IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Recognizing Emotions Evoked by Music using CNN-LSTM Networks on EEG signals

S. Sheykhivand[1], Z. Mousavi[2], T. Yousefi Rezaii[1], and A. Farzamnia[3] (Senior Member IEEE)

[1]Biomedical Engineering Department, Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran
[2]Department of Mechanical Engineering, Faculty of Mechanical Engineering, University of Tabriz, Tabriz, Iran
[3]Faculty of Engineering, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia

Corresponding author: A. Farzamnia (e-mail: ali-farzamnia@ieee.org) and T. Yousefi Rezaii (e-mail: yousefi@tabrizu.ac.ir).

This work was supported by Research and Innovation Management Center (PPPI) and Faculty of Engineering, Universiti Malaysia Sabah (UMS)

ABSTRACT Emotion is considered to be critical for the actual interpretation of actions and relationships. Recognizing emotions from EEG signals is also becoming an important computer-aided method for diagnosing emotional disorders in neurology and psychiatry. Another advantage of this approach is recognizing emotions without clinical and medical examination, which plays a major role in completing the Brain-Computer Interface (BCI) structure. Emotions recognition ability, without traditional utilization strategies such as self-assessment tests, is of paramount importance. EEG signals are considered the most reliable technique for emotions recognition because of the non-invasive nature. Manual analysis of EEG signals is impossible for emotions recognition, so an automatic method of EEG signals should be provided for emotions recognition. One problem with automatic emotions recognition is the extraction and selection of discriminative features that generally lead to high computational complexity. This paper was design to prepare a new approach to automatic two-stage classification (negative and positive) and three-stage classification (negative, positive, and neutral) of emotions from EEG signals. In the proposed method, directly apply the raw EEG signal to the convolutional neural network and long short-term memory network (CNN-LSTM), without involving feature extraction/selection. In prior literature, this is a challenging method. The suggested deep neural network architecture includes 10-convolutional layers with 3-LSTM layers followed by 2-fully connected layers. The LSTM network in a fusion of the CNN network has been used to increase stability and reduce oscillation. In the present research, we also recorded the EEG signals of 14 subjects with music stimulation for the process. The simulation results of the proposed algorithm for two-stage classification (negative and positive) and three-stage classification (negative, neutral and positive) of emotion for 12 active channels showed 97.42% and 96.78% accuracy and Kappa coefficient of 0.94 and 0.93 respectively. We also compared our proposed LSTM-CNN network (end-to-end) with other hand-crafted methods based on MLP and DBM classifiers and achieved promising results in comparison with similar approaches. According to the high accuracy of the proposed method, it can be used to develop the human-computer interface system.

INDEX TERMS Emotions Recognition, CNN, LSTM, EEG.

I. INTRODUCTION

Emotion is a physiological excitement mood that one finds in an emotional state. This theory supports other cognitive assessment theories that, in the experience of emotion, six dimensions of cognitive evaluation of situations are presented, including commitment, control, certainty, attention, control, and expectance of the situation and pleasure [1]. Also, there are several important definitions and theories about human emotions. According to James Long's theory, emotional experience is a response to physiological changes in the body. According to James Long's theory, emotional experience is a response to physiological changes in the body. Any emotion is an interpretation of its previous excitation. Therefore, the knowledge of the physiological reaction of every emotion is important to the emotion analysis [2]. Russell suggested one prevailing hypothesis, which stated that emotion consists of two arousals and valence elements [3]. Arousal denotes the emotional activation level, whereas valence identifies positivity or negativity. This representation depicts emotions systematically and is commonly used as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011882, IEEE Access

background knowledge in countless studies, including current work.

A large body of research has been devoted to exploring the neural correlates of emotions to build devices that can interpret emotions. In these efforts, different pictorial [4], musical [5-7], music video and video [8-10] triggers have elicited emotions. Music listening encompasses a range of psychological processes, such as multimodal and perception integration, syntactic processing and encoding of sense knowledge, sentiment, concentration, emotion, attention and social cognition [11]. Also, music will bring out strong emotions [12]. Emotional processing includes the various structures of the human brain and changes their operation [11]. It also induces some other physiological responses that are side effects of brain activity, such as changes in heart rate [13], skin conductance, and body temperature [14]. Researchers have used different modalities such as PET [15], fMRI [16], NIRS [17], and EEG [18] to study neural correlates of emotion. High temporal resolution, non-invasive availability, portability, and relatively low data processing costs have made EEG an effective candidate for studying the neural correlates of different cognitive functions such as emotion. Various computational methods based on EEG signals have been developed for the observation and analysis of automatic emotions recognition, which will be discussed below.

Balconi et al. [19] employed the EEG frequency bands in combination with hemodynamic testing to analyze brain reactions. Sammler et al. [13] found that listening to enjoyable music in the frontal midline region increases theta power of EEG signals. Balasubramanian et al. [7] also found a higher frontal midline theta band energy for liked songs. They also studied the strength of EEG components collected by decomposing the wavelet packet when listening to liked and disliked music. Zheng et al. [20] used EEG spectral power as a feature in the discovery and emotional identification of EEG channels by coarse canonical correlation analysis. Ozel et al. [21] implemented multivariate synchrosqueezing transform to extract EEG features for emotional state recognition. For one of the emotional states, they achieved 93% classification accuracy. Hasanzadeh et al. [22] used a nonlinear autoregressive model and genetic algorithm to predict emotional states by the EEG power range when listening to music. Soleymani et al. [23] developed a multimodal database to investigate emotions recognition. Based on this study, they measured the correlation between the electrode's EEG spectral power and valence scores and found that higher frequency components on the frontal, parietal, and occipital lobes had a higher correlation with the valence response based on self-evaluation. They also fused the PSD and facial features to improve the classification performance for emotions recognition. Lin et al. [6] validated the emotion-specific features based on EEG power spectral changes and assessed the relation between EEG dynamics and emotional states triggered by music. Koelstra et al. [24] introduced the dataset for emotion analysis using physiological signals (DEAP) for human affective states, and among 32 participants extracted the spectral power feature of five frequency bands. They noticed that the frontal and parietal lobe features can provide discriminative emotion-producing information. Chanel et al. [25], employed the Naive Bayes classifier to classify three emotion classes from specific frequency bands at specific electrode locations. Zheng et al. [26] provided the SJTU emotion EEG dataset (SEED) for emotions recognition. During their study, the authors extracted six different spectral features to analyze the neural signatures for positive, negative, and neutral emotions; by this, they found that the EEG patterns were relatively stable within and between sessions at sensitive frequency bands and brain regions. Therefore, Changes in EEG power spectral can predict distinct emotional states of subjects in different brain regions. Thammasat et al. [27] used physiological signals to extract power density spectra and fractal dimensions for emotion analysis and found that using less familiar music improved the accuracy of recognition regardless of whether the classifier was support vector (SVM) or multilayer perception. Kumagai et al. [28] evaluated the relationship between cortical response and music familiarity. They found that when listening to scrambled or unfamiliar music, the two peaks of the cross-correlation values were significantly larger than familiar ones. Such results collectively indicate that the cortical response to unfamiliar music is greater than familiar music and is, therefore, highly valued by BCIs programs for classification applications. Zhao et al. [29] analyzed the volunteers' EEG signals when they were watching effective films. SVM has been used as a classifier after extracting EEG features to recognize human emotions. Lu et al. [30] selected nine musical passages as stimulus and divided them into three categories according to the two-dimensional model of emotions using the variance test and t-test. To evaluate EEG signals, the power density spectra were derived from different bands, and the dimensionality reduction of the principal component analysis (PCA) was used to select the features. They found that the accuracy of SVM classifier emotions recognition was higher than other bands using average beta and gamma-band power information. Therefore, beta and gamma bands appear to be effective for emotional discrimination. Yoon and Chung [31] suggested a classifier based on Bayes' theorem that used supervised learning algorithms to classify human emotions based on EEG signals from volunteers. In this work, Fast Fourier Transform (FFT) analysis was used to extract the feature in recognition. Li et al. [32] extracted 816 features from 16 electrodes and improved the recognition rate by reducing CFS dimensions and machine learning. Signals related to electrode position of O2, Fp1, F3, T3, and Fp2 were also suggested to be most closely associated with mild depression. Yimin et al. [33] used a Correlation Based Feature Selection (CFS) model to extract features from EEG signals to classify different

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011882, IEEE Access

emotions (relaxation, happiness, sadness, and grief) into 8 subjects. They used BP, SVM, LDA, and C4.5 classifiers for the classification, and concluded that the C4.5 classifier works better for emotions recognition than other classifiers. Fatemeh et al. [34] used a fuzzy parallel cascade (FPC) to predict a continuous subjective assessment of the emotional content of music from the EEG signals on 15 subjects. They also compared the FPC model with LSTM and linear regression (LR) models. The RMSE of their model was about 0.089. Their proposed model RMSE was lower than the other models for estimating both valence and arousal. Panayu et al. [35] used CNN to extract features from the EEG signals for arousal and valence classification into 12 subjects. Their network architecture consisted of six layers of convolution. They compared their proposed algorithm with SVM and concluded that CNN was better in emotions recognition. Yang et al. [36] used a hybrid neural network that combined CNN and the Recurrent Neural Network (RNN) to automatic emotions recognition from EEG signals. In their experiments, these researchers used the DEAP benchmarking dataset. Also, in their proposed method, 1D EEG signals have been converted to 2D EEG frames. The reported accuracy for both valence and excitement classes is reported to be 90.80% and 91% respectively. Yang et al. [37] used a multi-column structured CNN model for emotions recognition. The DEAP database was also used by these researchers. The final accuracy reported for their multi-column model for the classification of valence and arousal is reported to be 90% and 90%, respectively. Chen et al. [38] used parallel hybrid convolutional recurrent neural networks to classify the binary emotions of EEG signals. The DEAP database was also used by these researchers. The final accuracy reported for the classification of valence and arousal is reported to be 93.64% and 93.26%, respectively. Chen Wei et al. [39] used the Dual-tree Complex Wavelet Transform (DT-CWT) for feature extraction on EEG signals. These researchers also used the Simple Recurrent Units (SRU) network to train emotion models. This method resulted in an average accuracy of 85.65%, 85.45%, and 87.99%, respectively, with low/high arousal, valence, and liking classes. The challenging step in emotions recognition is to select the discriminative features of different stages of emotion. In the works that are most available, first, statistical features are extracted, and then the best discriminatory features are selected manually or using common feature selection methods, which is a time-consuming method that requires high complication complexity. Also, for one case, the best features may not be regarded as optimal in another. Therefore, implementing an algorithm that learns the correct features corresponding to each case is essential. This will remain as the main benefit of this research.

In the proposed method, pre-processing operations were performed on the data after recording data with the auditory stimulation. Then a fusion of a deep convolutional network and an LSTM network is used to train two-stage classification (positive and negative) and three-stage classification of emotions (positive, neutral and negative). The approach suggested can be used as an end-to-end classifier, in which there is no need for a feature selection/extraction method and the correct features of each class will be automatically learned with a deep neural network.

The remainder of the paper is structured as follows: The experimental database based on the stimulation of the auditory, the related mathematical background of CNN and LSTM network are given in Section II. The proposed method is presented in Section III. The simulation results and comparison of the proposed method with the common methods are given in Section IV; finally, section V is as to the conclusion.

II. MATERIALS AND METHODS

In this section, we first introduce the experiments of collecting EEG data based on the stimulation of the auditory at the University of Tabriz. Then, the mathematical background of CNN and LSTM will be provided.

A. EEG COLLECTING:

A database of three positive, neutral and negative emotions was created to recognize emotions from the EEG signal. The nine-degree version of the Self-Assessment Manikin (SAM) test was also used in the testing process. In this test, a score below 3 is considered to be low and a score over 6 is considered high. Before recording the signal, all participants were asked to sign their consent form (no history of mental illness, no history of epilepsy, no use of psychiatric drugs, normal sleep at night, no fatty food, no pre-test caffeine, and no pre-test hair washing). Participants were asked to complete a beck depression inventory (BDI). After completing the questionnaire, participants who had scored more than 21 on the test were excluded from the processing and conclusion process by the psychological standards. To collect the database, 16 people (6 females and 10 males) between the ages of 20 and 28 were invited to participate in the experiment. Participants' EEG signal was recorded while listening to music. All EEG signals were recorded at 29°C between 9 and 11 a.m. to ensure that the participants were not tired. Also, to avoid EOG noise, all subjects were asked to keep their eyes closed during the signal recording process. The experiment was conducted using Encephalan 21-channel EEG recorder and Macbook air 2017 (Corei5 and 8 Ram). All channel data were referenced to the two reference electrodes A1 and A2 and digitized from an international 10-20 system-based 21-channel electrode cap at 250 Hz. Fig. 1. shows the EEG recording while the participant is listening to music. The descriptive results of the BDI and SAM tests have been shown in Table 1. For example, according to Table 1, subject 3 was excluded from the processing-process due to a mismatch in the SAM test. Subject 1 entered the process with a mean dimension of positive emotional

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011882, IEEE Access

capacity 9 (greater than 6) and a mean dimension of negative emotional induction (less than 9) and a BDI test Scale of 16

FIGURE 1. **EEG recording while listening to music.**

TABLE I
VALIDATION OF SUBJECTS IN THE EEG SIGNAL RECORDING PROCESS FOR EMOTIONS RECOGNITION.

Su	Sex	Age	BDI	Mean valence of induction for positive emotion	Mean arousal of induction for positive emotion	Mean valence of induction for negative emotion	Mean arousal of induction for negative emotion	Result of validation	Reason for removal subject
1	M	25	16	9	9	1.8	1	✓	-
2	M	24	22	6.8	6.2	3.6	2	✗	Beck Depression (21 <22)
3	F	27	19	6.2	7.4	4.2	4.6	✗	Mismatch of the control question in the SAM test
4	M	24	4	7.4	7.6	2.4	2.6	✓	-
5	M	24	0	5.8	5	4.4	5.6	✗	Mismatch of the control question in the SAM test
6	M	28	10	5.6	5.4	2	1.6	✗	The desired lack of induction in the positive emotional class
7	M	28	13	7.2	7.4	3.8	3.8	✗	The desired lack of induction in the negative emotional class
8	M	20	19	7.8	7.4	2.8	3	✓	-
9	M	26	9	7.4	7	3.4	5.4	✗	The desired lack of induction in the negative emotional class
10	F	23	9	6.8	6.6	3.8	3.2	✗	The desired lack of induction in the negative emotional class
11	F	25	22	7.8	8	4.5	3	✗	Beck Depression (21 <22)
12	F	27	1	8.6	8.6	2	1.2	✓	-
13	F	29	9	6	6	2	1.2	✓	-
14	M	26	8	8	8	1.8	1.6	✓	-
15	F	25	12	-	-	-	-	✗	Motion Noise
16	M	27	0	7.4	8	1.8	2	✓	-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011882, IEEE Access

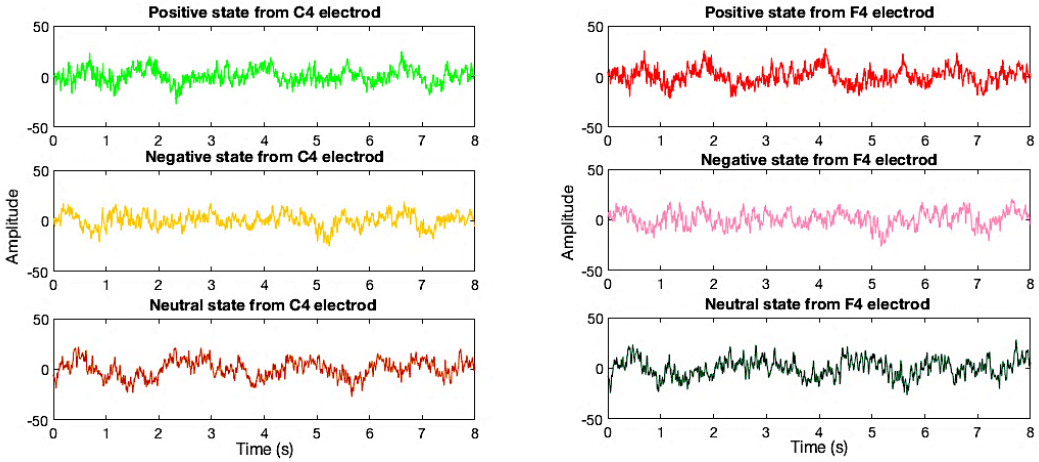

FIGURE 4. **Part of the EEG signal for positive, negative and neutral stages of C4 and F4 channels for subject 1.**

(16 < 21). Also, details of the validation result of the SAM test have been shown in Fig. 2. Music stimulation is used to stimulate positive and negative emotions in participants. Each music track is played for 1 minute and pauses for 15 seconds to prevent any transfer of emotion between the music tracks. It is also considered a neutral state in the processing process. Headphones are also used to play the songs (to induce more). After listening to the music stimulus, each subject immediately began listening and refilling out the questionnaire. Overall, the whole test process takes about 720 seconds. The theme and mood of music have a general and physiological effect on everyone with different mental and emotional mechanisms. But the magnitude and severity of this effect depend on the condition of the neurons, the mental history and the habit of the listener. According to the Iranian nationality of the participants, and also studies the sad theme has been chosen for negative emotion induction and the historical theme for positive emotions. Iranian music tracks were used for each

emotional theme. Table. 2 shows the details of each selected music and Fig. 3. shows the order of the played music tracks. Fig. 4. also shows samples of EEG signals for the three-stage of emotion for C4 and F4 channels on subject 1.

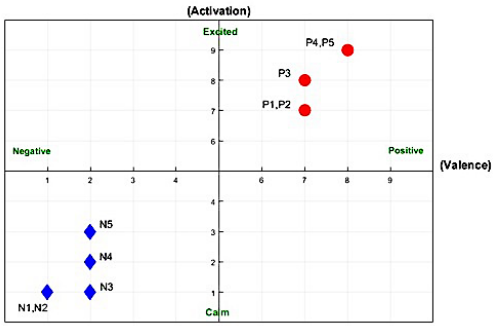

FIGURE 2. **Validation of SAM test.**

TABLE II
THE ORDER AND TYPE OF MUSIC USED TO STIMULATE EMOTIONS.

Abbreviation	Definition	Music Title
N1	First negative emotion	Esfahan prologue mohammad reza lotfi
P1	First positive emotion induction	Turkish music azarbaijan
N2	Second negative emotion	Homayoun prologue by Faramarz payvar
P2	Second positive emotion induction	Turkish music azarbaijan
P3	Third positive emotion induction	Bandari music
N3	Third negative emotion	Afshari piece by sohrab pournazeri
N4	Fourth negative emotion	Esfahan prologue mohammad reza lotfi
P4	Fourth positive emotion induction	Bandari music
N5	Fifteenth negative emotion	Dashti prologue by hosein alizade & keyhan kalhor
P5	Fifteenth negative emotion	Bandari music

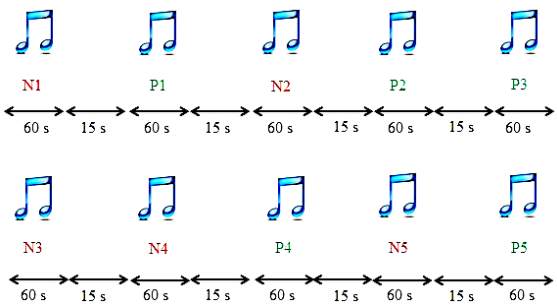

FIGURE 3. **The duration and order of the music tracks.**

B. DEEP CONVOLUTIONAL NEURAL NETWORK:

CNN is a better replacement for the traditional neural network, which is very effective and develops classification methods in the field of machine vision [40]. There are 2 phases for learning in CNN; feed-forward and backpropagation (BP) phase [41].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011882, IEEE Access

CNN consists of three main layers, namely, convolutional, pooling and fully connected (FC) layers [41-43]. The output of the convolution layer is called the feature mapping. In this research, the max-pooling layer has been used, which selects only the maximum values in each feature map. The dropout technique is used to avoid the overfitting; therefore, according to a probability, each neuron is thrown out of the network at each stage of training, which results in decrease network. The batch normalization (BN) layer is used to normalize the data inside the network. The BN transformation is given as follows:

(1)

$$\hat{\mathbf{y}}^{(\mathbf{l-1})} = \frac{\mathbf{y}^{*(\mathbf{l-1})} - \mu_B}{\sqrt{(\sigma_B^2 + \varepsilon}}$$
$$\mathbf{z}^{*(\mathbf{l})} = \gamma^{(l)}\hat{\mathbf{y}}^{(\mathbf{l-1})} + \beta^{(l)}$$

where $\mathbf{y}^{*(\mathbf{l-1})}$ is the input vector to the BN layer, $\mathbf{z}^{*(\mathbf{l})}$ is the output response related to a neuron in layer l, $\mu_B = \mathrm{E}[\mathbf{y}^{*(\mathbf{l-1})}]$, $\sigma_B^2 = \mathrm{var}[\mathbf{y}^{*(\mathbf{l-1})}]$, ε represents a small constant for numerical stability, $\gamma^{(l)}$ and $\beta^{(l)}$ are the parameters of scale and shift, respectively, which are obtained by learning. An activation function is applied after each layer. In this study, Relu and Softmax, as two-types of activation functions, are used. Relu, as it has been defined in (2), is used in the convolutional layers and has the ability to apply nonlinearity and sparseness to the network structure.

$$\mathbf{R}(d) = \begin{cases} d & if\ d > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The probable distribution of the output classes can be calculated by a Softmax activation function. Therefore, the Softmax function is used in the last FC layer and is defined as follows:

$$\sigma(\boldsymbol{\delta})_i = \frac{e^{\delta_i}}{\sum_{j=1}^{k} e^{\delta_j}} \quad \text{for } i = 1, \ldots k \text{ and } \boldsymbol{\delta} = (\delta_1, \ldots, \delta_k) \in R^k \quad (3)$$

Where $\boldsymbol{\delta}$ is the input vector and the output values $\sigma(\boldsymbol{\delta})$ are between 0 and 1, which their sum is equal to 1 [41-43].

C. LONG SHORT-TERM MEMORY (LSTM):

Recurrent neural networks (RNN) are widely used to deal with variable-length sequence inputs. The long-distance history is stored in a recurrent hidden vector, which depends on the previous hidden vector [44]. LSTM is one of the popular changes to the RNN [45]. This network is designed to solve the problem of gradient vanish problem and RNN instability. Unlike the RNN, which only calculates a balanced sum of input signals and then passes an activation function, each LSTM unit uses memory C_t at time t. The output of the h_t or the activation of the LSTM unit is $h_t = \Gamma_o.\tanh(C_t)$, where Γ_o is the output gate and it controls the amount of content delivered through memory. The output gate can be calculated by Equation (4).

$$\Gamma_o = \sigma(W_o.[h_{t-1}, X_t + b_0) \quad (4)$$

This equation σ is a sigmoid activation function. W_O is also a skew matrix. The memory cell C_t is also updated to Equation (5).

$$C_t = \Gamma_f.C_{t-1} + \Gamma_u.\hat{C}_t \quad (5)$$

Where $\hat{C}_t$ is new memory content and is obtained by the form of Equation (6).

$$\hat{C}_t = \tanh(W_C.[h_{t-1}, X_t] + b_C) \quad (6)$$

The amount of current memory to forget is controlled by Γ_f and that amount of new memory content, which needs to be added to the memory cell, is done by the Γ_u update gateway. This is done by calculating Equation (7) and (8) [46, 47].

$$\Gamma_f = \sigma(W_f.[h_{t-1}, X_t] + b_f) \quad (7)$$

$$\Gamma_f = \sigma(W_u.[h_{t-1}, X_t] + bu) \quad (8)$$

In this study, we used a fusion of CNN and LSTM networks to classifying 2-stage and 3-stage of emotion. We will see that fusion of these two networks will increase the accuracy and reduce the oscillation.

III. PROPOSED METHOD

Details of the proposed emotions recognition method based on CNN-LSTM are provided in this section. Fig. 5. shows the general structure of the proposed method.

A. PREPROCESSING AND DISCUSSION:

First, a Notch filter is applied to the data for removing the 50 Hz frequency of the power supply, second, a first-order low-pass Butterworth filter with a frequency of 0.5 to 45 Hz is applied to the data. Third, the data were normalized between 0 and 1. Considering that one of the goals of this study was to provide an algorithm based on the minimum number of EEG signal channels, it was necessary to identify the active channels; For this purpose, in the fourth step, according to [33], [34] and [35] only Pz, T3, C3, C4,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011882, IEEE Access

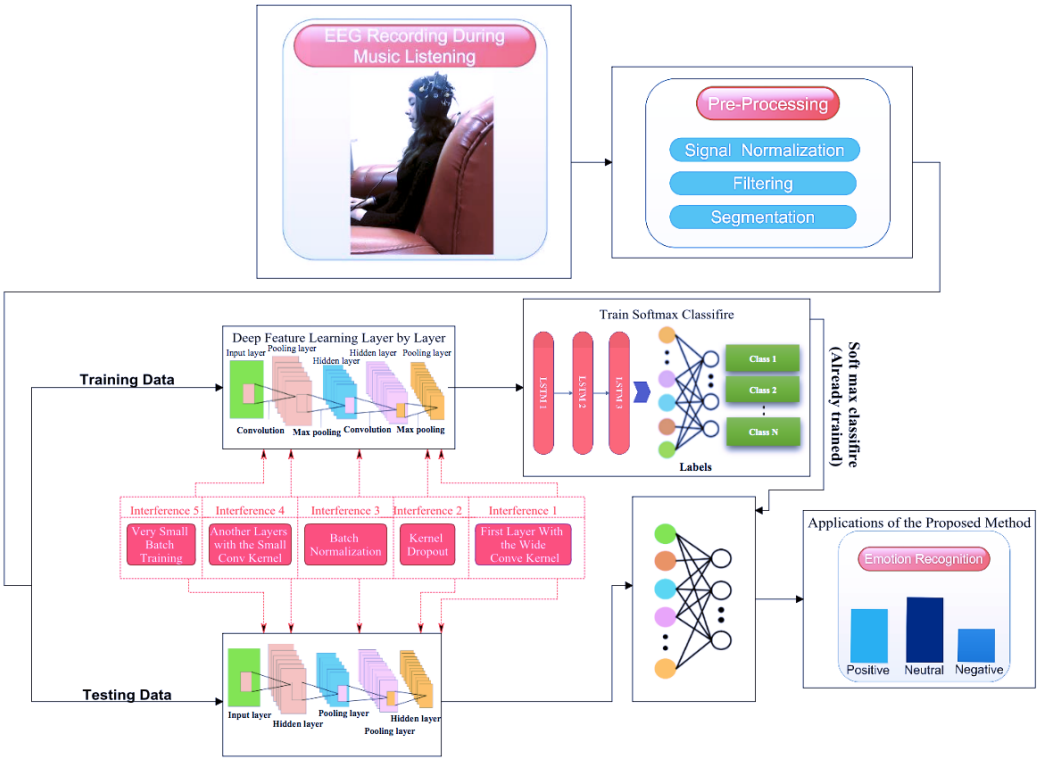

FIGURE 5. **The block diagram of the proposed method.**

T4, F7, F3, Fz, F4, F8, Fp1, and Fp2 electrodes are used for simulation and data processing. Fig. 6. shows the electrodes selected for simulation and data processing.

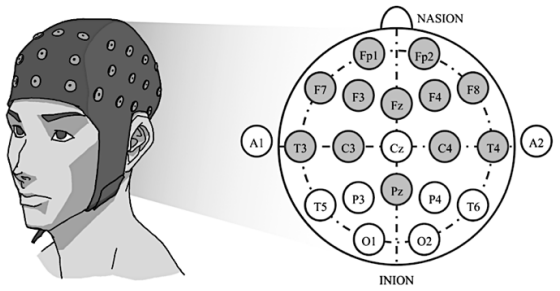

FIGURE 6. **The electrodes selected for simulation and processing of data.**

As can be seen from Fig. 3, the number of data related to the neutral class is less than the data of the positive and negative classes, which causes an imbalance between the data and may cause an overfitting problem. Furthermore, the lack of balance between the data of each class is a challenging situation, resulting in a bias in classification outcomes and degraded accuracy. The overlapping methods are used in the proposed approach to solve the challenges of unbalanced classes. In this process, all corresponding epochs of each emotion are concatenated to form a single long signal; then rectangular windows of specific duration and overlap are implemented in such a way that the number of epochs collected for each of the emotion classes is equivalent.

Fifth, in the proposed method for each channel, 5 minutes of recorded signal (according to Fig. 3.) is selected for each emotion. In this case, we have 2 data classes (negative and positive) with 75000 sampling points for each channel. Then, using the overlap technique to prevent overfitting, the data are split into 8-second intervals per channel. In fact, depending on the size of the shift, each electrode is divided into 2000 sampling points (8 seconds). For example, electrode e is the dimension of the input matrix (e×360×2000). Since we have 7 subjects and 2 classes (positive and negative), then the final dimension of the input matrix to the network will be equal to (2×7×360)×(e×2000). According to Fig. 3, we also considered the three-stage classification of emotion (positive, negative, neutral). This step (three-stage classification of emotion) has also been considered as before; eventually, the final dimension of the input matrix to the network will be equal to (3×7×360)×(e×2000). Fig. 7. shows this operation for the positive (5×60s=300s), negative (5×60s=300s) and neutral (8×15s=120s) classes of emotion. Due to the fact that the number of data related to the neutral class is less than the positive and negative classes, the overlap between the neutral class will be higher, the amount of shift in the neutral class is less than the positive and negative classes, so the overlap between the neutral class will be higher.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011882, IEEE Access

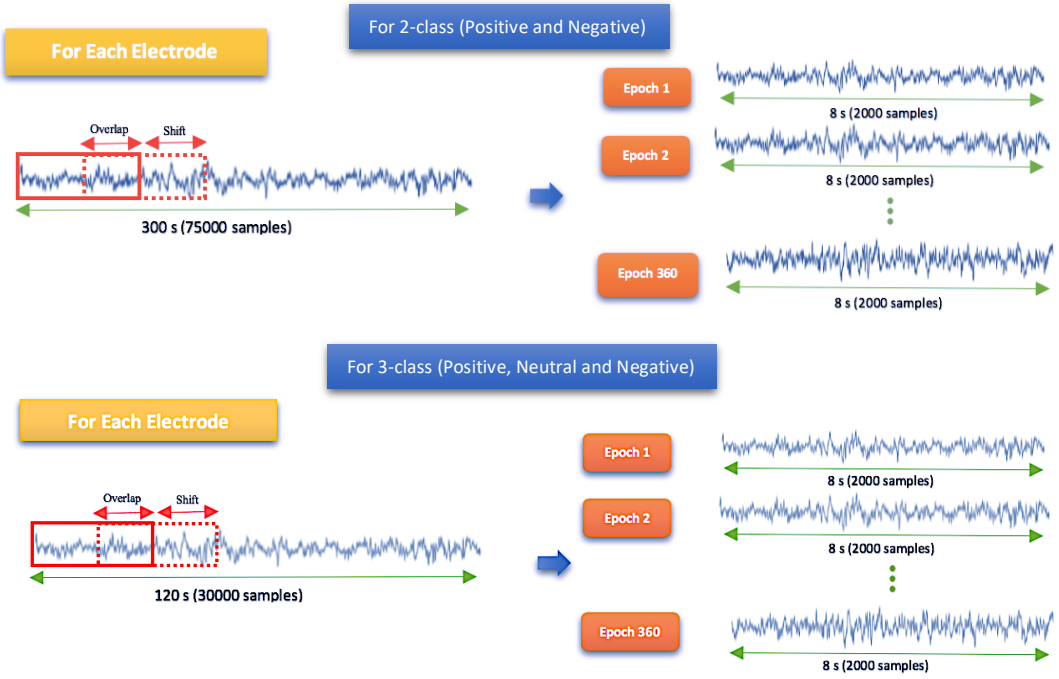

FIGURE 7. **Shows the overlap operation for 2-stage and 3-stage classification of emotion.**

B. PROPOSED NETWORK ARCHITECTURE:

In the proposed network architecture, we used a fusion of 10 convolution1-D layers and 3 LSTM layers. A cross-library in Python programming language is used to implement the proposed CNN-LSTM network. The CNN architecture has been also selected as follows: I. A dropout layer. II. A convolutional layer with nonlinear Leaky-Relu function, then a max-pooling layer followed by a batch normalization layer. III. The previous step architecture is repeated 9 times. IV. The output of the previous architecture is connected to a 2D matrix. V. The previous architecture output connects to the 3 layers of LSTM with Leaky-Relu nonlinear functions in series, then these layers followed by a batch normalization layer. VI. Two fully connected layers are used to access the output layer. Table 3 shows the details of the proposed deep neural network architecture. As it is shown in Table 3, the dimensionality reduction in the hidden layers continued from 24000 (12×2000) (the number of initial time features) to 100. Finally, the selected feature vector was linked to the fully connected layer with the nonlinear Softmax function. Fig. 8. shows the architecture of the suggested network.

C. PROPOSED DNN MODEL TRAINING AND ASSESSMENT:

All hyper-parameters for the proposed CNN-LSTM network are specifically adjusted to achieve the best rate of convergence and the procedure for trial and errors is followed to determine these hyper-parameters. Finally, the training process is performed by cross-entropy cost function and RmStrop optimizer [46] with a learning rate of 0.001 and batch size 10. The total number of parameters for two-class and three-class is equal to 167822 and 167923 respectively. Besides, the total number of samples for two-class and three-class of emotion is 5040 and 7560 respectively, of these, 60% are randomly selected for training the network (3024 for two-class and 4536 for three-class) and the remaining 40% (2016 for two-class and 3024 for three-class) are selected as the test set. 10% of the data are also used to validate the training set. After training the deep neural network, the proposed network model is evaluated using 40% of the total data. According to what has been explained before, Fig. 9. shows the EEG signal allocation data for the training and test set.

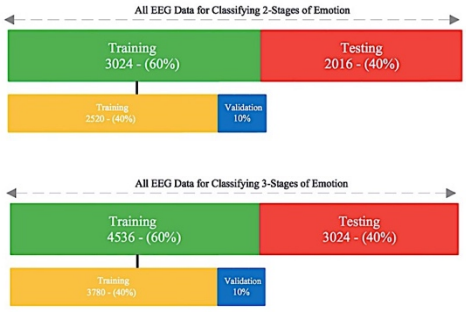

FIGURE 9. **EEG data allocation in the proposed algorithm for classifying 2-stage and 3-stage of emotions.**

IV. Simulation Result

In this section, simulation results of the proposed algorithm was presented for automatic emotion recognition. A laptop with 8 GB of RAM and 2.4 GHz Core i7 CPU were used to simulate the proposed algorithm. Fig. 10. shows the loss function of the proposed network for two-stage classification (negative

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011882, IEEE Access

TABLE III
SIZE OF FILTERS AND STEPS RECOMMENDED FOR SUGGESTED NETWORK.

L	Layer type	Activation function	Output Shape	Size of Kernel and Pooling	Strides	Number of filters	padding
0-1	Convolution1-D	Leaky ReLU	(None, 4000, 16)	120×1	6×1	16	**yes**
1-2	Max-Pooling1-D	-	(None, 2000, 16)	2×1	2×1	-	**no**
2-3	Convolution1-D	Leaky ReLU	(None, 2000, 32)	3×1	1×1	32	**yes**
3-4	Max-Pooling1-D	-	(None, 1000, 32)	2×1	2×1	-	**no**
4-5	Convolution1-D	Leaky ReLU	(None, 1000, 64)	3×1	1×1	64	**yes**
5-6	Max-Pooling1-D	-	(None, 500, 64)	2×1	2×1	-	**no**
6-7	Convolution1-D	Leaky ReLU	(None, 500, 80)	3×1	1×1	80	**yes**
7-8	Max-Pooling1-D	-	(None, 250, 80)	2×1	2×1	-	**no**
8-9	Convolution1-D	Leaky ReLU	(None, 250, 80)	3×1	1×1	80	**yes**
9-10	Max-Pooling1-D	-	(None, 125, 80)	2×1	2×1	-	**no**
10-11	Convolution1-D	Leaky ReLU	(None, 125, 80)	3×1	1×1	80	**yes**
11-12	Max-Pooling1-D	-	(None, 62, 80)	2×1	2×1	-	**no**
12-13	Convolution1-D	Leaky ReLU	(None, 62, 80)	3×1	1×1	80	**yes**
13-14	Max-Pooling1-D	-	(None, 31, 80)	2×1	2×1	-	**no**
14-15	Convolution1-D	Leaky ReLU	(None, 31, 80)	3×1	1×1	80	**yes**
15-16	Max-Pooling1-D	-	(None, 15, 80)	2×1	2×1	-	**no**
16-17	Convolution1-D	Leaky ReLU	(None, 15, 80)	3×1	1×1	80	**yes**
17-18	Max-Pooling1-D	-	(None, 7, 80)	2×1	2×1	-	**no**
18-19	Convolution1-D	Leaky ReLU	(None, 7, 80)	3×1	1×1	80	**yes**
19-20	Max-Pooling1-D	-	(None, 3, 80)	2×1	2×1	-	**no**
21-22	LSTM	Leaky ReLU	(None, 128)	-	-	-	
22-23	LSTM	Leaky ReLU	(None, 128)	-	-	-	
23-24	LSTM	Leaky ReLU	(None, 128)	-	-	-	
24-25	Fully-connected	Leaky ReLU	(None, 100)	-	-	-	
25-26	Fully-connected	Softmax	(None, 2-3)	-	-	-	

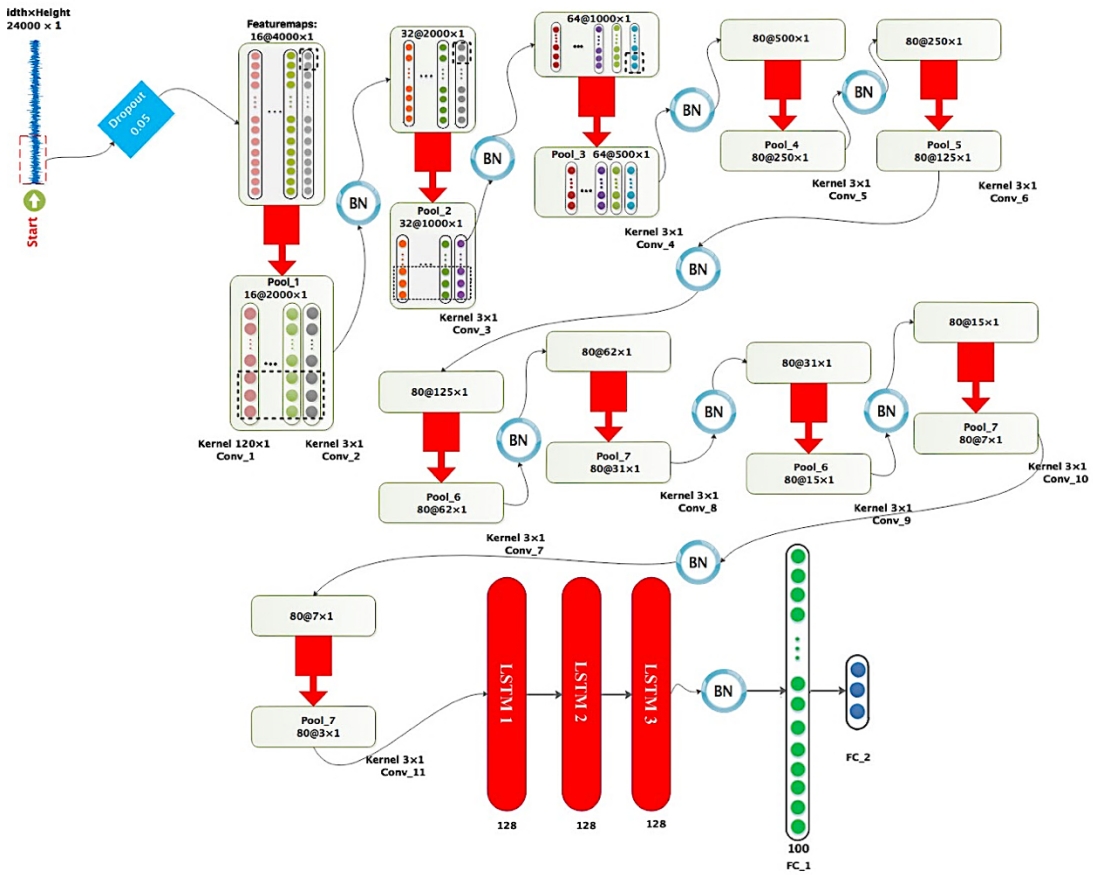

FIGURE 8. **Proposed CNN-LSTM Network Architecture.**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011882, IEEE Access

and positive) and three-stage classification (negative, neutral and positive) of emotion. As shown in Fig. 10. (a), the network error for two-stage classification of emotion decreases by increasing the number of iteration and reach its steady-state value from about 130th iteration. as well as Fig. 10. (b) shows the steady-state value for the three-stage classification of emotion from about 145th iteration. Fig. 11. shows the accuracy of the proposed method for two-stage classification (negative and positive) and three-stage classification (negative, neutral and positive) of emotion in 400 iterations for validation of the data. As shown in Fig. 11, the accuracy of the proposed method for two-stage classification (negative and positive) and three-stage classification (negative, neutral and positive) of emotion reaches 97.42% and 96.78%, respectively, at about 200 iterations. To further analysis the suggested method, the confusion matrix for two-stage and three-stage classification has been given in Fig. 12. The accuracy obtained from the proposed method is also promising. Also, Fig. 13. shows the Bar-chart diagram of sensitivity, specificity, and accuracy of two-stage and three-stage classification. Furthermore, the precision, sensitivity and specificity values of the two-stage and three-stage classification are very promising in this figure. Table 4 shows the F-measure obtained for a 2-stage (positive and negative),and 3-stage (positive, negative, and neutral) classification of emotion. According to Table 4, the F-Measure for the 2-and 3-class stages is 97.42% and 95.24%, respectively. Fig. 14. also shows the t-SEN chart for the raw signal, Conv5, and Softmax layers for two-stage classification and three-stage classification of emotion. As it is clear from the last layer, almost all the samples are separated for the evaluation set, indicating the optimal efficiency of the suggested method for two-stage classification and three-stage classification.

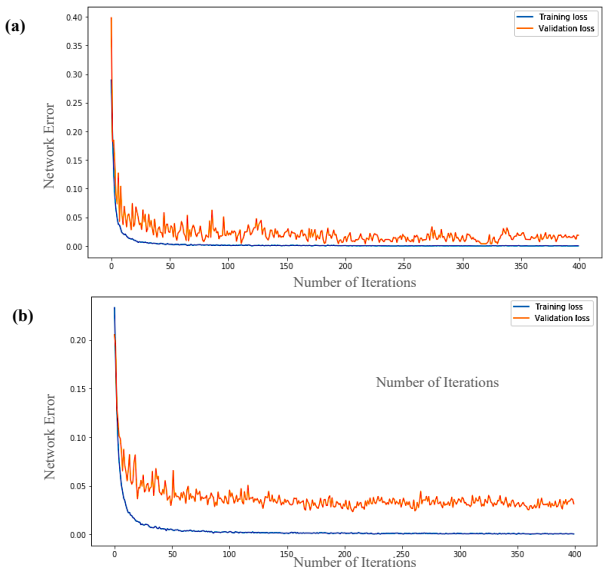

FIGURE 10. The proposed network error for (a). 2-stage and (b). 3-stage classification of emotion.

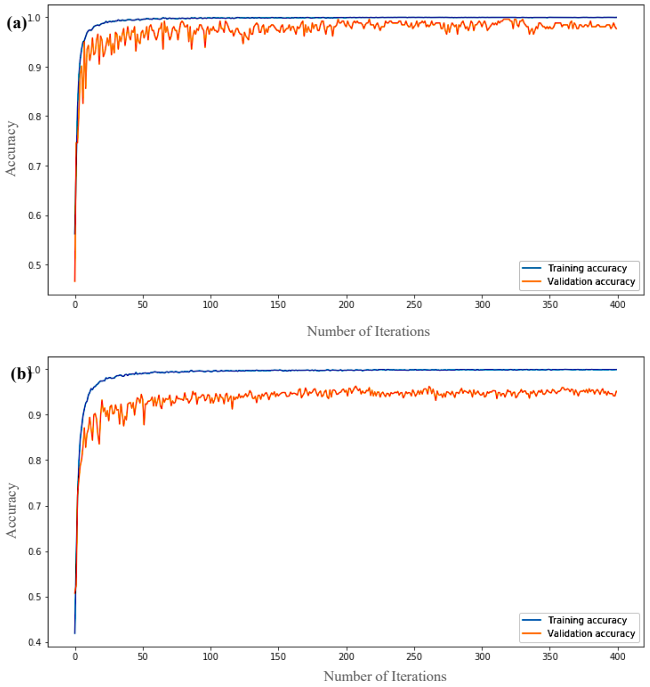

FIGURE 11. The proposed network accuracy for (a). 2-stage and (b). 3-stage classification of emotion.

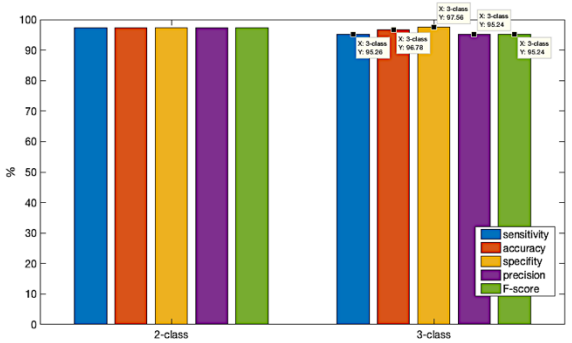

FIGURE 13. The bar-chart diagram for 2-stage and 3-stage classification of emotion.

In order to show the performance of the proposed LSTM-CNN method with different data types as input, the accuracy of the classification is obtained using the other common methods for 3-stage emotion recognition. In this regard, time data and several manual features from time data, along with DBM and MLP, are selected as comparative methods [43, 47, 48]. The number of hidden layers is considered 3 for DBM and MLP, and the learning rate is chosen as 0.001. Also, for CNN, the proposed architecture in Table 3 was selected regardless of the LSTM layers. The parameters of the minimum, maximum, skewness, crest factor, variance, root mean square (RMS), mean, and kurtosis are chosen as the hand crafted features of the time domain (time features). The classification accuracy of the different methods based on the feature learning from raw data and the manual features are presented in Fig. 15. The reliability of CNN, DBM and MLP reaches 90%, 79% and 73%, respectively after 180 iterations. As can be seen from Fig. 15, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011882, IEEE Access

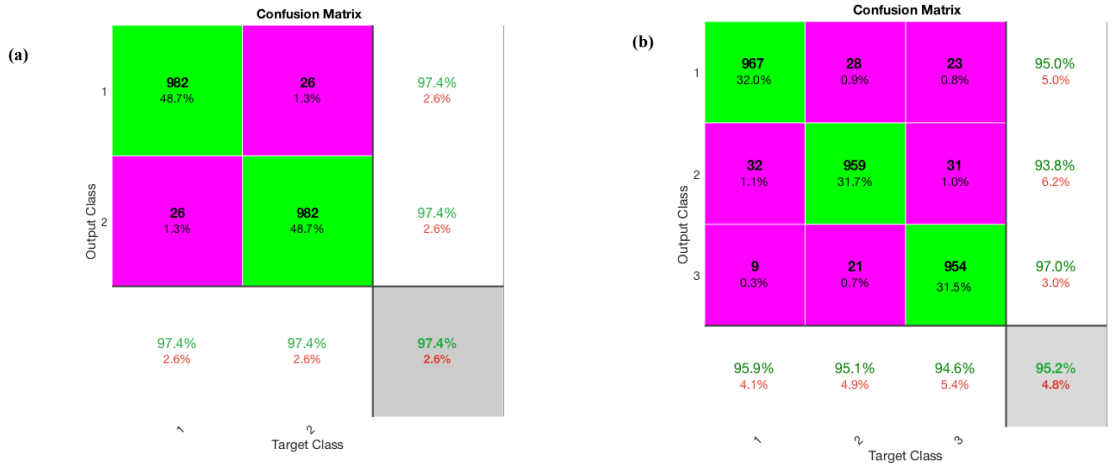

FIGURE 12. Shows the confusion matrix for (a). 2-stage and (b). 3-stage classification of emotion.

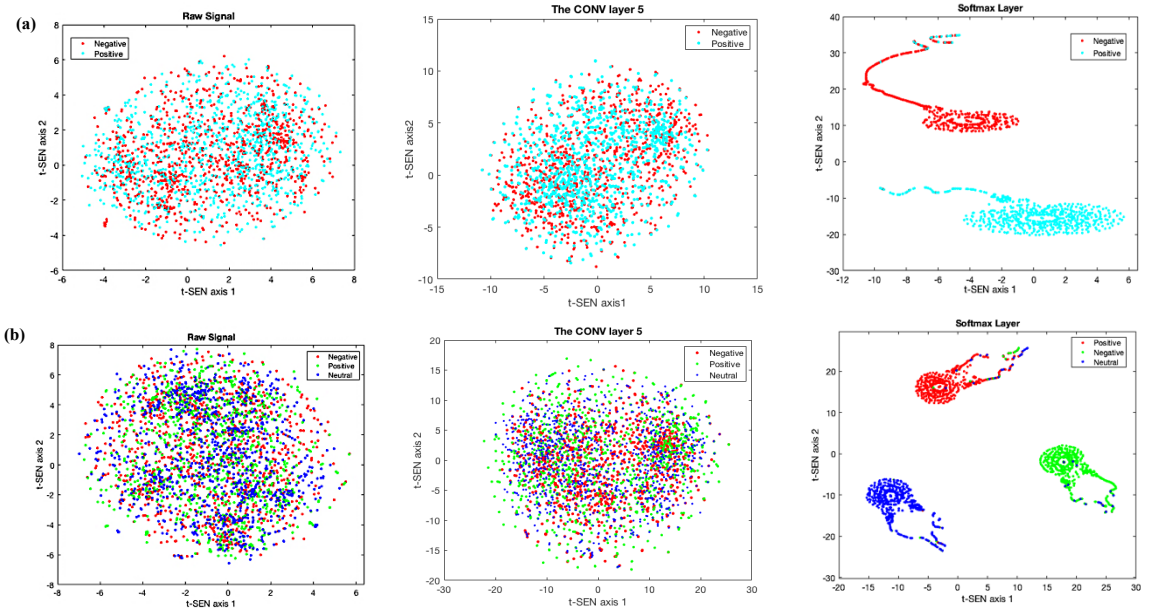

FIGURE 14. The t-SEN chart of the proposed method for (a). 2-stage and (b). 3-stage classification of emotion.

FIGURE 15. The performance of the suggested method compared to the CNN, MLP and DBM networks for classifying 3-stage of emotion.

TABLE V
COMPARISON OF THE SUGGESTED NETWORK'S COMPUTATIONAL COMPLEXITY WITH CNN, MLP AND DBM FOR CLASSIFYING 3-STAGE OF EMOTION IN 180 ITERATIONS.

	P-M (LSTM-CNN)		CNN		DBM		MLP	
Class	Train	Test	Train	Test	Train	Test	Train	Test
2-Stage	5400 s	5 s	5002 s	3 s	3011 s	4.5 s	909 s	2.5 s
3-Stage	12600 s	6 s	11200 s	3.5 s	6009 s	4.5 s	1201 s	2 s

performance of the proposed network is promising compared to CNN, DBM and MLP and the proposed algorithm converge to the desired value faster. Also, the computational complexity of running time for training and test phases is given for the proposed network (for 3-stags) as well as CNN, DBM and MLP networks in Table 5.

TABLE IV
F-MEASURE OBTAINED FOR A 2-STAGE (POSITIVE AND NEGATIVE) AND 3-STAGE (POSITIVE, NEGATIVE AND NEUTRAL) CLASSIFICATION OF EMOTION.

2-class	**Positive**	**Negative**	**Average**
F-measure	97.42	97.42	**97.42**

3-class	**Positive**	**Negative**	**Neutral**	**Average**
F-measure	95.46	94.48	95.78	**95.24**

As can be seen from Table 5, the running time of the proposed network is approximately comparable to that of DBM and CNN, but, the MLP has overtaken all three of the proposed network, CNN and DBM methods. Nevertheless, this is at the cost of reduced accuracy of classification.

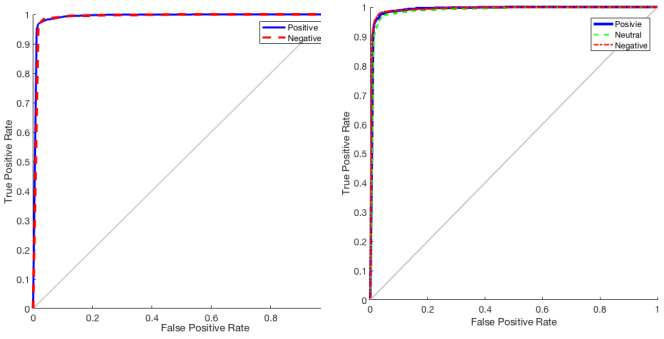

FIGURE 16. The ROC diagram for 2-stage and 3-stage classification of emotion.

To further research the efficiency of the proposed method, the ROC diagram is provided in Fig. 16. Common approaches such as wavelet transformation, empirical mode decomposition, etc. have been used in most previous works to extract the essential features of the signal, involving some common problems regarding the parameters of the extraction feature such as selecting the mother wavelet type, the number of decomposition levels, etc. One of the most important advantages of the proposed method compared to the other methods is that the extraction of a feature is done automatically on the basis of deep learning and no feature selection procedure is needed. Table 6. Shows the kappa coefficient of the proposed method for classifying 2-stage and 3-stage of emotion.

TABLE VI
SHOWS THE KAPPA COEFFICIENT OF THE PROPOSED METHOD FOR AUTOMATIC 2-STAGE AND 3-STAGE CLASSIFICATION OF EMOTION.

Class	**2-stage**	**3-stage**
Kappa	0.9484	0.9306

In order to evaluate the performance of the proposed, CNN, DBM and MLP method against observation noise, the white Gaussian noise of SNR -4 to 20 dB is added as the measurement noise to the EEG signals and the classification accuracy for all methods is reported in Fig. 17. As it can be seen, the classification performance of the proposed method is considerably robust to the measurement noise in a wide range of SNR, so that the accuracy is still more than 90% for SNR -4 to 20 dB.

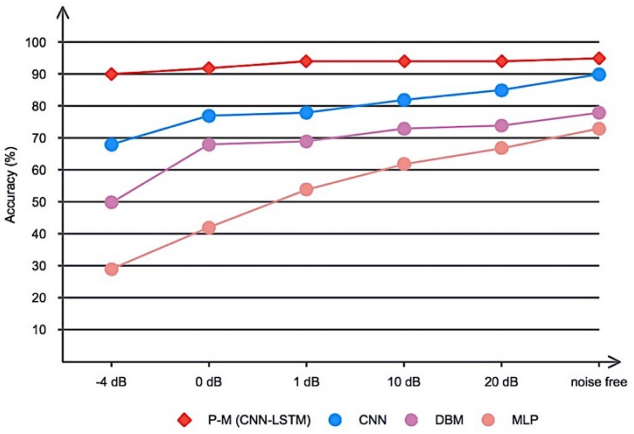

FIGURE 17. Accuracy of the proposed network versus SNR in additive white Gaussian noise for emotions recognition.

V. CONCLUSION

In this work, a new method for emotions recognition is presented using a fusion of the CNN and LSTM networks. The proposed network consisted of the 10-CNN and 3-LSTM layers. As it was observed, the fusion of these networks increases the accuracy and stability of the proposed algorithm. Also, we achieved 97.42% and 95.23% accuracy for 2-stage and 3-stage of emotion for 12 active channels, also the Kappa Cohen's coefficients for 2-stage and 3-stage of emotion are 0.96 and 0.93, respectively, which is very promising compared to the previous emotions recognition approaches, we also compared our proposed LSTM-CNN network (end-to-end) with other hand-crafted methods based on MLP and

DBM classifiers and achieved promising results compared to similar methods, as well as, it is shown that the proposed network is robust to the measurement noise of level as much as 1 dB. Despite the contributions, this work has some limitations, as with other previous studies. First, proposed network parameters, such as the learning rate and training algorithm parameters were selected mostly based on the experience or the trial-and-error method. Therefore, it would be better to develop a more systematic method for selecting the appropriate parameters. Second, the training time of the proposed algorithm is relatively high, which can be solved using a graphical processing unit (GPU) systems. In addition, deep learning is more dependent on powerful computation compared to traditional machine learning methods, and will spend much more time on general training. In order to save computational complexity; in future research, we could consider introducing some pre-training models or combining transfer learning methods to accelerate model training. We also intend to use a number of more emotional states for classification purposes in future studies. The proposed method is also expected to be used in BCI applications.

REFERENCES

[1] R. Nicole, "Title of paper with only first word capitalized, J," Name Stand. Abbrev, 1987.
[2] T. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE translation journal on magnetics in Japan, vol. 2, no. 8, pp. 740-741, 1987.
[3] J. A. Russell, "A circumplex model of affect," Journal of personality and social psychology, vol. 39, no. 6, p. 1161, 1980.
[4] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," IEEE Transactions on information Technology in Biomedicine, vol. 14, no. 2, pp. 186-197, 2009.
[5] A. M. Bhatti, M. Majid, S. M. Anwar, and B. Khan, "Human emotion recognition and analysis in response to audio music using brain signals," Computers in Human Behavior, vol. 65, pp. 267-275, 2016.
[6] Y.-P. Lin et al., "EEG-based emotion recognition in music listening," IEEE Transactions on Biomedical Engineering, vol. 57, no. 7, pp. 1798-1806, 2010.
[7] G. Balasubramanian, A. Kanagasabai, J. Mohan, and N. G. Seshadri, "Music induced emotion using wavelet packet decomposition—An EEG study," Biomedical Signal Processing and Control, vol. 42, pp. 115-128, 2018.
[8] Y. Ding, X. Hu, Z. Xia, Y.-J. Liu, and D. Zhang, "Inter-brain EEG feature extraction and analysis for continuous implicit emotion tagging during video watching," IEEE Transactions on Affective Computing, 2018.
[9] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 60-75, 2017.
[10] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," IEEE transactions on affective computing, vol. 3, no. 1, pp. 42-55, 2011.
[11] S. Koelsch, Brain and music. John Wiley & Sons, 2012.
[12] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," Psychology of Music, vol. 39, no. 1, pp. 18-49, 2011.
[13] D. Sammler, M. Grigutsch, T. Fritz, and S. Koelsch, "Music and emotion: electrophysiological correlates of the processing of pleasant and unpleasant music," Psychophysiology, vol. 44, no. 2, pp. 293-304, 2007.
[14] L.-O. Lundqvist, F. Carlsson, P. Hilmersson, and P. N. Juslin, "Emotional responses to music: Experience, expression, and physiology," Psychology of music, vol. 37, no. 1, pp. 61-90, 2009.
[15] A. J. Blood and R. J. Zatorre, "Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion," Proceedings of the National Academy of Sciences, vol. 98, no. 20, pp. 11818-11823, 2001.
[16] K. Mueller et al., "Investigating the dynamics of the brain response to music: A central role of the ventral striatum/nucleus accumbens," NeuroImage, vol. 116, pp. 68-79, 2015.
[17] S. Moghimi, A. Kushki, S. Power, A. M. Guerguerian, and T. Chau, "Automatic detection of a prefrontal cortical response to emotionally rated music using multi-channel near-infrared spectroscopy," Journal of neural engineering, vol. 9, no. 2, p. 026022, 2012.
[18] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," IEEE Transactions on Affective Computing, 2017.
[19] M. Balconi, E. Grippa, and M. E. Vanutelli, "What hemodynamic (fNIRS), electrophysiological (EEG) and autonomic integrated measures can tell us about emotional processing," Brain and cognition, vol. 95, pp.67-76, 2015.
[20] W. Zheng, "Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis," IEEE Transactions on Cognitive and Developmental Systems, vol. 9, no. 3, pp. 281-290, 2016.
[21] P. Ozel, A. Akan, and B. Yilmaz, "Synchrosqueezing transform based feature extraction from EEG signals for emotional state prediction," Biomedical Signal Processing and Control, vol. 52, pp. 152-161, 2019.
[22] F. Hasanzadeh and S. Moghimi, "Emotion estimation during listening to music by EEG signal and applying NARX model and genetic algorithm," presented at the National Conference of Technology, Energy & Data on Electrical & Computer Engineering, 2015.
[23] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," IEEE Transactions on Affective Computing, vol. 7, no. 1, pp. 17-28, 2015.
[24] S. Koelstra et al., "Deap: A database for emotion analysis; using physiological signals," IEEE transactions on affective computing, vol. 3, no. 1, pp. 18-31, 2011.
[25] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals," in International workshop on multimedia content representation, classification and security, 2006: Springer, pp. 530-537.
[26] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," IEEE Transactions on Affective Computing, 2017.
[27] N. Thammasan, K. Moriyama, K.-i. Fukui, and M. Numao, "Familiarity effects in EEG-based emotion recognition," Brain informatics, vol. 4, no. 1, pp. 39-50, 2017.
[28] Y. Kumagai, M. Arvaneh, and T. Tanaka, "Familiarity affects entrainment of EEG in music listening," Frontiers in human neuroscience, vol. 11, p. 384, 2017.
[29] G. Zhao, Y. Zhang, and Y. Ge, "Frontal EEG asymmetry and middle line power difference in discrete emotions," Frontiers in behavioral neuroscience, vol. 12, p. 225, 2018.
[30] J. Lu, D. Wu, H. Yang, C. Luo, C. Li, and D. Yao, "Scale-free brain-wave music from simultaneously EEG and fMRI recordings," PloS one, vol. 7, no. 11, 2012.
[31] H. J. Yoon and S. Y. Chung, "EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm," Computers in biology and medicine, vol. 43, no. 12, pp. 2230-2237, 2013.
[32] X. Li, B. Hu, S. Sun, and H. Cai, "EEG-based mild depressive detection using feature selection methods and classifiers,"

Computer methods and programs in biomedicine, vol. 136, pp. 151-161, 2016.

[33] F. Hasanzadeh, M. Annabestani, and S. Moghimi, "Continuous Emotion Recognition during Music Listening Using EEG Signals: A Fuzzy Parallel Cascades Model," arXiv preprint arXiv:1910.10489, 2019.

[34] Y. Hou and S. Chen, "Distinguishing Different Emotions Evoked by Music via Electroencephalographic Signals," Computational intelligence and neuroscience, vol. 2019, 2019.

[35] P. Keelawat, N. Thammasan, M. Numao, and B. Kijsirikul, "Spatiotemporal Emotion Recognition using Deep CNN Based on EEG during Music Listening," arXiv preprint arXiv:1910.09719, 2019.

[36] Y. Yang, Q. Wu, M. Qiu, Y. Wang and X. Chen, "Emotion Recognition from Multi-Channel EEG through Parallel Convolutional Recurrent Neural Network," 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, 2018, pp. 1-7.

[37] Yang, H., Han, J., & Min, K. (2019). A Multi-Column CNN Model for Emotion Recognition from EEG Signals. Sensors (Basel, Switzerland), 19(21), 4736.

[38] Chen, J., Jiang, D., Zhang, Y., & Zhang, P. (2020). Emotion recognition from spatiotemporal EEG representations with hybrid convolutional recurrent neural networks via wearable multi-channel headset. Computer Communications, 154, 58-65.

[39] Wei, C., Chen, L. L., Song, Z. Z., Lou, X. G., & Li, D. D. (2020). EEG-based emotion recognition using simple recurrent units network and ensemble learning. Biomedical Signal Processing and Control, 58, 101756.

[40] Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. http://www. deeplearningbook.org.

[41] Hung, S.L., Adeli, H., 1993. Parallel backpropagation learning algorithms on cray Y-MP8/ 864 supercomputers. Neurocomputing 5 (6), 287–302.

[42] Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R., 2012. IMproving neural networks by preventing co-adaptation of feature detectors. Arxiv Prepr. Arxiv. 1207.0580.

[43] Mousavi, Z., et al. "Deep convolutional neural network for classification of sleep stages from single-channel EEG signals." Journal of neuroscience methods (2019): 108312.

[44] Graves, "Generating sequences with recurrent neural networks," arXiv preprint arXiv:1308.0850, 2013.

[45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[46] Wichrowska, Olga, Niru Maheswaranathan, Matthew W. Hoffman, Sergio Gomez Colmenarejo, Misha Denil, Nando de Freitas, and Jascha Sohl-Dickstein. "Learned optimizers that scale and generalize." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3751-3760. JMLR. org, 2017.

[47] Salakhutdinov, R., Hinton, G., 2009. Deep boltzmann machines. Artificial Intelligence and Statistics. pp. 448–455.

[48] Hsu, Y.L., Yang, Y.T., Wang, J.S., Hsu, C.Y., 2013. Automatic sleep stage recurrent neural classifier using energy features of eeg signals. Neurocomputing 104 (0), 105–114.