Research Paper

A Hybrid EEG-based Emotion Recognition Approach Using Wavelet Convolutional Neural Networks and Support Vector Machine

Sara Bagherzadeh[1], Keivan Maghooli[1*], Ahmad Shalbaf[2*], Arash Maghsoudi[1]

1. Department of Biomedical Engineering, Sciences and Research Branch, Islamic Azad University, Tehran, Iran.
2. Department of Biomedical Engineering and Medical Physics, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

Use your device to scan and read the article online

Citation Bagherzadeh, S., Maghooli, K., Shalbaf, A., & Maghsoudi, A. (2023). A Hybrid EEG-based Emotion Recognition Approach Using Wavelet Convolutional Neural Networks and Support Vector Machine. *Basic and Clinical Neuroscience, 14*(1), 87-102. http://dx.doi.org/10.32598/bcn.2021.3133.1

http://dx.doi.org/10.32598/bcn.2021.3133.1

Article info:
Received: 21 Dec 2020
First Revision: 06 Mar 2021
Accepted: 14 Apr 2021
Available Online: 01 Jan 2023

Keywords:
Emotion recognition, Electroencephalogram, Continuous wavelet transform, Convolutional neural network, Feature extractor, Support vector machine

ABSTRACT

Introduction: Nowadays, deep learning and convolutional neural networks (CNNs) have become widespread tools in many biomedical engineering studies. CNN is an end-to-end tool, which makes the processing procedure integrated, but in some situations, this processing tool requires to be fused with machine learning methods to be more accurate.

Methods: In this paper, a hybrid approach based on deep features extracted from wavelet CNNs (WCNNs) weighted layers and multiclass support vector machine (MSVM) was proposed to improve the recognition of emotional states from electroencephalogram (EEG) signals. First, EEG signals were preprocessed and converted to Time-Frequency (T-F) color representation or scalogram using the continuous wavelet transform (CWT) method. Then, scalograms were fed into four popular pre-trained CNNs, AlexNet, ResNet-18, VGG-19, and Inception-v3 to fine-tune them. Then, the best feature layer from each one was used as input to the MSVM method to classify four quarters of the valence-arousal model. Finally, the subject-independent leave-one-subject-out criterion was used to evaluate the proposed method on DEAP and MAHNOB-HCI databases.

Results: Results showed that extracting deep features from the earlier convolutional layer of ResNet-18 (Res2a) and classifying using the MSVM increased the average accuracy, precision, and recall by about 20% and 12% for MAHNOB-HCI and DEAP databases, respectively. Also, combining scalograms from four regions of pre-frontal, frontal, parietal, and parietal-occipital and two regions of frontal and parietal achieved the higher average accuracy of 77.47% and 87.45% for MAHNOB-HCI and DEAP databases, respectively.

Conclusion: Combining CNN and MSVM increased the recognition of emotion from EEG signals and the results were comparable to state-of-the art studies.

*** Corresponding Author:**
Keivan Maghooli, PhD.
Address: *Department of Biomedical Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.*
Tel: *+98 (912) 2993212*
E-mail: *keivanmaghooli91@gmail.com*

*** Corresponding Author:**
Ahmad Shalbaf, *PhD.*
Address: *Department of Biomedical Engineering and Medical Physics, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran.*
Tel: *+98 (912) 2993212*
Email*: Shalbaf@sbmu.ac.ir*

Highlights

• Electroencephalogram (EEG) is a suitable measure to study emotion due to high temporal resolution, inexpensiveness and comfort recording for the user.

• A hybrid approach based on Wavelet convolutional neural networks and multiclass support vector machine is proposed to improve recognition of emotional states from EEG signals.

• Combining scalograms from frontal and parietal regions achieved the higher average accuracy of 77.47% and 87.45% for MAHNOB-HCI and DEAP databases, respectively.

Plain Language Summary

Emotions are an important function of the human brain that affects our decision and behavior, for example, when we are angry or sad, we can decide to do dangerous acts or, when we are bored we could learn lessons hardly. These emotions can be tracked and monitored by signals called electroencephalogram (EEG) from the scalp surface. In this study, EEG signals are represented by a time-frequency method, named wavelet transform. This method transforms EEG frequency information into time samples. Then, two machine learning methods are combined to recognize emotions from time-frequency images (the scalogram). The first method is based on deep learning techniques (convolutional neural networks) and extracts deep features from images and the second method discriminates them among emotions. Two public EEG databases, DEAP and MAHNOB-HCI were used to evaluate the proposed method. Results showed that scalograms from the combination of the two frontal and parietal regions achieved the highest accuracies compared to all regions.

1. Introduction

Emotions are brain states evoked in response to external and internal stimulations. Watching pictures and movies, listening to music, and smelling odors are external stimuli (Alarcao & Fonseca, 2017). Among these stimuli, watching movie clips is an efficient method to elicit human emotions. Movie clips can affect people like situations that occur in real life, as they have the two dimensions of sound and moving images simultaneously and benefit from these two dimensions to evoke people's emotions. Changes in the body, facial expressions, and physiological changes are examples of actions, which occur following emotional states. People are faced with different emotions in daily life, such as happiness, sadness, surprise, excitement, etc. These emotions can be represented with the two-dimensional model named the valence-arousal model (Russell, 1980). Valence means pleasantness from the emotion and ranges from negative to positive and arousal means the excitation of emotion that ranges from low to high. This model has four quarters, the first quarter includes emotions, such as excitement and happiness with high valence and low arousal (quarter 4, Q4) values. The second quarter includes emotions, like fear or anger with high valence and high arousal (quarter 1, Q1) values. Emotions, such as sadness, boredom, and depression are in the third quarter, having low valence and low arousal (quarter 3, Q3) values. Contentment and calmness are the fourth emotions, which have low valence and high arousal (quarter 2, Q2) values (García-Martínez et al., 2019). For emotion recognition, electroencephalogram (EEG) is well accepted due to its high correlation with emotional states in psychological studies, high temporal resolution, simple recording, and, being a noninvasive, and portable method (Alarcao & Fonseca, 2017; Rolls, 2015). EEG devices could be set up easily in real-time and are widely used in clinical applications (Afshani et al., 2019; Shalbaf et al., 2018).

In recent years, a number of EEG-based approaches have been proposed for emotion recognition during watching movie clips. Soleymani and Pantic (2013) proposed a system based on the Fast Fourier Transform (FFT) method and the support vector machine (SVM) for classifying three classes of valence and arousal. Koelstra and Patras (2013) extracted the power spectral density (PSD) from EEG signals. Then, appropriate electrodes using the recursive feature elimination (RFE) method are selected, and used the Gaussian Naïve Bayes (GNB) method to classify two classes of valence and arousal. Soleymani et al., (2015) detected emotions continuously using PSD of EEG sub-bands and four methods of long-short-term-memory recurrent neural networks (LSTM-RNN), multi-linear regression (MLR), continuous conditional random fields (CCRF), and support vector regression (SVR) for arousal and valence. Zhu et al. (2014) used a power spectrum of standard frequency bands and SVM to classify emotions considering binary valence and arousal values. Huang et al. (2016) proposed an emotion recognition system using spectral power (SP), sequential forward floating search, fusion in level and decision, K-Nearest Neighbor (KNN), and SVM from EEG signals. Nonlinear features, such as fractal dimension (FD), correlation dimension (CD), and Poincare plot are another way to investigate emotional states from EEG. Soroush et al. extracted these nonlinear features from EEG signals and classified four emotional states (valence and arousal quarters) using machine learning classifiers. As observed from these studies, traditionally, machine learning methods have been used for feature selection and classification. But over the past few years, there has been a developing interest in the utilization of deep learning methods, such as the convolutional neural network (CNN). CNN is a novel deep-learning method, which extracts low- and high-level features, reduces feature size, and finally classifies (Craik et al., 2019; Guo et al., 2016; Bengio et al., 2017; Roy et al., 2019). It has been widely applied in computer vision studies, especially in medical applications (Suzuki, 2017; Lundervold & Lundervold, 2019; Sun et al., 2017). Recently, this method has been utilized to process EEG signals for non-emotion (Faust et al., 2018; Zhang et al., 2019; Craik et al., 2019) and emotion studies (Yang et al., 2018). Yang et al. (2018) extracted the nonlinear feature of recurrence quantification analysis (RQA) from EEG channels and used the parallel convolutional neural networks (CNNs) to classify two emotional classes based on valence and arousal concepts, separately.

The contribution of this paper is divided into three parts:

1- Finding effective brain regions involved in recognizing emotional states using two-dimensional time-frequency representations of EEG signal and pre-trained CNN models. Two-dimensional time-frequency representations of EEG signals in the local or global form are fed into various pre-trained CNN models to fine-tune procedure and benefit from extracted deep features.

2- Improving an emotion recognition framework based on the fusion of deep features extracted from pre-trained CNN models and classification using the multiclass support vector machine (MSVM) method.

3- Evaluating the proposed emotion recognition framework through a subject-independent approach. Leave-one-subject-out cross-validation (LOSO CV) is used to evaluate the proposed framework on two publicly accessible datasets of MAHNOB-HCI and DEAP that were recorded during watching movie clips and music video clips.

2. Materials and Methods

MAHNOB-HCI database

In this paper, EEG signals of the MAHNOB-HCI database were used (Soleymani et al., 2011). These EEG signals were recorded during watching 20 different video clips from 27 subjects (16 females and 11 males) with an age range of 19 to 40 years and with different education levels, cultures, and languages. In this study, after watching, EEG signals from seven subjects were removed due to the high levels of artifacts, and finally, 32 channels from 20 subjects were used in the processing step. EEG signals were recorded using active electrodes with the Biosemi device according to the international 10-20 system of electrode placement and divided into nine anatomical brain regions according to Table 1. Other details about the database are reported in Table 2. After watching each video clip, subjects filled out a self-assessment questionnaire to evaluate the valence and arousal concepts and rated them by the Self Manikin Assessment (SAM) values from one to nine (one for low and nine for high). The labels were considered based on these values and from the used 230 trials, 57 trials belonged to the first quarter of the valence-arousal model (Q1), 59, 52, and 62 trials to the second (Q2), third (Q3), and fourth (Q4) quarters, respectively.

DEAP database

In the DEAP database, 32 EEG channels were recorded according to a 10-20 international recording system from 32 subjects (16 males and 16 females in the age range of 17 to 37 years) while watching a music video (Koelstra et al., 2011). The original sampling frequency was 512 Hz. Also, 40 music videos with a length of 60 seconds were used to evoke four emotional states and a neutral state. After watching music videos, subjects rated the value of valence and arousal based on SAM from one to nine scales. The class labels were considered as a previous explanation in the MAHNOB-HCI database and other details about the database are reported in Table 2.

Table 1. EEG signals of the brain regions

Region (Symbol)	Electrodes
Pre-frontal (FP)	FP1, FP2, AF3, AF4
Frontal (F)	F7, F3, FZ, F4, F8
Frontal-Central (F-C)	FC5, FC1, FC2, FC6
Central (C)	C3, C4, CZ
Central-Parietal (C-P)	CP5, CP1, CP2, CP6
Parietal (P)	P7, P3, PZ, P4, P8
Temporal (T)	T7, T8
Parietal-Occipital (P-O)	PO3, PO4
Occipital (O)	O1, OZ, O2

NEUROSCIENCE

Preprocessing EEG signal

Preprocessing is a crucial step in pattern recognition and especially, EEG signal processing. Usually, EEG signals are affected by the subject's factors, such as eye or body movements, or environmental factors, like power line noise. Muscle artifacts or eye movement contaminated high-frequency components; thus, these artifacts were removed by a basic Finite Impulse Response (FIR) low pass filter with a 45 Hz cut-off frequency. Head or body movements contaminated low frequency (below 0.5 Hz); thus, a high pass filter was used to re-

Table 2. Detail of MAHNOB-HCI and DEAP databases

MAHNOB-HCI	Description	
	Length of signal	34.9-117 (sec)
	Number of channels	32 (10-20 international electrode placements)
	Sampling frequency	256 Hz
	Number of subjects	27
	Stimulation type	Short video clips
	Number of video clips (emotional states)	20 (4)
	Valence and arousal rate	1 (low)-9 (high)
DEAP	Length of signal	60 (sec)
	Number of channels	32 (10-20 international electrode placements)
	Sampling frequency	512 (Hz)
	Number of subjects	32
	Stimulation type	Short music clips
	Number of music video clips (emotional states)	40 (4)
	Valence and arousal rate	1 (low)-9 (high)

NEUROSCIENCE

move these artifacts with a 0.5 Hz cut-off frequency. Because filters are not really ideal and there was leakage in near-frequency components, a Notch filter was used to remove the 50 Hz frequency component of power line noise. Also, bad channels were removed manually in the EEGLAB toolbox. All preprocessing steps were done using the EEGLAB toolbox. EEG signals from the DEAP database were recorded at 512 Hz sampling frequency, and for simplicity and to reduce samples, they were down-sampled to 128 Hz in the EEGLAB toolbox. Moreover, EEG signals from the MAHNOB-HCI database were not recorded by the reference electrode, and then, signals were re-referenced by the averaging method before filtering.

Converting EEG signals to a time-frequency representation

Time-frequency (TF) methods, like continuous wavelet transform, can convert a 1-D EEG signal to a 2-D representation or image and capture the variation of the spectral content of a signal over time. One dimension of our image is time and the other is the spectral content of a signal. This image represents EEG power changes in frequency and time. It represents a signal as a linear combination of basic functions called wavelets. This method convolves the signal x(t) with a set of wavelets (Chaudhary et al., 2019) (Equation 1):

$$1.\ W_{(a.b)}[x(t)]=\frac{1}{|a|^{1/2}}\int_{-\infty}^{+\infty} x(t)\ Ø\hat{}*(=\frac{t-b}{a})dt$$

Where, a is the scale (real and positive integer), b is the translational value (real integer), ω is a window, and Ø is the mother wavelet, which is in time and frequency domains.

Convolutional neural network and pre-trained versions

CNN is one of the most powerful tools of deep learning methods in the computer vision field. This network contains three different layers, convolutional (C), pooling (P), and fully connected (FC) (Guo et al., 2016; Bengio et al., 2017). Feature maps are created in convolutional layers using kernels. Pooling layer-down samples feature maps using maximum or average operators and fully connected layers to the classification operation. Drop-out and batch normalization techniques are introduced to overcome the overfitting problem in this neural network. Pre-trained CNNs are networks that are trained previously on very large amounts of images, like the ImageNet database consisting of many categories (Krizhevsky et al., 2012). ImageNet is a known image database for visual object recognition projects that starts with 1.2 million images from 1000 different categories from animals (dogs, cats, lions, ….) to objects (desks, pens, chairs, …). Because the pre-trained CNN was trained in a huge database with several categories, it can be useful to solve several classification problems even in biomedical signal processing studies; for example, it was used to diagnose schizophrenia from EEG signal (Shalbaf et al., 2020). Indeed, the parameters (weights, layers, and biases) of a pre-trained CNN will be used to solve the new problem. This work reduces the requirement for several recorded samples, decreases training time, and can be manipulated in low-cost and cheaper hardware. AlexNet, VGGNet, Inceptions, and ResNet are the four popular pre-trained CNNs that were used in this study due to their specific characteristics.

AlexNet is a simple CNN with a few convolutional layers, which has won the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) (Krizhevsky et al., 2012). It has five convolutional layers for extraction of low- and high-level features using several 11×11, 5×5, and 3×3 filters. Also, it has three max pooling layers to downsample the extracted features in previous convolutional layers and three fully connected layers for classification. It uses the rectified linear unit (ReLU) as an activation function after each convolutional and fully connected layer. AlexNet has 61 million parameters (from training on ImageNet) and allows 227×227 color images as input. VGGNet is the runner-up of ILSVRC2014 and has been introduced by Simonyan and Zisserman (2014). This network has two versions with different stacked convolutional layers, VGG-16 and VGG-19. VGG-16 has three stacked three convolutional layers and VGG-19 has three stacked four convolutional layers. In this paper, VGG-19 was used due to better performance. It has 19 uniform convolutional layers with several 3×3 filters and allows color images with the size of 224×224 and has 144 million parameters, which were from training VGG-19 in the ImageNet database. After winning of AlexNet in OLSVRC2012, the residual network (ResNet) is the winner of ILSVRC2015 (He et al., 2016).

ResNet has many stacked identity shortcut connections that help to solve the vanishing gradient problem of earlier CNNs. Researchers found that deeper CNNs face the vanishing gradient problem, i.e. when there are so many layers, repeating multiplication makes the gradient value to be near zero and it will be vanished in updating procedure. Therefore, the performance will be degraded following each additional layer, and ResNet overcomes this problem through its architecture. ResNet has some versions with various numbers of convo-

Table 3. Comparison of detail of the used pre-trained CNNs

Net	Convolutional Layers	Parameters on ImageNet (Millions)	Image Input Size
AlexNet	8	61	227×227×3
VGG-19	19	144	224×224×3
Inception-v3	48	23.9	299×299×3
ResNet-18	18	11.7	224×224×3

NEUROSCIENCE

lutional layers and filter sizes. ResNet-18 is one of its versions that has the lowest convolutional layer among other versions. It has 18 convolutional layers with several 3×3 filters that are arranged in stacked form side by side batch normalization, pooling, and layers. ResNet-18 requires a 224×224 color image as input and has 11.7 million parameters trained in the ImageNet database. After winning the ResNet-18 in ILSVRC2015, the third version of Inception Net (Inception-v3) was first Runner Up in ILSVRC 2015 (Szegedy et al. 2016). Inception-v3 has several stacked inception modules, which are parallel convolutional layers. This network reduced the number of connections, without degrading the efficiency of the network. Inception-v3 is a 48-convolutional layer CNN that in each layer has several 3×3 and 5×5 filters with different slide and padding characteristics. Finally, Inception-v3 has 23.9 million parameters trained in the ImageNet database and requires color images with a size of 299×299 as input. Table 3 compares these pre-trained CNNs in detail.

Transfer learning approach

Transfer learning is an approach with two scenarios that helps effectively in the deep learning field. The first scenario is fine-tuned, that is, transfer learning employs a pre-trained reference model trained previously for a specific classification task and adapts it using a smaller database for a new application. The second scenario is deep learning as a feature extractor, that is, using parameters of some deep layers from a pre-trained reference model trained previously for a specific classification task as features and then feeding them into a classifier, like SVM. These need shorter time and fewer samples for the training procedure. Also, they resolve the problem of providing powerful hardware. Here, the two scenarios were used to improve the performance of the recognition system, i.e. first popular pre-trained CNNs (AlexNet, ResNet-18, VGG-19, and Inception-v3) fine-tuned with the scalogram of two mentioned databases separately, and then best layers are selected as extracted features. Because these CNNs were trained in ImageNet to solve the classification problem with 1000 classes; therefore, the fully connected and classifier layers were replaced by new layers to solve the problem with four classes. Then, all layers from the beginning of CNNs were tuned and classification was done once with the softmax layer and once with MSVM. Softmax is a simple function in the last layer, which decides the probability of belonging the input to one of the classes. The MSVM can solve classification problems strongly. SVM is a supervised method of classification in the machine learning field. It minimizes error iteratively by maximizing marginal hyperplane and benefits from linear and non-linear kernels. It has binary and multiclass versions that here MSVM with the Gaussian kernel was used to classify the four emotional states Q1, Q2, Q3, and Q4. This classifier has been successfully used in EEG signal-processing studies (Chaudhary et al., 2019) (Craik & Contreras-Vidal, 2019). Figure 1 shows the fine-tuning procedure.

Evaluation phase

Four pre-trained CNNs were evaluated with three measures of average (overall) accuracy, precision, and recall (Sokolova & Lapalme, 2009) through the leave-one-subject-out cross-validation (LOSO CV) approach. In this approach, images from 31 subjects are used to fine-tune existing CNNs, and images from another subject are used as the test set to calculate the three mentioned measures this procedure repeats for the other 31 subjects and finally, the average value and standard deviation are reported. LOSO CV is subject-independent because there are no samples from one subject in both test and train sets. Accuracy, precision, and recall are calculated as follows (Sokolova & Lapalme, 2009) (Equations 2-4):

$$2.\ Accuracy=\frac{\sum_{i=1}^{l}\frac{tp_i+tn_i}{+tp_i+tn_i+fp_i+fn_i}}{l}$$

$$3.\ Precision=\frac{\sum_{i=1}^{l}\frac{tp_i}{+tp_i+fp_i}}{l}$$

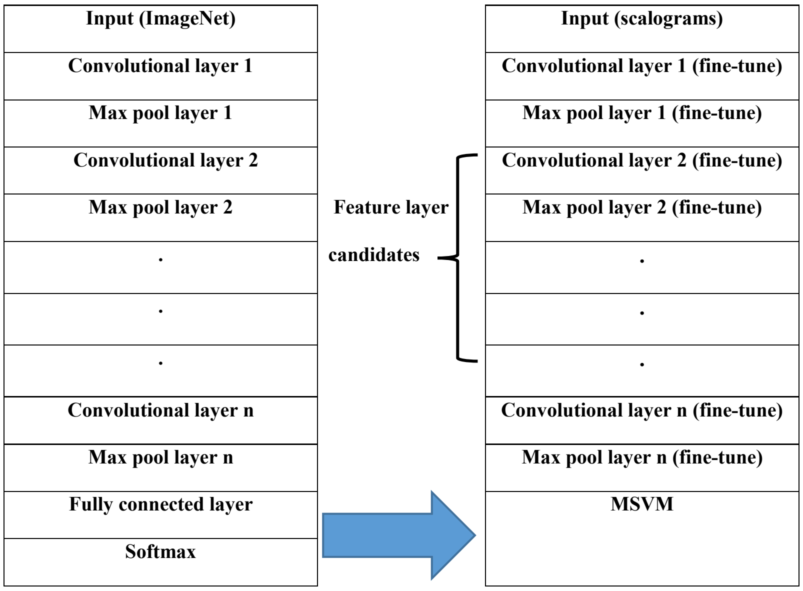


Figure 1. Fine-tuning and optimized feature selection procedures

In the fine-tuning procedure, the scalograms are used as input to the CNNs model, and layers are updated while the fully connected layer and softmax layer are replaced by new ones (1000 classes replaced by 4 classes). Then, the fine-tuned layers are examined to select the best feature layer and then be classified using an MSVM classifier.

$$4.\ Recall=\frac{\sum_{i=1}^{l}\frac{tp}{+tp_i+fn_i}}{l}$$

Where, tp_i, tn_i, fp_i andfn_i are true positive, true negative, false positive, and false negative elements for th emotional class from the confusion matrix obtained from each method.

Summary

Figure 2 shows the flowchart of our method. MAHNOB-HCI and DEAP EEG signals were used; these signals were recorded during watching 20 and 40 video clips and music videos with different emotional states (which cover four-quarters of the valence-arousal emotional model), respectively. In preprocessing step, EEGs were re-referenced by the averaging method and passed through the low pass, high pass, and notch filters, and other artifacts were removed manually by the EEGLAB toolbox in MATLAB software. Then, scalogram images were built using the CWT method from each defined brain region fed into each pre-trained CNNs, and the parameters were tuned. Then, to improve the recognition of emotional states, deep extracted features from different convolutional layers of fine-tuned CNNs were examined, and the best feature layers were applied to MSVM classifier. Also, for improving recognition performance, scalograms from brain regions were combined and results were reported using three subject-dependent and subject-independent evaluation approaches in tables.

3. Results

Thirty-two channels of EEG signals from 230 trials recorded from 20 subjects from the MAHNOB-HCI database and 32 channels from the same number of trials from the DEAP database were preprocessed using the EEGLAB toolbox in MATLAB software (version 2019a). Preprocessing steps were described in the summary section and shown in Figure 2. Then, EEG signals were converted to scalogram images by the CWT method by Morse wavelet. Generally, the Morlet wavelet is used to process EEG signals but we examined all available wavelets (Morse, Morlet, and Bomp), and among their scalogram images, Bomp had lower resolution and there were no considerable differences between the other two. Also, the Morse wavelet can vary two parameters to change time and frequency spread, then, Morse was used to make a scalogram. Thirty-two scalogram images were built from all channels of each subject. Figures. 3 and 4 show the average scalogram for four quarters of the valence-arousal model for MAHNOB-HCI and DEAP databases, respectively. Horizontal and vertical axes represent time (second) and frequency (Hz) contents, respectively.

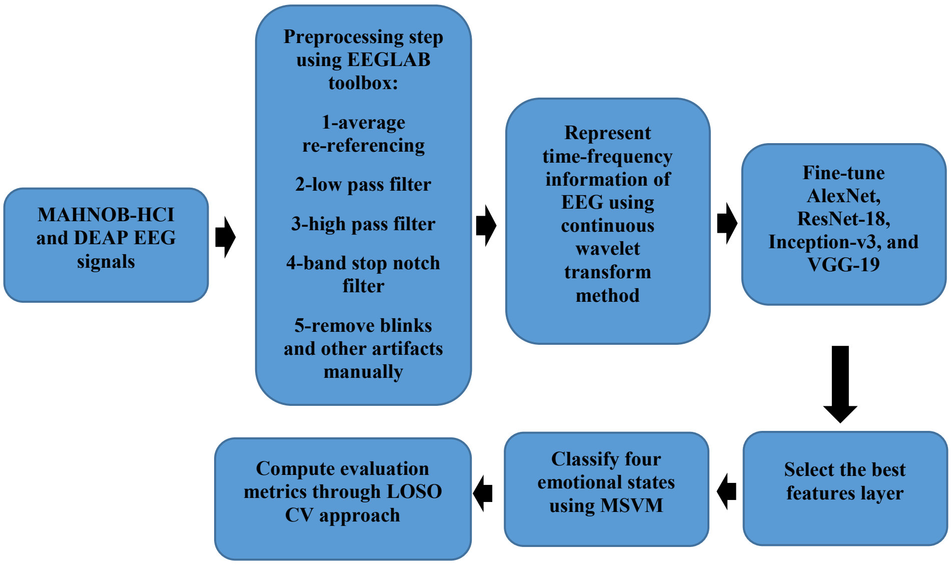

MAHNOB-HCI and DEAP EEG signals
Preprocessing step using EEGLAB toolbox:
1-average re-referencing
2-low pass filter
3-high pass filter
4-band stop notch filter
5-remove blinks and other artifacts manually
Represent time-frequency information of EEG using continuous wavelet transform method
Fine-tune AlexNet, ResNet-18, Inception-v3, and VGG-19
Select the best features layer
Classify four emotional states using MSVM
Compute evaluation metrics through LOSO CV approach

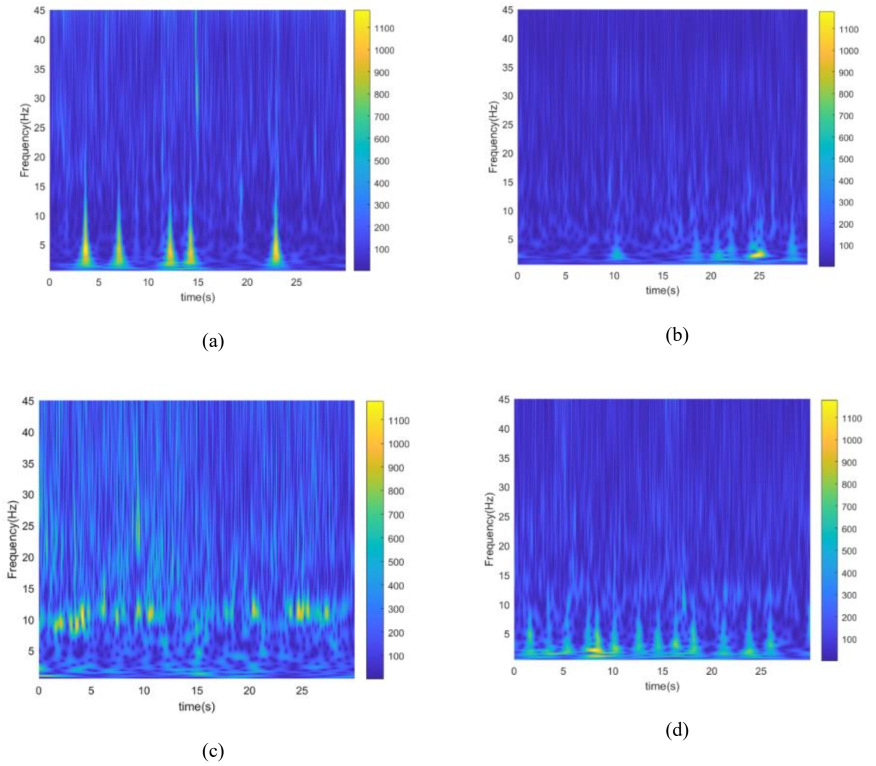
Figure 2. The flowchart of the proposed method

NEUROSCIENCE

Four pre-trained CNNs, Inception-v3, VGG-19, ResNet-18, and AlexNet were fine-tuned, i.e. the fully connected and softmax layers were replaced by new ones with four classes. The initial learning rate, squared gradient decay factor, max epochs, and mini-batch size were 0.0004, 0.99, 20, and 32, respectively. The adaptive moment estimation optimizer (ADAM) was used for training networks. Fine-tuning was performed on 90% of scalogram images of each defined brain region (Table 1) and then, the average accuracy was computed on residual scalograms. All processing steps were done with MATLAB software version 2019a. All codes were implemented on a laptop with an Intel ® CoreTM i7-6500U CPU @2.50 GHz.

Figure 5 shows the average accuracies for nine defined brain regions on MAHNOB-HCI (a) and DEAP (b) databases using the AlexNet, VGG-19, ResNet-18, and Inception-v3 for the LOSO CV evaluation criterion. Comparing MAHNOB-HCI figures, maximum accuracies were achieved for ResNet-18 in the range of 18-31%. Also, the pre-frontal and parietal brain regions achieved higher accuracies using all CNNs. Observing DEAP figures showed that scalograms from two brain regions frontal and parietal using ResNet-18 had achieved higher accuracies of 77% and 74%, respectively. Totally, higher accuracies were achieved for each database using ResNet-18, followed by Inception-v3, VGG-19, and AlexNet.

Due to the low accuracy of recognizing the four mentioned emotional states, the idea of using deep features from these pre-trained CNNs was examined. Also, the classification was done using the MSVM to improve emotion recognition performance. Therefore, each convolutional and pooling layer from the beginning of AlexNet, ResNet-18, VGG-19, and Inception-v3 was examined as a feature vector and fed into an MSVM, and then, each layer, in which had the highest accuracy was selected for further analysis. Table 4 reports the best selected layer for each fine-tuned CNN. The third convolutional layer from AlexNet, named 'Conv3', the first triplet convolutional layer from VGG-19, named 'Conv3_1', the fourth convolutional layer of Inception-v3 named 'Conv2d_4', and the end part of the first residual block of the ResNet-18 named 'Res2a' achieved the highest accuracies for all brain regions in MAHNOB-HCI and DEAP databases. Figure 6 shows accuracy values for brain regions using different fusions of CNN-MSVM for MAHNOB-HCI (up) and DEAP (down) databases for the LOSO CV criterion. As it can be observed, the accuracies using the mentioned procedure increased by nearly 18~24% and 10~12% in MAHNOB-HCI and DEAP databases, respectively. For example, scalograms from the pre-frontal region using ResNet-18-MSVM achieved 56% and 78.5% for MAHNOB-HCI and DEAP databases, respectively, while it was 31.4% and 67.4% using ResNet-18. Also, among all regions, pre-frontal, frontal, and parietal regions achieved higher accuracies for MAHNOB-HCI (video clips), while, frontal and parietal regions achieved higher accuracies for the DEAP database (music videos).



(a)

(b)

(c)

(d)

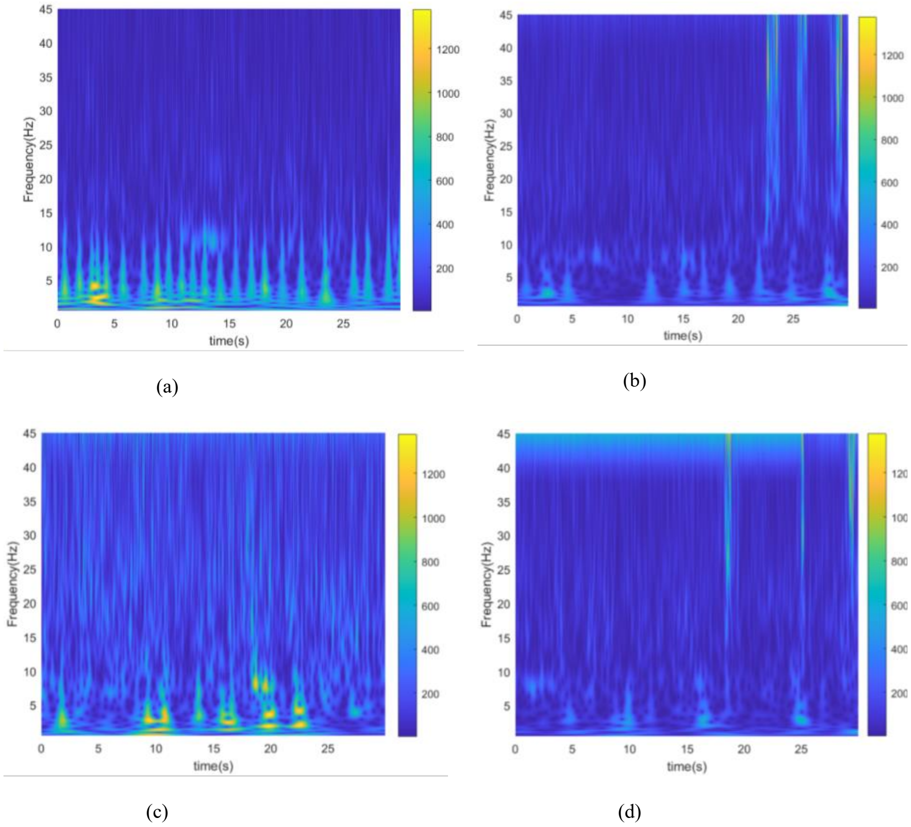
Figure 3. Scalogram images from 4 quarters of valence-arousal model for MAHNOB-HCI database

A) Scalogram averaged among subjects for Q1; B, Q2; C, Q3; and D, Q4.

Furthermore, to find effective brain regions involved in the recognition of emotional states, scalograms from different anatomical regions were combined into two, three, or more forms, i.e. in the first stage, scalograms from all possible two regions were investigated to recognize four desired emotional states, and then in stage two, scalograms from all possible three regions were investigated, and up to the end. Because the highest accuracies were achieved using the extracted features from layer 'Res2a' of ResNet-18 and MSVM classifier, this hybrid method was used for further analysis. Combining scalograms from two possible regions achieved the highest accuracy for the DEAP database and combining three, four, and more possible regions caused no higher accuracy for the DEAP database; thus, only the best results are shown. However, combining scalograms from four brain regions achieved the highest accuracy for the MAHNOB-HCI database. Table 5 reports the six best combinations of two brain regions for classifying the four mentioned emotional classes using the mentioned proposed method and LOSO CV evaluation criterion in the DEAP database. Combining scalograms from two frontal and parietal regions of the DEAP database obtained the highest average accuracy, precision, and recall of 87.76%, 87.31%, and 87.77%, respectively.

Table 4. Best selected layer using MSVM for pre-trained CNNs in selected features procedure for two databases

CNN	Layer Number	Name
AlexNet	10	Conv3
VGG-19	4	Conv3_1
Inception-v3	13	Conv2d_4
ResNet-18	12	Res2a



(a)

(b)

(c)

(d)

Figure 4. Scalogram images from 4 quarters of valence-arousal model for DEAP database

a, Scalogram averaged among subjects for Q1; B, Q2; C, Q3; and D, Q4.

Table 5. Features from Res2a' layer+MSVM classifier

Combining Scalograms From 2 Regions (%)	Mean±SD		
	Accuracy	Precision	Recall
Frontal, central	87.76±0.81	87.31±1.82	87.77±0.78
Frontal, parietal	87.45±2.49	87.34±2.44	87.46±2.50
Frontal, pre-frontal	84.38±2.65	84.18±2.57	84.39±2.55
Frontal-central, parietal	82.15±2.53	82.10±2.63	82.16±3.49
Parietal-occipital, occipital	80.48±2.72	80.36±2.50	80.49±2.70
Frontal, central-parietal	78.30±2.75	78.19±2.69	78.32±2.57
Frontal, central	77.09±2.89	76.68±2.64	77.15±2.65

Six highest accuracies for the combination of scalograms from four brain regions obtained from the ResNet18-MSVM* in the DEAP database.*Features from Res2a' layer+MSVM classifier=footnote.

Table 6. Features from Res2a′ layer+MSVM classifier

Combining Scalograms From 4 Regions (%)	Mean±SD		
	Accuracy	Precision	Recall
Pre-frontal, frontal, parietal, parietal-occipital	77.47±3.40	77.40±3.62	77.52±2.63
Pre-frontal, frontal, frontal-central, parietal-occipital	75.20±3.35	75.16±3.37	75.27±3.40
Pre-frontal, frontal, frontal-central, central	72.66±3.50	72.59±3.32	72.69±3.41
Pre-frontal, frontal, frontal-central, parietal	69.75±3.45	69.58±3.370	69.81±3.75
Pre-frontal, frontal, frontal-central, temporal	69.30±3.50	69.26±3.52	69.38±3.62
Frontal, frontal-central, central, parietal	66.52±3.37	66.36±3.30	66.61±3.53

Six highest accuracies for the combination of scalograms from four brain regions obtained from the ResNet18-MSVM* in the MAHNOB-HCI database. *Features from Res2a′ layer+MSVM classifier=footnote.

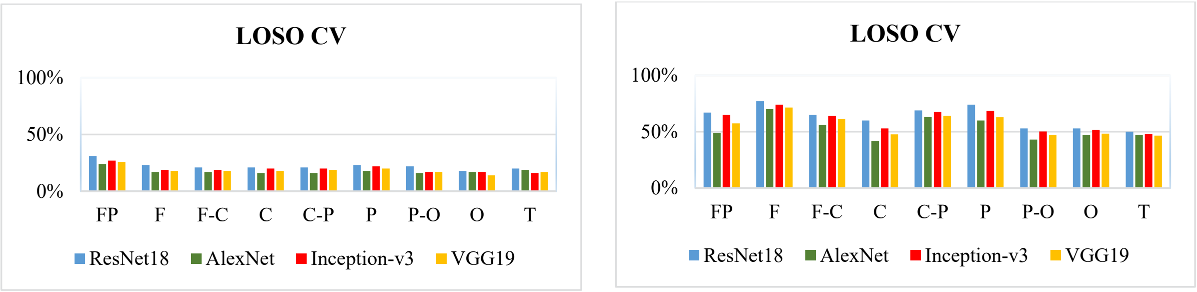
	Q1	Q2	Q3	Q4	
Q1	145	5	6	4	
Q2	5	148	4	3	
Q3	4	1	145	7	
Q4	5	5	4	146	
Acc					91.25%

(a)

	Q1	Q2	Q3	Q4	
Q1	132	8	11	9	
Q2	10	127	9	14	
Q3	9	13	130	8	
Q4	12	7	6	135	
Acc					81.87%

(b)

Table 7. Confusion matrixes of highest results using ResNet-18-MSVM for (a) DEAP and (b) MAHNOB-HCI databases



(a)

(b)

Figure 5. Average accuracy of four emotional recognition using the AlexNet, VGG-19, ResNet-18 and inception-v3 on scalogram images of brain regions for (a) MAHNOB-HCI and (b) DEAP databases using LOSO CV criterion. These results are obtained before selecting optimized features.

Abbreviations FP: Pre-frontal; F: Frontal; F-C: Frontal-central; C: Central; C-P: Central-parietal; P: Parietal; P-O: Parietal-occipital; O: Occipital; T: Temporal.

Table 8. LOVO CV=Leave-one-video-out cross validation

References	Processing Method	Database	Number of Classes	Accuracy (%)
Koelstra & Patras, 2013	PSD, RFE, GNB, 10-fold CV	MAHNOB-HCI	2	66.7 (arousal), 80 (valence)
Zhu et al., 2014	Power spectrum, SVM, LOVO CV[1]	MAHNOB-HCI	2	55.72 (valence), 60.23 (arousal)
Huang et al., 2016	SP, KNN, SVM, LOSO CV	MAHNOB-HCI	2	63 (valence), 65.1 (arousal)
Soroush et al., 2018	Nonlinear features (CD, FD, ...), ICAs, modified Dempster-Shafer theory of evidence, 10-fold CV	DEAP	4	90.54
Soroush et al., 2020	Poincare plane, MSVM, KNN, MLP, 10-fold CV	DEAP	4	89.76
Yang et al., 2018	RQA, Parallel Convolutional Recurrent Neural Network, 5-fold CV	DEAP	3	92.24
Shen et al., 2020	Differential entropy, 4-d-Convolutional recurrent neural network, 5-fold CV	DEAP	2	94.22 (valence), 94.58 (arousal)
Our method	CWT method, ResNet-18-MSVM, LOSO CV	MAHNOB-HCI, DEAP	4	77.47±3.40 (MAHNOB-HCI), 87.45±2.49 (DEAP)

NEUROSCIENCE

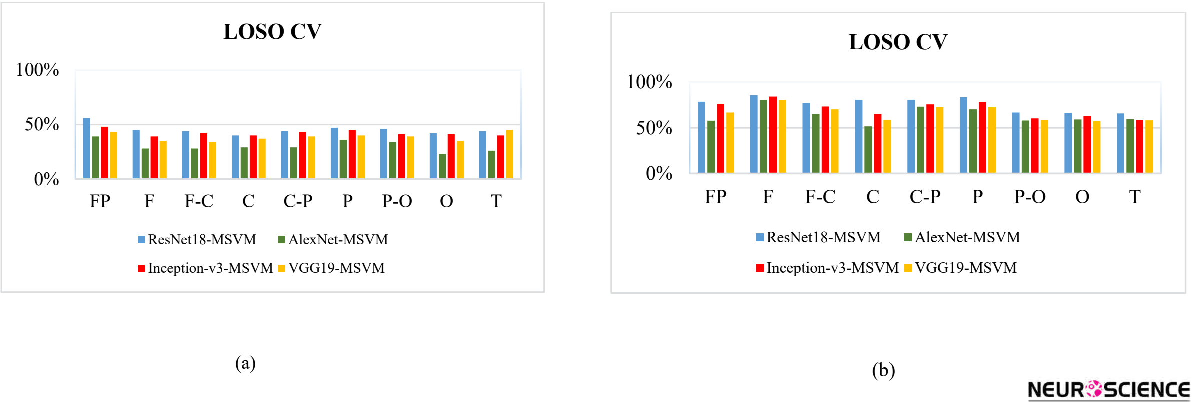
Comparison of this study and related studies with same database. 1LOVO CV=Leave-one-video-out cross validation=Footnote.

Table 6 shows the six highest results from combining scalograms of possible four brain regions from the MAHNOB-HCI database using ResNet-18-MSVM and LOSO CV evaluation criterion. Combination scalograms from four pre-frontal, frontal, parietal, and parietal-occipital regions achieved the highest average accuracy, precision, and recall for the MAHNOB-HCI database and LOSO CV equal to 77.47%, 77.40%, and 77.52%, respectively.

Table 7 reports confusion matrixes of the highest results for ResNet-18-MSVM and LOSO CV evaluation criterion and both databases.

4. Discussion

Four emotional states from EEG signals were recognized through scalogram images using the CWT method and extracted features from fine-tuned pre-trained CNNs



(a)

(b)

NEUROSCIENCE

Figure 6. Average accuracy recognition four emotional classes using the AlexNet-MSVM, VGG-19-MSVM, ResNet-18-MSVM and Inception-v3-MSVM from scalogram images of various brain regions for (A) MAHNOB-HCI and (B) DEAP databases using LOSO CV criterion. These results were obtained after feature selection method using MSVM and Gaussian kernel

Abbreviations: FP: Pre-frontal; F: Frontal; F-C: Frontal-central; C: Central; C-P: Central-parietal; P: Parietal; P-O: Parietal-occipital; O: Occipital; T: Temporal.

and MSVM classifiers. CNNs were trained in a very huge image database (ImageNet) with 1000 categories but their categories included objects, animals, and other things except for biomedical signals. Therefore, fine-tuning of CNNs parameters based on scalogram images of four emotional states from EEG signals help the network parameters to be compatible with the specific emotion recognition problem. Using MSVM as the classifier is reasonable, as this was the best method to discriminate classes before the improvement of deep learning methods. Also, as can be observed from Figures 5 and 6, extracted deep features from fine-tuned CNNs and MSVM classifiers improved results by nearly 12% and 20% for DEAP and MAHNOB-HCI databases, respectively. Selected features for each CNN were from earlier deep layers, these layers extract low-level features than deeper layers. Also, among different pre-trained CNNs+MSVM configurations, the ResNet-18+MSVM achieved the highest accuracies for all brain regions. ResNet-18 consisted of multiple residual units that are stacked identity maps and shortcuts, while, Inception-v3 has multiple parallel convolutional layers in its Inception units, or AlexNet and VGG-19 have simple convolutional layers. Therefore, according to the results on accuracy for all evaluation criteria, it seems that extracted features from the residual unit perform better than the Inception module or simple form of the convolutional layer to solve this emotion recognition task. Then, Inception-v3 and VGG-19 had better results compared to the AlexNet.

One of our goals was to find effective brain regions to recognize emotional states using the proposed method; thus, we considered all brain regions and possible combinations of them at several levels. According to Tables 5 and 6, the combination of scalograms from pre-frontal, frontal, parietal, and parietal-occipital regions achieved the highest average accuracy among other combinations for the MAHNOB-HCI database that was recorded during watching ordinary video clips. This means that these regions are the most related regions in recognition of the four mentioned emotional classes. Also, frontal and parietal regions had higher accuracy for all evaluation criteria for the DEAP database that was recorded during watching music videos. According to neuroimaging studies, the limbic system is responsible for emotions (Rolls, 2015). The limbic system is in the amygdaloid nuclear complex, ventral nuclei of Gudden, and central gray and dorsal raphe nucleus (Morgane et al., 2005). In addition, our findings about the best regions are consistent with related studies with other methods (Alarcao & Fonseca, 2017; Rolls, 2015). Moreover, these results show that the type of stimulation (music videos or ordinary video clips) influences the involved brain regions.

Among evaluation metrics, subject-dependent criteria, such as 10-fold cross-validation (10-fold CV) may use samples of one subject for both train and test sets and cause higher accuracy, but subject-independent criteria, such as LOSO CV, only use samples from one subject at the test set as unknown samples and samples from other subjects as a train set, this causes you to consider inter-subject varieties. However, this can cause lower accuracy for LOSO CV, but the decrease shows how different the emotional states are between subjects. As we know, emotions vary somewhat across subjects, genders, years, and cultures; however, researchers from MAHNOB-HCI and DEAP databases tied to consider these factors, but aside from these factors, it is difficult to recognize human emotions and distinguish them. The highest average accuracy was 87.45% for DEAP and 77.43% for MAHNOB-HCI, while inter-subjects differences were 2.49% and 3.40%, respectively. Also, LOSO CV proved the generalization ability of the proposed method and the results were reasonable and acceptable.

In Table 8, the results of this study are compared with related studies that used EEG signals of MAHNOB-HCI and DEAP databases. Among these studies, only Zhu et al. (2014) and Huang et al. (2016) used the LOSO CV criterion and others used the 10-fold CV criterion. As observed, the accuracy of our study was higher than these two studies, they used the traditional machine learning methods of extraction of features that prove the preference of the proposed method. Extracted features from deep layers provide appropriate and discriminative features than features, such as PSD (or other spectral features) (Koelstra & Patras, 2013; Zhu et al., 2014; Huang et al., 2016). Also, our values are higher than those of Koelstra & Patras (2013), evaluated using the 10-fold CV criterion. However, our values are lower than other studies mentioned in Table 8 (Soroush et al., 2020; Soroush et al., 2018; Yang et al., 2018; Shen et al., 2020), but the increase is due to their evaluation criteria. As mentioned above, in the subject-dependent 10-fold CV evaluation criterion, as the model can see each sample through folds, the performance is usually higher than LOSO CV. Another cause of the difference in the results is the number of discriminated classes, all of these studies, except Soroush et al., (2018) and Soroush et al., (2020) had two or three emotional classes.

5. Conclusion

In this paper, four emotional states were recognized using a hybrid EEG-based approach: scalogram images built by the CWT method, extracted deep features from popular pre-trained CNNs, and MSVM classifier. Among extracted features from pre-trained CNNs, deep features of the early convolutional layer of ResNet-18 were selected, and combining some brain regions and well performance of MSVM caused improvement of the emotion recognition system from EEG signal. The results were promising and can be used in other fields of neuroscience. In the future study, we will consider new methods to build images based on brain connectivity measures.

Ethical Considerations

Compliance with ethical guidelines

There were no ethical considerations to be considered in this research.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

Conceptualization, methodology, investigation, writing, review, and editing: All authors; Writing the original draft: Sara Bagherzadeh; Supervision: Keivan Maghooli, Ahmad Shalbaf, and Arash Maghsoudi.

Conflict of interest

The authors declared no conflict of interest.

Acknowledgments

We would like to thank the Islamic Azad University, Science and Research Branch, due to the provided support

References

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adeli, H. (2018). Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in Biology and Medicine, 100,* 270–278. [DOI:10.1016/j.compbiomed.2017.09.017] [PMID]

Afshani, F., Shalbaf, A., Shalbaf, R., & Sleigh, J. (2019). Frontal-temporal functional connectivity of EEG signal by standardized permutation mutual information during anesthesia. *Cognitive Neurodynamics, 13*(6), 531-540. [DOI:10.1007/s11571-019-09553-w] [PMID] [PMCID]

Alarcao, S. M., & Fonseca, M. J. (2017). Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing, 10*(3), 374-393. [DOI:10.1109/TAFFC.2017.2714671]

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* Massachusetts: MIT press. [Link]

Chaudhary, S., Taran, S., Bajaj, V., & Sengur, A. (2019). Convolutional neural network based approach towards motor imagery tasks EEG signals classification. *IEEE Sensors Journal, 19*(12), 4494-4500. [DOI:10.1109/JSEN.2019.2899645]

Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: A review. *Journal of Neural Engineering, 16*(3), 031001. [PMID]

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., & Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine, 161,* 1-13. [PMID]

García-Martínez, B., Martinez-Rodrigo, A., Alcaraz, R., & Fernández-Caballero, A. (2019). A review on nonlinear methods using electroencephalographic recordings for emotion recognition. *IEEE Transactions on Affective Computing, 12*(3), 801 - 820. [DOI:10.1109/TAFFC.2018.2890636]

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing, 187,* 27-48. [DOI:10.1016/j.neucom.2015.09.116]

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition.* Paper presented at: IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27-30 June 2016. [DOI:10.1109/CVPR.2016.90]

Huang, X., Kortelainen, J., Zhao, G., Li, X., Moilanen, A., & Seppänen, T., et al. (2016). Multi-modal emotion analysis from facial expressions and electroencephalogram. *Computer Vision and Image Understanding, 147,* 114-124. [DOI:10.1016/j.cviu.2015.09.015]

Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., & Ebrahimi, T., .et al. (2011). Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing, 3*(1), 18-31. [DOI:10.1109/T-AFFC.2011.15]

Koelstra, S., & Patras, I. (2013). Fusion of facial expressions and EEG for implicit affective tagging. *Image and Vision Computing, 31*(2), 164-174. [DOI:10.1016/j.imavis.2012.10.002]

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84-90. [DOI:10.1145/3065386]

Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik, 29*(2), 102-127. [PMID]

Morgane, P. J., Galler, J. R., & Mokler, D. J. (2005). A review of systems and networks of the limbic forebrain/limbic midbrain. *Progress in Neurobiology, 75*(2), 143-160. [DOI:10.1016/j.pneurobio.2005.01.001] [PMID]

Rolls, E. T. (2015). Limbic systems for emotion and for memory, but no single limbic system. *Cortex, 62,* 119-157. [DOI:10.1016/j.cortex.2013.12.005]

Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T. H., & Faubert, J. (2019). Deep learning-based electroencephalography analysis: A systematic review. *Journal of Neural Engineering, 16*(5), 051001. [DOI:10.1088/1741-2552/ab260c]

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and Social Psychology, 39*(6), 1161-1178. [DOI:10.1037/h0077714]

Shalbaf, A., Shalbaf, R., Saffar, M., & Sleigh, J. (2020). Monitoring the level of hypnosis using a hierarchical SVM system. *Journal of Clinical Monitoring and Computing, 34*(2), 331-338. [DOI:10.1007/s10877-019-00311-1]

Shalbaf, A., Bagherzadeh, S., & Maghsoudi, A. (2020). Transfer learning with deep convolutional neural network for automated detection of schizophrenia from EEG signals. *Physical and Engineering Sciences in Medicine,43,* 1229–1239. [DOI:10.1007/s13246-020-00925-9]

Shen, F., Dai, G., Lin, G., Zhang, J., Kong, W., & Zeng, H. (2020). EEG-based emotion recognition using 4D convolutional recurrent neural network. *Cognitive Neurodynamics, 14*(6), 815-828. [DOI:10.1007/s11571-020-09634-1]

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. [DOI:10.48550/arXiv.1409.15561]

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2011). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing, 3*(1), 42-55. [DOI:10.1109/T-AFFC.2011.25]

Soleymani, M., & Pantic, M. (2013). *Multimedia implicit tagging using EEG signals.* Paper presented at: 2013 IEEE International Conference on Multimedia and Expo (ICME), San Jose, CA, USA, 15-19 July 2013. [DOI:10.1109/ICME.2013.6607623]

Soleymani, M., Asghari-Esfeden, S., Fu, Y., & Pantic, M. (2015). Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing, 7*(1), 17-28. [DOI:10.1109/TAFFC.2015.2436926]

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437. [DOI:10.1016/j.ipm.2009.03.002]

Soroush, M. Z., Maghooli, K., Setarehdan, S. K., & Nasrabadi, A. M. (2018). A novel approach to emotion recognition using local subset feature selection and modified Dempster-Shafer theory. *Behavioral and Brain Functions, 14*(1), 17. [DOI:10.1186/s12993-018-0149-4] [PMID] [PMCID]

Soroush, M. Z., Maghooli, K., Setarehdan, S. K., & Nasrabadi, A. M. (2020). Emotion recognition using EEG phase space dynamics and Poincare intersections. *Biomedical Signal Processing and Control, 59,* 101918. [DOI:10.1016/j.bspc.2020.101918]

Sun, W., Tseng, T. B., Zhang, J., & Qian, W. (2017). Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Computerized Medical Imaging and Graphics, 57,* 4-9. [DOI:10.1016/j.compmedimag.2016.07.004] [PMID]

Suzuki, K. (2017). Overview of deep learning in medical imaging. *Radiological Physics and Technology, 10*(3), 257-273. [DOI:10.1007/s12194-017-0406-5] [PMID]

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). *Rethinking the inception architecture for computer vision.* Paper presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA 27-30 June 2016. [DOI:10.1109/CVPR.2016.308]

Yang, Y. X., Gao, Z. K., Wang, X. M., Li, Y. L., Han, J. W., & Marwan, N., et al. (2018). A recurrence quantification analysis-based channel-frequency convolutional neural network for emotion recognition from EEG. *Chaos (Woodbury, N.Y.), 28*(8), 085724. [DOI:10.1063/1.5023857] [PMID]

Zhang, X., Yao, L., Wang, X., Monaghan, J. J., Mcalpine, D., & Zhang, Y. (2021). A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. *Journal of Neural Engineering, 18*(3), 10.1088/1741-2552/abc902. [DOI:10.1088/1741-2552/abc902][PMID]

Zhu, Y., Wang, S., & Ji, Q. (2014). *Emotion recognition from users' eeg signals with the help of stimulus videos.* Paper presented at: 2014 IEEE International Conference on Multimedia and Expo (ICME), Chengdu, China, 14-18 July 2014. [DOI:10.1109/ICME.2014.6890161]

This Page Intentionally Left Blank