.778	0.889	0.779	0.882	0.764
Decision tree	0.847	0.694	0.853	0.701	0.871	0.764	0.882	0.764	0.889	0.778

The bold values represent the highest accuracies.

Table 9: The effect of feature reduction on the accuracy.

	Scenario			
Method	Four-class		Two-class	
	Acc	Kappa	Acc	Kappa
Without feature reduction	0.806	0.612	0.831	0.662
PCA	0.837	0.674	0.859	0.718
LDA	0.852	0.704	0.872	0.744
PCA + LDA	0.873	0.746	0.897	0.794
Static SSDR	0.869	0.738	0.885	0.770
Optimized SSDR	**0.928**	**0.856**	**0.953**	**0.906**

The bold values represent the highest accuracies.

Table 10: Performance comparison of the proposed method and recently introduced ones.

Authors	Modality	Accuracy
Wu et al. [37]	EEG + EOG	0.866 (two classes)
Hatipoglu Yilmaz and Kose [38]	EEG + EOG	0.915 (two classes)
Ma et al. [39]	EEG + PPS	0.923 (two classes)
Qiu et al. [40]	EEG + PPS	0.856 (two classes)
Li et al. [41]	EEG + PPS	0.949 (two classes)
Zhang et al. [7]	EEG + PPS	0.847 (two classes)
Zhang et al. [8]	EEG + PPS	0.901 (two classes)
Jalal and Peer [12]	PPS	0.842 (four classes)
Cimtay et al. [19]	EEG, GSR, facial	0.915 (four classes)
Proposed method	EEG + PPS	0.953 (two classes) 0.928 (four classes)

5. Conclusion

This paper proposed a new method for emotion recognition from multimodal signals, including EEG in 32 channels and PPS in eight channels. The proposed method employs the 2D-DOST to analyze the relations between the multimodal signals. Then, a CNN was used to extract the deep local features from the absolute of the 2D-DOST. After feature reduction by SSDR, a classifier determines the emotion by solving an optimization problem. The results showed that the extracted deep features by the Inception-v3 network and their classification by the Gaussian SVM reached the highest accuracy equal to 0.953 and 0.928, respectively, for two- and four-class scenarios on the DEAP dataset. Several fusion schemes to combine the EEG and PPS signals were examined, and it was observed that the scheme $[\mathbf{X}_{EEG}; \mathbf{X}_{PPS}]$ has the maximum accuracy. Also, it was shown that optimized SSDR has higher accuracy than the frequently used feature reduction schemes such as PCA and LDA. The results indicate the efficiency of multimodal emotion recognition compared to the unimodal approach. Also, the proposed method outperforms the recently introduced methods.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Behrooz Zali-Vargahan was responsible for conceptualization, investigation, methodology, software, and writing the original draft; Asghar Charmin was responsible for conceptualization, methodology, validation, and supervision; Hashem Kalbkhani was responsible for conceptualization, software, visualization, and review writing and editing; and Saeed Barghandan was responsible for the methodology and review writing and editing.

References

[1] X. Zheng, X. Yu, Y. Yin, T. Li, and X. Yan, "Three-dimensional feature maps and convolutional neural network-based emotion recognition," *International Journal of Intelligent Systems*, vol. 36, no. 11, pp. 6312–6336, 2021.

[2] B. Zali-Vargahan, A. Charmin, H. Kalbkhani, and S. Barghandan, "Deep time-frequency features and semi-supervised dimension reduction for subject-independent emotion recognition from multi-channel EEG signals," *Biomedical Signal Processing and Control*, vol. 85, Article ID 104806, 2023.

[3] Y. Zhou, F. Li, Y. Li et al., "Progressive graph convolution network for EEG emotion recognition," *Neurocomputing*, vol. 544, Article ID 126262, 2023.

[4] S. Liu, Y. Zhao, Y. An et al., "A global to local feature 395 aggregation network for EEG emotion recognition," *Biomedical Signal Processing and Control*, vol. 396 85, Article ID 104799, 2023.

[5] X. Zheng, X. Liu, Y. Zhang, L. Cui, and X. Yu, "A portable HCI system-oriented EEG feature extraction and channel selection for emotion recognition," *International Journal of Intelligent Systems*, vol. 36, no. 1, pp. 152–176, 2021.

[6] D. Li, L. Xie, Z. Wang, and H. Yang, "Brain emotion perception inspired EEG emotion recognition with deep reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2023.

[7] Y. Zhang, C. Cheng, and Y. Zhang, "Multimodal emotion recognition using a hierarchical fusion convolutional neural network," *IEEE Access*, vol. 9, pp. 7943–7951, 2021.

[8] Y. Zhang, C. Cheng, and Y. Zhang, "Multimodal emotion recognition based on manifold learning and convolution neural network," *Multimedia Tools and Applications*, vol. 81, no. 23, pp. 33253–33268, 2022.

[9] H. Huang, Z. Hu, W. Wang, and M. Wu, "Multimodal emotion recognition based on ensemble convolutional neural network," *IEEE Access*, vol. 8, pp. 3265–3271, 2020.

[10] S. Chen, J. Tang, L. Zhu, and W. Kong, "A multi-stage dynamical fusion network for multimodal emotion recognition," *Cognitive Neurodynamics*, vol. 17, no. 3, pp. 671–680, 2022.

[11] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Computer Methods and Programs in Biomedicine*, vol. 140, pp. 93–110, 2017.

[12] L. Jalal and A. Peer, "Emotion recognition from physiological signals using continuous wavelet transform and deep learning," in *Proceedings of the HCI International 2022-Late Breaking Papers. Multimodality in Advanced Interaction Environments: 24th International Conference on Human-Computer Interaction, HCII 2022*, pp. 88–99, Springer, Berlin, Germany, July, 2022.

[13] Y. Dai, X. Wang, P. Zhang, and W. Zhang, "Wearable biosensor network enabled multimodal daily-life emotion recognition employing reputation-driven imbalanced fuzzy classification," *Measurement*, vol. 109, pp. 408–424, 2017.

[14] M. J. Al-Dujaili and A. Ebrahimi-Moghadam, "Speech emotion recognition: a comprehensive survey," *Wireless Personal Communications*, vol. 129, no. 4, pp. 2525–2561, 2023.

[15] S. Liu, P. Gao, Y. Li, W. Fu, and W. Ding, "Multi-modal fusion network with complementarity and importance for emotion recognition," *Information Sciences*, vol. 619, pp. 679–694, 2023.

[16] N. Saleem, J. Gao, R. Irfan et al., "DeepCNN: spectro-temporal feature representation for speech emotion recognition," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 2, pp. 401–417, 2023.

[17] S. Wang, J. Qu, Y. Zhang, and Y. Zhang, "Multimodal emotion recognition from EEG signals and facial expressions," *IEEE Access*, vol. 11, pp. 33061–33068, 2023.

[18] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography," *Biomedical Signal Processing and Control*, vol. 70, 2021.

[19] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168865–168878, 2020.

[20] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. Wahby Shalaby, "A 3D-convolutional neural network framework with ensemble learning techniques for multimodal emotion recognition," *Egyptian Informatics Journal*, vol. 22, no. 2, pp. 167–176, 2021.

[21] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen, and V. Makarenkov, "A hybrid latent space data fusion method for multimodal emotion recognition," *IEEE Access*, vol. 7, pp. 172948–172964, 2019.

[22] R. A. Jaswal and S. Dhingra, "Empirical analysis of multiple modalities for emotion recognition using convolutional neural network," *Measurement: Sensors*, vol. 26, 2023.

[23] D.-H. Kang and D.-H. Kim, "1D convolutional autoencoder-based PPG and GSR signals for real-time emotion classification," *IEEE Access*, vol. 10, pp. 91332–91345, 2022.

[24] W. Mellouk and W. Handouzi, "CNN-LSTM for automatic emotion recognition using contactless photoplythesmographic signals," *Biomedical Signal Processing and Control*, vol. 85, 2023.

[25] J. A. Domínguez-Jiménez, K. C. Campo-Landines, J. C. Martínez-Santos, E. J. Delahoz, and S. H. Contreras-Ortiz, "A machine learning model for emotion recognition from physiological signals," *Biomedical Signal Processing and Control*, vol. 55, 2020.

[26] S. Koelstra, C. Muhl, M. Soleymani et al., "Deap: a database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2012.

[27] R. G. Stockwell, L. Mansinha, and R. Lowe, "Localization of the complex spectrum: the S transform," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 998–1001, 1996.

[28] M. Das and S. Ari, "Analysis of ECG signal denoising method based on S-transform," *IRBM*, vol. 34, no. 6, pp. 362–370, 2013.

[29] Z. Zidelmal, A. Amirou, D. Ould-Abdeslam, A. Moukadem, and A. Dieterlen, "QRS detection using S-Transform and Shannon energy," *Computer Methods and Programs in Biomedicine*, vol. 116, no. 1, pp. 1–9, 2014.

[30] R. S. Choraś, "Time-frequency analysis of image based on stockwell transform," in *Image Processing and Communications Challenges 5*, pp. 91–97, Springer, Heidelberg, Germany, 2014.

[31] Y. Wang and J. Orchard, "Fast discrete orthonormal Stockwell transform," *SIAM Journal on Scientific Computing*, vol. 31, no. 5, pp. 4000–4012, 2009.

[32] S. Drabycz, R. G. Stockwell, and J. R. Mitchell, "Image texture characterization using the discrete orthonormal S-transform," *Journal of Digital Imaging*, vol. 22, no. 6, pp. 696–708, 2009.

[33] J. Zhang, Y. Li, and J. Yin, "Modulation classification method for frequency modulation signals based on the time–frequency distribution and CNN," *IET Radar, Sonar and Navigation*, vol. 12, no. 2, pp. 244–249, 2018.

[34] J. Nie, Y. Xiao, L. Huang, and F. Lv, "Time-frequency analysis and target recognition of HRRP based on CN-LSGAN, STFT, and CNN," *Complexity*, vol. 2021, Article ID 6664530, 10 pages, 2021.

[35] M. A. Ozdemir, D. H. Kisa, O. Guren, and A. Akan, "Hand gesture classification using time–frequency images and transfer learning based on CNN," *Biomedical Signal Processing and Control*, vol. 77, Article ID 103787, 2022.

[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, Las Vegas, NV, USA, June, 2016.

[37] X. Wu, W.-L. Zheng, Z. Li, and B.-L. Lu, "Investigating EEG-based functional connectivity patterns for multimodal emotion recognition," *Journal of Neural Engineering*, vol. 19, no. 1, Article ID 016012, 2022.

[38] B. Hatipoglu Yilmaz and C. Kose, "A novel signal to image transformation and feature level fusion for multimodal emotion recognition," *Biomedical Engineering/Biomedizinische Technik*, vol. 66, no. 4, pp. 353–362, 2021.

[39] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual LSTM network," in *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 176–183, Nice, France, October, 2019.

[40] J.-L. Qiu, W. Liu, and B.-L. Lu, "Multi-view emotion recognition using deep canonical correlation analysis," in *Proceedings of the Neural Information Processing: 25th International Conference, ICONIP 2018*, vol. 25, pp. 221–231, Springer, Krong Siem Reap, Cambodia, December, 2018.

[41] Q. Li, Y. Liu, F. Yan, Q. Zhang, and C. Liu, "Emotion recognition based on multiple physiological signals," *Biomedical Signal Processing and Control*, vol. 85, Article ID 104989, 2023.