February 2020

Classifying Emotions with EEG and Peripheral Physiological Data Using 1D Convolutional Long Short-Term Memory Neural Network

Rupal Agarwal
University of South Florida

Classifying Emotions with EEG and Peripheral Physiological Data Using 1D Convolutional Long Short-Term Memory Neural Network

by

Rupal Agarwal

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Major Professor: Marvin Andujar, Ph.D.
Shaun Canavan, Ph.D.
Paul A. Rosen, Ph.D.

Date of Approval:
March 11, 2020

Keywords: EEG, DEAP, Affective Computing, Deep Neural Networks

Dedication

To my family and friends who have supported me throughout my life.

Acknowledgments

I would like to express my heartfelt gratitude to my advisor Dr. Marvin Andujar for his continuous support and guidance throughout my thesis project. It was a pleasure doing my thesis under his supervision. His advice and constructive feedback have helped me learn and grow. And, I would also like to extend my sincere thanks to Dr. Shaun Canavan and Dr. Paul A. Rosen for agreeing to be a part of my defense committee and taking me in the right direction with all their valuable inputs.

Table of Contents

List of Tables

List of Figures

Abstract

Recognizing emotions is very important while building robust and interactive Affective Brain-Computer Interfaces as it allows the machines to have some degree of emotional intelligence with the help of which they can understand the changing emotional state of users. In the past, emotions have been recognized via unimodal data such as electroencephalography (EEG) signals, speech, facial expressions or peripheral physiological signals. However, emotions are complex as they are a combination of human behavior, thinking and feeling. Therefore, as compared to unimodal methods, multi-modal techniques, recognize emotions with more reliability. This thesis aims to recognize and classify human emotions into high/low arousal and high/low valence using a multi-modal approach. The different modalities used are EEG, blood pressure, respiration, skin temperature, eye movements, muscle movements and skin conductance. The data is taken from a publicly available dataset called DEAP. The experiments are performed using the 1D Convolutional LSTM network and its performance is then compared with three baseline Machine Learning (ML) algorithms – Support Vector Machine, K-Nearest Neighbor and Random Forest. To investigate further, the emotion classification performance of different regions of the brain such as frontal lobe, parietal lobe, temporal lobe, occipital lobe, left hemisphere and right hemisphere are also compared. The model achieved an average accuracy of 91.19% for valence and 91.51% for arousal when used with a combination of EEG and peripheral data. The overall results show that the proposed neural network outperforms the traditional ML algorithms and gives a high emotion classification accuracy.

Chapter 1: Introduction

1.1 Motivation and Problem Statement

Emotion is a complex psychological subjective experience that constitutes a physiological and behavioral response to internal and external stimuli [1]. It plays an important role in decision making and human interpersonal communication. Apart from its importance in human to human communication, the study of emotions in Human-Computer Interaction has also gained a lot of attention in the last few years. Past research has shown that humans tend to pass off their interpersonal behavioral patterns onto their computers [2]. Therefore, to facilitate a more natural interaction, computers must observe the changing emotional states of users. For example, if a speech recognition interface is considered, it should not only focus on what is said but also how it is said [3]. Understanding this essential information will be very beneficial in building effective communication between humans and machines.

Emotions can be expressed via both verbal and non-verbal means. Non-verbal means include facial expressions, body gestures, body postures and neurophysiological signals such as Electroencephalogram (EEG) signals, respiration, heart rate and skin conductance. In the past, a lot of research has been done in the field of emotion recognition, using data such as facial images [67], body gestures [25] and speech [13], but these modalities are not reliable as they can be easily faked. For example, people may smile even if they are angry or disgusted [4]. Other challenges with these modalities include auditory noise, poor lighting conditions in facial images and the use of facial accessories like glasses which occlude the face. All these problems hamper the accurate estimation of emotions.

Considering this, recently, a lot of attention is being directed towards using physiological data for emotion recognition. It is because they are more accurate, reliable and are beyond the voluntary control of people [5]. These signals are involuntarily produced in response to the Autonomous Nervous System (ANS) of the body and therefore, they cannot be controlled. Although, the physiological signals are popularly used for emotion recognition, there are some problems associated with them. They are complex because it is difficult to comprehend their nature. They cannot be visually perceived and are often contaminated with noise from the recording devices and artefacts of the human body. All these factors make it relatively difficult to annotate the raw physiological data with specific emotions and get the ground truth information from it. However, their advantages such as accuracy, reliability, spontaneity and a strong relationship with the underlying emotions cannot be ignored, and hence, they are often used in studies dealing with emotion recognition.

In this thesis, a hybrid of 1D convolution [6] and recurrent neural network (LSTM) [7] was developed to classify emotions. The experiments were conducted utilizing the online DEAP [8] dataset. This dataset had 32 EEG channels and eight peripheral physiological channels containing the data of six types of signals such as eye movements, muscle movements, blood pressure, respiration, skin conductance and temperature. The network was trained on these 40 channels with different combinations. To gain in-depth information on how different regions of the brain can classify emotions, the network was trained on data belonging to different brain lobes and hemispheres. The input was preprocessed to remove noise and then it was passed through a 12-layer deep neural network. The performance of the deep neural model was also compared against three baseline machine learning algorithms – Support Vector Machine [9], K-Nearest Neighbor

[10] and Random Forest [11]. The results showed that the proposed neural model outperformed the traditional ML classifiers and gave a high emotion classification accuracy.

1.2 Contributions

The contributions of this thesis are described in detail below –

(i) Previous studies have used 1D CNN-LSTM neural architecture for recognizing emotions through auditory and visual modalities [12] or via speech modality[13], [14], but, to the best of our knowledge, this is the first study that has used a 1D CNN-LSTM neural model for recognizing emotions based on physiological signals of the human body and has shown the state of the art performance on DEAP data.

(ii) This study applied minimum preprocessing techniques on input data and was able to achieve better results than the current state of the art studies which have done rigorous preprocessing on the same data.

(iii) Most of the studies have reported their emotion classification accuracies based on the entire EEG dataset. However, it should be noted that emotional activity is dominant in some regions of the brain and it should not be generalized across the entire brain as different parts of the brain produce different quality of data. We believe this is the first study that has reported emotion accuracies for different parts of the brain (4 brain lobes and 4 different regions of two brain hemispheres) and has analyzed the differences in the results for the DEAP dataset.

Chapter 2: Related Work

In recent years, researchers have been using physiological data, especially EEG, to recognize emotions as it gives an objective, reliable and detailed information on the emotional status and brain activity of humans. In a research study, Koelstra et. al [8], introduced a dataset (DEAP) for emotion analysis using EEG and peripheral physiological signals. Bayes classifier was used for classification which gave 62% accuracy for arousal and 57.6% accuracy for valence. Using the DEAP [8] dataset, Alhagry et. al [15], developed a Long-Short Term Memory neural network model to classify EEG data into high/low arousal, valence and liking and got an average accuracy of 85.65%, 85.45%, and 87.99%, respectively. In another research [16], a systematic comparison between different classifiers including KNN, SVM and Random Forest (RF) was done, and it was found out that KNN gave the best results with an average accuracy of 89.72% for high/low valence and high/low arousal.

Many studies have shown that multimodal signals give more detailed and comprehensive information on emotions as compared to unimodal signals [17]. A multimodal approach was adopted in [18], in which Bimodal Deep Autoencoder (BDAE) was used to recognize emotions by extracting shared representations from EEG signals and eye features. The two datasets used for this method were – SEED [19] and DEAP [8]. Experimental results showed that fusion of EEG and eye features improved the positive emotion recognition accuracy as compared to the scenario when only eye features were used. Similarly, negative emotion accuracy also improved after using both EEG and eye features as compared to when only EEG features were used. In another research, Said et. al [20] used Deep Autoencoder architecture to compress and classify EEG and EMG

signals. Results showed that the proposed approach performed better as compared to single modality approaches and gave an accuracy of 78.1%.

In another study [21], the arousal dimension of human emotions was classified into three classes, namely, calm, neutral and excited using EEG and peripheral physiological signals (galvanic skin response, thoracic movements, blood pressure and respiration). The data was acquired from 4 participants and it was classified using Naïve Bayes Algorithm and a classifier based on Fisher Discriminant Analysis (FDA). Results showed that the FDA was less sensitive to correlated features and hence, it outperformed Naïve Bayes.

Zheng et al. [22], built a fusion model after extracting different features including Power Spectral Density and Differential Entropy from EEG and eye-tracking data. Both, feature level fusion and decision level fusion techniques were used to fuse data. SVM was used for training which gave classification accuracies of 73.59% and 72.98%, respectively.

In [23], a multimodal fusion model was proposed to classify 13 emotions into three dimensions: arousal, valence and dominance. Daubechies Wavelet Transform [24] was used for analyzing the physiological signals and maximum classification accuracy of 85.46% was achieved using Support Vector Machine.

Recent advancements in emotion recognition have shown that deep learning methods perform better than traditional machine learning algorithms as they can handle large datasets more efficiently and can extract more robust and complex features from data [25]. Ranganathan et. al [25], proposed a Convolutional Deep Belief Networks to classify 23 different emotions using a database of multimodal recordings like facial expressions, body gestures, vocal expressions and physiological signals. In this study, 10 users participated, and their data was recorded six times: three times in standing position and three times in sitting position. Performance of Convolutional

Deep Belief Networks was compared against the SVM algorithm and it was found that the former gave better results while classifying emotions. In another study [26], a multiple-fusion-layer based ensemble classifier of stacked autoencoder (MESAE) was developed using physiological signals. MESAE gave better results as compared to state-of-the-art classifiers like KNN, SVM and Naïve Bayes due to its higher generalization capability.

Another study [27] used a transfer learning approach to classify emotions using the DEAP dataset. Time and frequency domain features from EEG signals and hand-crafted features from peripheral physiological signals were fed into a fine-tuned AlexNet [28] model. The accuracies obtained for valence and arousal were 85.5% and 87.30%, respectively.

In another research [29], EEG and EOG signals were used to detect fatigue during driving. Deep Autoencoder was used for fusing the features and it gave a high correlation coefficient (0.85) and low root mean square error (0.09).as compared to decision level and feature level fusion strategies. In a recent study, Fabiano and Canavan [72] developed a fusion technique and used a feed-forward neural network to classify emotions into two classes low/high arousal and low/high valence. They used the DEAP dataset for this purpose and achieved an average accuracy of 95.5% for the valence dimension and 95.27% for the arousal dimension.

Rayatdoost and Soleymani [30] reported that for the valence dimension, similar EEG activity patterns exist across different datasets and showed that deep convolutional networks improve classification performance as compared against traditional machine learning algorithms.

To summarize, the above-mentioned studies have used different types of physiological data, feature extraction strategies and classification algorithms for emotion recognition. To improve the classification accuracy for emotion recognition, in this thesis, a hybrid structure of 1-D Convolutional Neural Network and Recurrent Neural Network (LSTM) is proposed, which uses

EEG data and peripheral physiological data of DEAP dataset as input. The performance of the proposed neural network is also compared against state-of-the-art machine learning classifiers like KNN, SVM, and RF. Furthermore, the detailed classification performance of different brain lobes and different brain hemispheres is also highlighted. The rest of this thesis is organized as follows: Chapter 3 gives an overview of the theory of emotion, different parts of the human brain and physiological signals. Chapter 4 describes the DEAP dataset in detail. It is then followed by Chapter 5 which explains the adopted methodology. Chapter 6 gives the results and analysis and Chapter 7 concludes this study with some thoughts on future work.

Chapter 3: Background

This chapter gives a short, but, detailed description of the theory of emotion, brain lobes and brain hemisphere and different types of physiological signals.

3.1 Theory of Emotion

Emotion is a feeling or state of mind that occurs spontaneously and is accompanied by physiological changes in the human body. In general, emotions are reactions to thoughts, memories or events that occur in our surroundings.

Psychologists have defined several theories of emotion, however, the most common is the two-dimensional valence and arousal circumplex model [31]. This model proposes that emotions are distributed in two-dimensional space with valence representing the horizontal axis and arousal representing the vertical axis such that each emotion is a combination of varying degrees of these two dimensions. In this model, valence is expressed on a continuum from pleasure to displeasure (from positive to negative) and arousal is expressed as different degrees of excitement (from low to high excitement) [32]. Figure 3.1 represents the valence and arousal circumplex model of affect. As can be seen in this figure, if someone has high arousal and positive valence then they will be excited or alert. Similarly, if they are depressed then they will have low arousal and negative valence. Thus, different emotions can be represented in terms of valence and arousal.

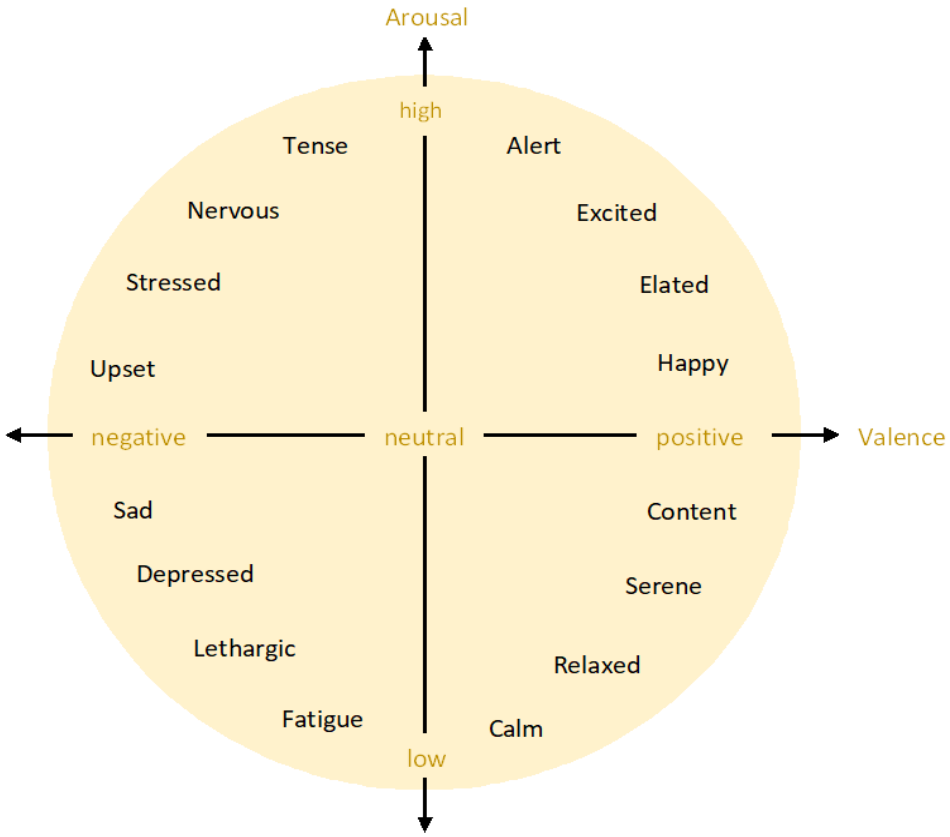

Figure 3.1 Valence and arousal circumplex model of affect.

3.2 Brain Lobes and Hemispheres

The dominant part of the brain is a highly convoluted surface layer called the cerebral cortex. It can be divided into two hemispheres – left and right. Cortex is further divided into four lobes – frontal lobe, temporal lobe, parietal lobe and occipital lobe. Each of these four lobes includes portions of the left and right hemispheres. The location of the four lobes in the brain is shown in Figure 3.2 and their functionality is described below –

(i) ***Frontal Lobe*** **–** It is responsible for conscious thought, intelligence, emotions and voluntary movement of limbs [33].

(ii) ***Temporal Lobe*** **–** It is related to long-term memory, visual memory, comprehension of speech and sense of smell and sound.

(iii) ***Parietal Lobe –*** It is responsible for merging the information received from various sense organs. It also processes and analyzes the shape and size of different objects [33].

(iv) ***Occipital Lobe –*** It is present in the rear part of the brain and is responsible for the sense of vision.

Many neurophysiological studies have reported that emotional activity occurs mainly in two areas of the brain – the pre-frontal cortex and the amygdala (frontal part of the temporal lobe) [34]. Researchers have also shown that the left hemisphere is more closely related to positive emotions (high valence) while the right hemisphere is more closely related to negative emotions (low valence) [35], [36], [37].

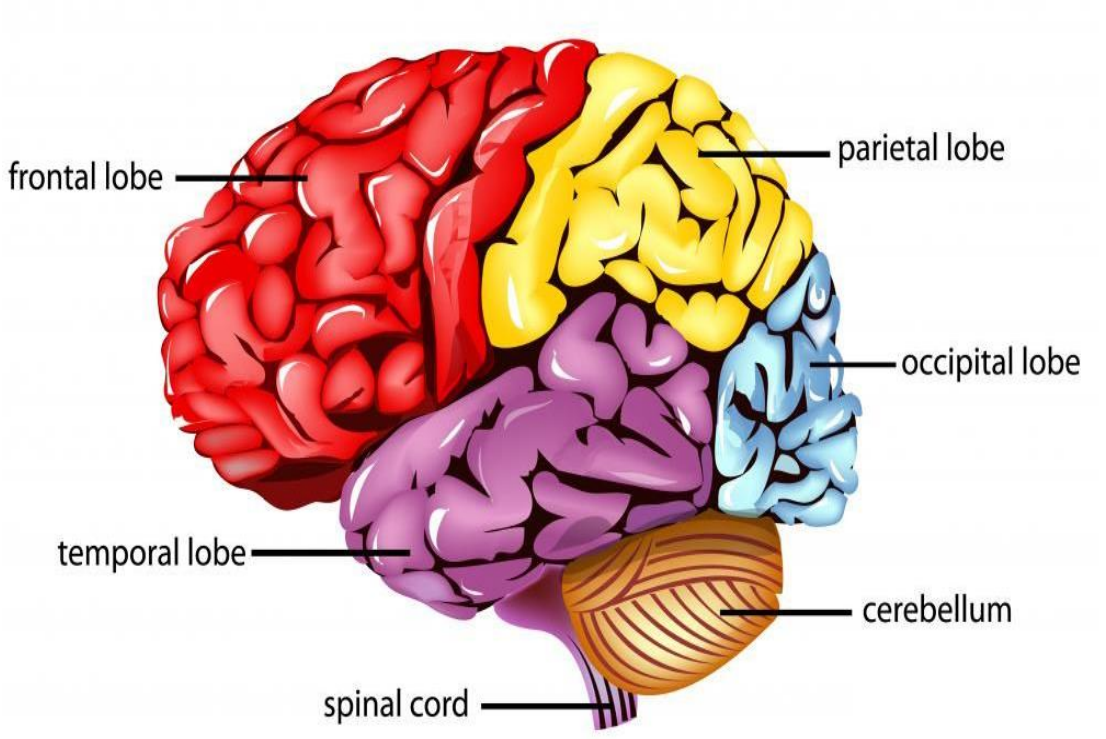

Figure 3.2 Different types of lobes in the human brain [34].

3.3 Physiological Signals

Whenever a person gets excited or experiences a change in emotion, various physiological changes take place in the body and in response to these changes, several physiological signals are generated [4]. The Physiological signals arising from the human body are spontaneous and cannot be controlled. These signals convey useful information regarding emotions and have often been analyzed for emotion recognition. The physiological signals can be broadly divided into two categories. Both are described in the following subsections.

3.3.1 Electroencephalogram (EEG) Signals

EEG signals represent the electrical activity of the brain and are measured via a non-invasive technique in which multiple electrodes are placed on the surface of the head. The electrodes are placed on standard locations on the scalp, according to the 10-20 International System [38]. This system defines the relationship between the location of electrodes and the underlying area of the cerebral cortex. The adjacent electrodes are placed at either 10% or 20% of the total distance from the front to the back of the skull or from left to right of the skull. Figure 3.3 shows the placement of electrodes on the scalp according to the 10-20 International System. As can be seen in the figure, electrodes are named according to the lobes in which they are present. For example, F stands for frontal, T for temporal, P for parietal and O for occipital. Letter ‘C’ stands for Central. There is no central lobe and it is just used for reference purposes. Also, even-numbered electrodes are placed in the right hemisphere while the odd-numbered electrodes are placed in the left hemisphere. Hence, “P7” refers to an electrode placed in the parietal part of the left hemisphere. ‘A’ refers to electrodes placed on the earlobe and ‘z’ refers to electrodes placed on the midline.

Many neurophysiological studies have reported that EEG is highly correlated with emotions. It gives a direct measurement of the underlying emotional activity occurring in the brain and it cannot be concealed or manipulated. It is measured in microvolts and for a typical adult it ranges from 10 to 100 µV [39]. It is related to the activity of the Central Nervous System (CNS) of our body, [4] which includes the brain and spinal cord. Measuring EEG is useful as it describes how different neurons in the brain communicate with each other with the help of electrical signals.

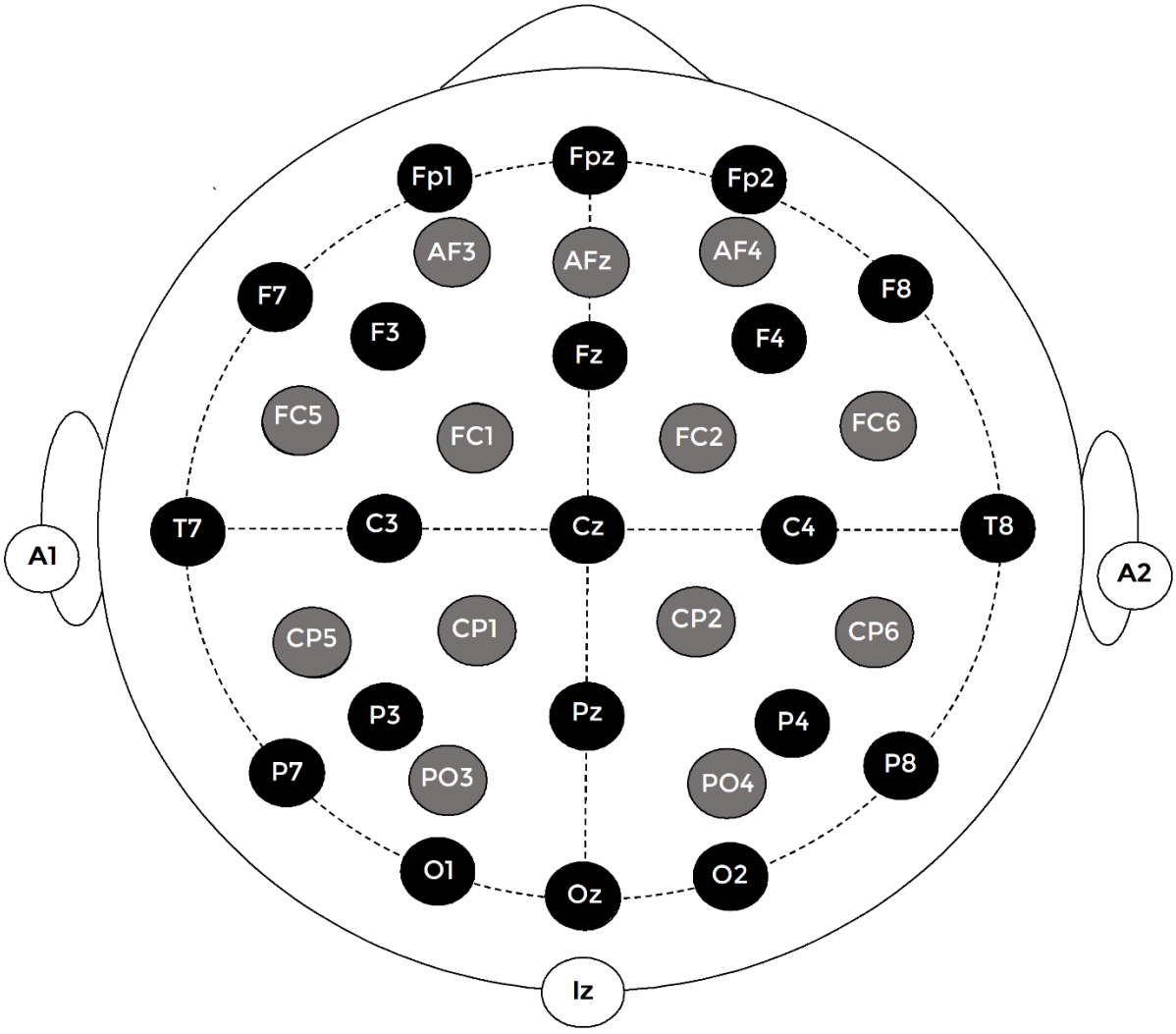

Figure 3.3 10-20 International system.

3.3.2 Peripheral Physiological Signals

These signals are elicited by the activity of the Autonomous Nervous System in our body [4]. The following list describes the various types of peripheral physiological signals and their importance in emotion recognition –

(i) ***Respiration*** **–** It depicts the thoracal activity of the body. The rate and depth of breathing convey a lot of information about the affective and emotional state of the body. Studies have shown that rapid and deep breathing is an indicator of anger or happiness, slow breathing represents a relaxed affective state, negative valence and low arousal is described by irregular breathing, shallow respiration represents depression [40] and shallow and rapid breathing indicates fear [41].

(ii) ***Temperature*** **–** Skin temperature is measured from the fingertip and is a useful indicator of affective state. A relaxed body state dilates the blood vessels which in turn makes the fingertip warm. On the other hand, anxiety, and stress decreases the temperature of the fingertip as the blood vessels constrict [41].

(iii) ***Electrocardiogram (ECG)–*** It represents the cardiac activity or heart rate of the body. Heart rate has been popularly used for detecting arousal levels in the body by using a comparison between sympathetic (increase) and parasympathetic (decrease) frequency bands of the ANS system [41]. ECG is also known to be a good indicator of mental effort and stress in humans [40].

(iv) ***Electromyogram (EMG)–*** It describes the muscle movements of the body like muscle contraction and relaxation. It is measured by attaching electrodes to the skin. A stressed body leads to highly tensed muscles whereas contraction in muscles is related to high

valence [41]. In addition to this, facial muscles also depict a lot of information about emotional states [42].

(v) ***Galvanic Skin Response (GSR)*** **–** It is another important signal which represents the variations in the conductivity or electrical activity of the skin. These changes in electrical conductivity arise because of the presence of sweat glands and it is a good indicator of arousal levels of the sympathetic nervous system [43]. Some studies have also shown that skin conductivity increases when the body undergoes stress and effort and it decreases when the body is in a relaxed state [41].

(vi) ***Electrooculography (EOG)*** **–** It depicts the eye movements of the body. Eye movements and blinks give important information about fatigue and anxiety [44]. Eye gaze is an indicator of the user's attention level [45] and social engagement [46].

Chapter 4: Dataset - DEAP

The dataset used in this study is DEAP [8]. It is a large publicly available multimodal dataset which is often used by researchers for the analysis of affective states. It is a collection of different physiological signals which include emotional and cognitive information. The dataset consists of Electroencephalogram (EEG) signals and peripheral physiological signals which include Galvanic Skin Response (GSR), respiration amplitude, skin temperature, EOG (eye movements), EMG (muscle movements) and blood pressure. While data recording, these signals were captured at a sampling rate of 512 Hz.

The data was collected from 32 participants (16 males and 16 females), aged between 19 and 37 years. To elicit emotions, each participant watched 40 one-minute long clips of music videos. After watching each video clip, the participants reported their perceived emotional state on a scale from 1 to 9 by using the Self-assessment manikins (SAM) [47] tool. They gave their ratings in terms of valence, arousal, dominance and like/dislike.

The dataset contains 40 channels which include 8 channels of peripheral physiological data and 32 channels of EEG data. The 32 channeled EEG data was obtained from 32 electrodes which were placed according to the 10-20 International system [38]. The different types of EEG and peripheral physiological channels used in DEAP are shown in Table 4.1.

Table 4.1 EEG (1-32) and peripheral (32-40) channels used in DEAP dataset [8].

Channel Number	Name of the Channel
1	Fp1
2	AF3
3	F3
4	F7
5	FC5
6	FC1
7	C3
8	T7
9	CP5
10	CP1
11	P3
12	P7
13	PO3
14	O1
15	Oz
16	Pz
17	Fp2
18	AF4
19	Fz
20	F4
21	F8
22	FC6
23	FC2
24	Cz
25	C4
26	T8
27	CP6
28	CP2
29	P4
30	P8
31	PO4
32	O2
33	hEOG (horizontal EOG, $hEOG_1$ – $hEOG_2$)
34	vEOG (vertical EOG, $vEOG_1$ – $vEOG_2$)
35	zEMG (Zygomaticus Major EMG, $zEMG_1$ – $zEMG_2$)
36	tEMG (Trapezius EMG, $tEMG_1$ – $tEMG_2$)
37	GSR (Ohm)
38	Respiration belt
39	Plethysmograph
40	Temperature

Each participant had a data array (physiological recordings) and a label array (valence, arousal, dominance and liking ratings). Table 4.2 shows the dimensions and content of data and label array. In the table, 8064 data points represent 63 seconds of recordings sampled at 128 Hz.

Table 4.2 Dimensions and content of each participant's array.

Array Type	**Array Dimensions**	**Array Contents**
Data	40 × 40 × 8064	Videos × Channels × Data Points
Labels	40 ×4	Videos × Labels (valence, arousal, dominance, and liking)

In this research, only valence and arousal ratings were taken into consideration as these two dimensions most commonly describe human emotional states [31]. In addition to this, all 40 channels of physiological data recordings were used. Overall, the physiological recordings represented the input data and the video ratings represented the output labels.

Chapter 5: Methodology

5.1 Data Preprocessing

Before the data can be used for classification, it must be processed. Preprocessing helps to transform noisy raw data into a relevant format that is reliable and more suitable for analysis. EEG data which is taken from the scalp has very low spatial resolution and is contaminated with different artefacts such as eye movements, muscle movements, etc. These artefacts are not produced by the brain and they reduce the quality of EEG signals. Similarly, peripheral physiological signals which are taken from the body are also contaminated with noise. Both artefacts and noise affect the analysis of signals and therefore they must be removed.

In this study, a preprocessed version of the DEAP dataset was used. On this data, some preprocessing steps had already been applied such as EOG artefact removal, downsampling (from 512 Hz to 128 Hz) and bandpass filtering (4.0 – 45.0 Hz). However, in this study, the preprocessed dataset was not used directly and instead a few more preprocessing steps were applied to it which are discussed in detail in the following subsections.

5.1.1 Baseline Removal

In the input data, for each participant and each video, 63 seconds of recordings were present in which the first 3 seconds represented pre-trail baseline data. These baseline signals were recorded when the participant was not watching the video and therefore, they are not useful for this study. They were removed to get 60 seconds of recording for each video, thereby, reducing the number of data recordings from 8064 (128 Hz × 63 seconds) to 7680 (128 Hz × 60 seconds).

5.1.2 Normalization

The input features in the raw data can have a different range of values. For example, one input feature can have values ranging from 0 to 1 while another can have values that range from 0 to 100. In such a case, the feature vector with a wider range of values will dominate the results. Normalization helps to make all the feature vectors fall within a similar range which in turn improves the efficiency of the predictive classifier [48].

In this study, the min-max normalization technique was used [49], [50] which linearly transformed all the elements in a feature to fit within the range [0-1]. It was applied to all the features individually. Mathematical formula of min-max normalization is given below –

$$x_{new} = \frac{(x - x_{min})}{(x_{max} - x_{min})} \tag{5.1}$$

In the above equation, x_{min} and x_{max} represent the minimum and maximum of feature vector x, respectively.

In addition to the input data, the output data was also processed. Originally, the valence and arousal ratings were in the form of real numbers from 1 to 9. For classification purposes, these ratings were converted into two classes. The threshold value was chosen as 5 [18]. Hence, for both valence and arousal, if the rating was less than 5, it was encoded to 0 (low valence/arousal) and if the rating was equal to or greater than 5, then it was encoded to 1 (high valence/arousal). Therefore, each video could be classified into high valence or low valence for valence dimension and high arousal or low arousal for the arousal dimension.

After preprocessing, the dimensions of the input data were 32 participants × 40 videos × 7680 data recordings × 40 channels (features). Also, for each participant, output dimensions were 40 videos × 2 output labels (arousal and valence).

5.2 Classification Models

A hybrid deep learning model was developed, for classifying emotions. It was a composition of two deep neural networks – Convolutional Neural Networks (1D-CNN) and Recurrent Neural Networks (RNN). For the convolutional part, 1D convolution layers were used and for the RNN part, Long Short-Term Memory units were used. As opposed to ML algorithms, which require hand-crafted features for classification, the deep neural networks, especially CNN and RNN, learn and extract hidden features in the data on their own. The EEG and peripheral signals are recorded over time and therefore, they contain temporal dependencies. These dependencies can be successfully extracted with the help of a 1D CNN-LSTM network. The 1D-CNN units extract spatial and temporal features from the data and identify cross-channel correlations. On the other hand, the LSTM units, extract the temporal dependencies and model the contextual information [51].

Apart from using the deep learning technique, this study also used three ML classifiers – SVM, KNN and RF to separately train the data and classify emotions. The details of all the classification algorithms are given in the following subsections.

5.2.1 Proposed 1D-Convolutional LSTM Architecture

The proposed hybrid deep neural network comprised of 12 layers including multiple convolution, pooling, regularization, lstm, and dense layers. The first layer of the network is the 1D convolution layer. It takes the input data and generates a convolution kernel which convolves with the input layer and produces output in the form of tensors [6].

The activation function used in this layer is ReLU (Rectified Linear Unit) [52] which is most commonly used in convolutional neural networks [28], [53] due to its reliability [52]. It helps

in eliminating the vanishing gradient problem and offers better performance than sigmoid and tanh activation functions [54]. Its formula is given below [55] –

$$f(x) = max(0, x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (5.2)$$

The next layer used is 1D Max Pooling. It helps to avoid overfitting by downsampling the feature maps which are generated after convolution in the previous layers. It reduces the number of parameters and the amount of computation by using the strongest activation unit in the pooling region [56] and discarding the rest. It only passes the most present (max) features in the previous feature maps onto the next layer.

After max pooling, a dropout layer is used. Dropout [57] is a powerful regularization technique that reduces overfitting. It randomly drops neurons along with their connections from the network during training due to which resulting neurons become independent. It helps to make a robust neural model that becomes insensitive to the specific weight of neurons and produces better generalization. In the proposed model, the dropout rate was set at 0.2 which means that 20% of the inputs neurons were randomly dropped. There are three sets of convolutional layers, max-pooling layers, and dropout layers placed one after the other in the model.

After the last dropout layer, two LSTM layers are added. LSTM network has been proved to be efficient in handling data with temporal dependencies. It is better than Recurrent Neural Networks as it can overcome error backflow and vanishing gradient problems [7]. It has three gates input gate, output gate and forget gate which can control and update the cell states. The forget gate uses a sigmoid activation function [58] which decides what information to discard. The output of the forget gate is given by the equation below –

$$f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \quad (5.3)$$

The input gate consists of two activation functions – the sigmoid function which updates the values with new information and the tanh function [55] which creates vectors for the newly updated values. Their functionality is represented with the help of the following two equations –

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \tag{5.4}$$

$$V_t = \tanh(W_c \times [h_{t-1}, x_t] + b_c) \tag{5.5}$$

Using equations 5.3, 5.4 and 5.5, the cell state can be updated as follows –

$$C_t = f_t * C_{t-1} + i_t * V_t \tag{5.6}$$

Based on the updated cell state, the sigmoid layer in the output gate calculates the output. Its functionality is described by the following two equations –

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \tag{5.7}$$

$$h_t = o_t * \tanh(C_t) \tag{5.8}$$

In the above equations, σ denotes sigmoid activation function, tanh denotes hyperbolic tangent activation function, x_t is the input vector, W_f, W_i, W_c, W_o are weight matrices, b_f, b_i, b_c, b_o are bias vectors and h_{t-1} denotes the past hidden state [15].

Finally, the output of the LSTM layer is passed to a dense or fully connected layer. In a dense layer, each input neuron is connected to each output neuron with the help of a weight [59]. This layer takes input from the previous layer and generates a N-dimensional vector as output, where N represents the number of target classes. In this model, only one dense layer, with one neuron is used. Sigmoid [58] is used as the activation function in this layer because this is a binary classification problem [15]. It classified the data, obtained from the previous layer, into two classes 0 and 1. Its mathematical formula is given below [60] –

$$S(x) = \frac{1}{(1 + e^{-x})} \tag{5.9}$$

The model is compiled using RMSprop optimizer. It is a type of stochastic gradient descent which divides the learning rate with a running average of its recent magnitude [61]. The learning rate of the model is kept as 0.0001 [62]. Since the number of target classes are two, the loss function used during model compilation is binary_crossentropy [63]. The model converged after training for 20 epochs and its performance was evaluated using accuracy metric. Figure 5.1 shows the architecture of the proposed model.

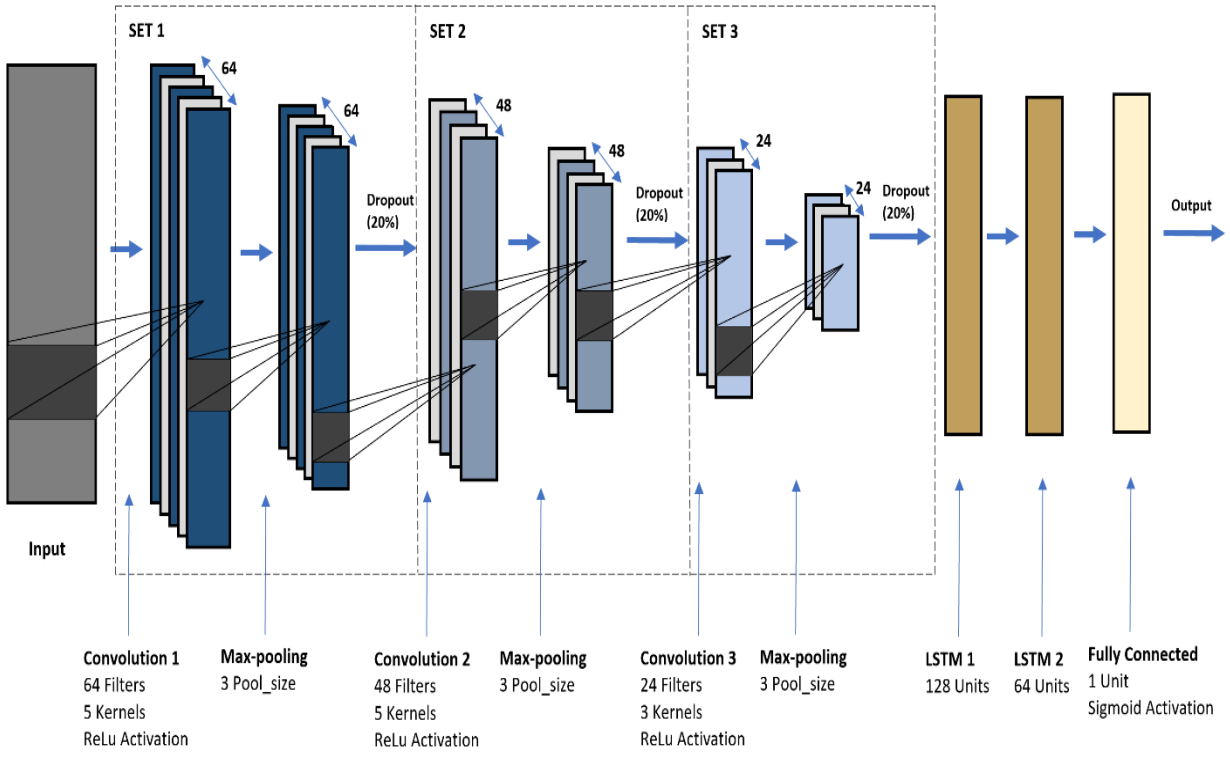

Figure 5.1 Proposed 1D CNN-LSTM architecture.

5.2.2 Machine Learning Classifiers

To evaluate the performance of the proposed 1D CNN-LSTM based deep learning approach against the traditional machine learning approaches, further training was done using the following three classifiers –

(i) ***Support Vector Machine (SVM)*** **–** Support Vector Machines [9] are supervised learning methods, used for classification, regression and outlier detection. They map non-linear input data points to high dimensional space and then construct a linear decision boundary or hyperplane to separate these data points into two or more classes [9]. They do so with the help of different types of kernel functions, the most common of which are 'Linear' and 'Radial Basis Function' kernel. Because of their high generalization ability, many researchers have used them for classifying emotions [16], [22]. Following some previous studies [25], [64], in this study also, SVM with RBF kernel was used as a baseline model for comparison.

(ii) ***K-Nearest Neighbor (KNN)*** **–** KNN is another supervised, statistics-based algorithm [10] which is used for classification and regression. For every new test instance, KNN identifies the closest K-samples in the training set using the Euclidean distance between the test sample and the training samples. The most common class value of K-training samples determines the target class of the test instance [10]. It is also commonly used for emotion recognition [65].

(iii) ***Random Forest (RF)*** **–** Random Forest [11] is a powerful ensemble algorithm that performs classification and regression by generating a forest of decision trees during training. The target class is determined by taking the mode of results of all the trees in the forest. It can handle high-dimensional data as it uses only a subset of features while

building decision trees. As compared to other classifiers, it takes very little time to train [66]. All these factors make Random Forest a very popular classification algorithm. In this project, an RF classifier with 100 decision trees was used for classification.

5.3 Training and Testing

The input data that was used for training had a total of 40 different features – 32 EEG features and 8 peripheral physiological features. Moreover, each video had two types of output – arousal and valence. In this study, training was completed in three different phases and in each phase different combinations of the channels/features were used to create input feature vectors. The feature vectors for the three phases were –

(i) ***Phase I*** **–** In this phase, three types of feature vectors were created. The first feature vector contained all the 40 features (EEG + peripheral). The second vector contained EEG features (32 channels) and the third feature vector had peripheral physiological features (8 channels).

(ii) ***Phase II*** **–** In phase II, input features were divided according to different brain lobes. The four brain lobes considered were – frontal, parietal, temporal and occipital. For each brain lobe, two feature vectors were made. The first vector had EEG data whereas the second vector had both EEG and peripheral data.

(iii) ***Phase III*** **–** In this phase, features were divided according to different brain hemispheres. There are two hemispheres in our brain – left and right. In this study, each of the two hemispheres was further divided into two parts – frontal and parietal-temporal-occipital. Thus, features were split according to four regions – left frontal, right frontal, left parietal-temporal-occipital and right parietal-temporal-occipital. For

each region, two feature vectors were made. The first vector had EEG data and the second one had both EEG and peripheral data.

Table 5.1, 5.2 and 5.3 lists the input feature vectors and the type of channels used for phases I, II and III, respectively. Moreover, in each phase, two sets of experiments were conducted. In the first experiment, training and testing were done on each participant individually. Data was divided in the ratio of 4:1, such that, out of 40 videos, 32 videos were used in the training set and the remaining 8 videos were present in the test set. Furthermore, training was done using four algorithms (proposed 1D CNN-LSTM model, SVM, KNN and RF) which have been described in the previous section. In the second experiment, training and testing were done by taking the data of all the participants. The training set had 80% of the participants and the test set had the remaining 20% of the participants. The training was done using three algorithms (proposed 1D CNN-LSTM model, KNN and RF). Moreover, for all the experiments, models were trained two times, once each for valence and arousal.

Combining phases I, II and III, there were a total of 19 types of feature vectors (3 from phase I, 8 from phase II and 8 from phase III). In the first experiment, when training was done on all subjects individually, a total of 4,864 training models (32 participants × 19 feature vectors × 4 algorithms × 2 output dimensions valence/arousal) were created while in the second experiment, 114 models were created (19 feature vectors × 3 algorithms × 2 output dimensions valence/arousal).

Table 5.1 Types of features/channels used for phase I.

Feature Vector Name	Number of Features	Types of Features/Channels
EEG + Peripheral	40	40 features (32 EEG + 8 peripheral)
EEG	32	32 EEG channels
Peripheral	8	8 Peripheral channels

Table 5.2 Types of features/channels used for phase II.

Feature Vector Name	Number of Features	Types of Features/Channels
Frontal Lobe	13	Fp1, F3, AF3, F7, FC5, FC1, Fp2, AF4, Fz, F4, F8, FC6 and FC2
Frontal Lobe + Peripheral	21	Fp1, F3, AF3, F7, FC5, FC1, Fp2, AF4, Fz, F4, F8, FC6, FC2 and 8 peripheral
Parietal Lobe	11	CP5, CP1, P3, P7, PO3, Pz, CP6, CP2, P4, P8, and PO4
Parietal Lobe + Peripheral	19	CP5, CP1, P3, P7, PO3, Pz, CP6, CP2, P4, P8, PO4 and 8 peripheral
Temporal Lobe	5	T7, T8, C3, Cz and C4
Temporal Lobe + Peripheral	13	T7, T8, C3, Cz, C4 and 8 peripheral
Occipital Lobe	3	O1, O2 and Oz
Occipital Lobe + Peripheral	11	O1, O2, Oz and 8 peripheral

Table 5.3 Types of features/channels used for phase III.

Feature Vector Name	Number of Features	Types of Features/Channels
Left Frontal	8	Fp1, AF3, F3, F7, FC5, FC1, C3 and Cz
Left Frontal + Peripheral	16	Fp1, AF3, F3, F7, FC5, FC1, C3, Cz and 8 peripheral
Right Frontal	8	FC2, FC6, F8, F4, AF4, FP2, Fz and C4
Right Frontal + Peripheral	16	FC2, FC6, F8, F4, AF4, FP2, Fz, C4 and 8 peripheral
Left Parietal-Temporal-Occipital	8	O1, PO3, P7, P3, CP1, CP5, T7 and Oz
Left Parietal-Temporal-Occipital + Peripheral	16	O1, PO3, P7, P3, CP1, CP5, T7, Oz and 8 peripheral
Right Parietal-Temporal-Occipital	8	O2, PO4, P8, P4, CP2, CP6, T8 and Pz
Right Parietal-Temporal-Occipital + Peripheral	16	O2, PO4, P8, P4, CP2, CP6, T8, Pz and 8 peripheral

Chapter 6: Results and Analysis

The training was completed in three phases and in each phase two sets of experiments were conducted. In the first experiment (Experiment 1), training and testing were done on each participant individually and final accuracy was calculated by taking the average of all the results. In the second experiment (Experiment 2), data of 80% of participants was used for training and testing was done on the data of the remaining 20% of participants. This was done to test the generalization capabilities of the proposed model.

6.1 Results – Phase I

Table 6.1 shows the average accuracy of all the experiments of phase I. The results show that the proposed 1D CNN-LSTM model outperformed all the machine learning classifiers and gave higher accuracy for all the experiments. For both valence and arousal dimensions, the proposed model gave higher results with EEG + peripheral physiological data (multimodal data). This is true for both experiments 1 and 2. This result conforms with other studies that show that multimodal data leads to higher accuracy as compared to unimodal data [67]. In almost all the experiments, arousal achieved higher accuracy than valence. Moreover, out of all the three types of features, EEG features gave the worst accuracy.

In the case of Machine Learning classifiers, KNN gave better results than SVM and RF in experiment 1. Similarly, in experiment 2, KNN gave higher accuracy than RF classifier.

6.2 Results – Phase II

Table 6.2 shows the average results for phase II in which data of different brain lobes (with and without peripheral physiological data) was used for classification. Again, it can be seen that

the proposed neural network model achieved higher accuracy as compared to machine learning models.

In experiment 1, when training was done on each participant individually, the brain lobe data (EEG data) combined with peripheral data gave better results than brain lobe data alone. This is true for all the brain lobes. Similarly, in experiment 2, the combination of brain lobe and peripheral data gave higher accuracy than brain lobe data.

Out of all the machine learning classifiers, KNN gave the highest accuracy for both arousal and valence dimensions.

6.3 Results – Phase III

Table 6.3 shows the average results for phase III in which data from different regions of the two brain hemispheres (with and without peripheral physiological data) was used for classification. It can be seen that the proposed neural model outperformed the machine learning classifiers for both the experiments and for both arousal and valence dimensions. Also, KNN gave better accuracy than other machine learning classifiers.

Another interesting observation is that when left frontal and right frontal features are individually combined with peripheral features, they produced the highest accuracy results for arousal and valence respectively. These values are higher than the results obtained in phase I and phase II, thereby, showing that left and right frontal regions can classify emotions more efficiently than other regions of the brain.

Table 6.1 Average results for phase I.*

Types of Features	Types of Models	Experiment 1		Experiment 2	
		Valence	*Arousal*	*Valence*	*Arousal*
EEG + Peripheral	***1D CNN-LSTM***	**91.19**	**91.51**	**70.28**	**71.04**
	SVM	70.65	70.29	-	-
	KNN	86.40	86.64	68.78	68.90
	RF	84.73	84.69	66.56	66.28
EEG	***1D CNN-LSTM***	**63.02**	**67.34**	**56.57**	**58.92**
	SVM	60.05	62.46	-	-
	KNN	61.37	65.57	54.52	53.37
	RF	60.85	64.63	55.10	53.80
Peripheral	***1D CNN-LSTM***	**87.23**	**89.95**	**69.45**	**70.92**
	SVM	76.36	76.52	-	-
	KNN	85.63	86.68	68.52	68.64
	RF	85.69	85.17	66.19	67.26

The results are expressed as a percent (%).

Table 6.2 Average results for phase II. *

Types of Features	Types of Models	Experiment 1		Experiment 2	
		Valence	*Arousal*	*Valence*	*Arousal*
Frontal Lobe	***1D CNN-LSTM***	**62.53**	**67.37**	**59.57**	**58.92**
	SVM	60.03	62.50	-	-
	KNN	61.15	65.40	54.11	55.88
	RF	60.54	64.56	53.39	55.96
Frontal Lobe + Peripheral	***1D CNN-LSTM***	**92.74**	**91.32**	**71.93**	**71.08**
	SVM	72.43	71.43	-	-
	KNN	88.11	88.23	68.05	68.86
	RF	87.38	87.21	66.65	66.35
Parietal Lobe	***1D CNN-LSTM***	**62.33**	**67.30**	**57.57**	**57.84**
	SVM	60.10	62.63	-	-
	KNN	60.66	65.10	55.04	54.93
	RF	60.06	64.25	53.63	54.27
Parietal Lobe + Peripheral	***1D CNN-LSTM***	**88.21**	**91.44**	**69.01**	**69.99**
	SVM	72.78	72.35	-	-
	KNN	87.06	87.17	66.80	67.69
	RF	86.49	86.26	66.09	67.57
Temporal Lobe	***1D CNN-LSTM***	**62.51**	**67.26**	**59.45**	**58.72**
	SVM	60.44	63.20	-	-
	KNN	60.67	64.53	54.35	55.82
	RF	59.99	63.86	53.74	56.03
Temporal Lobe + Peripheral	***1D CNN-LSTM***	**91.47**	**91.81**	**71.51**	**70.92**
	SVM	74.22	73.08	-	-
	KNN	88.88	88.94	68.11	68.69
	RF	87.61	87.31	67.69	66.50
Occipital Lobe	***1D CNN-LSTM***	**61.75**	**67.16**	**56.57**	**58.92**
	SVM	60.43	62.88	-	-
	KNN	59.48	64.17	52.14	54.27
	RF	58.43	62.76	52.44	55.62
Occipital Lobe + Peripheral	***1D CNN-LSTM***	**90.89**	**91.49**	**68.57**	**69.92**
	SVM	74.94	74.65	-	-
	KNN	88.98	88.68	67.81	68.18
	RF	86.33	86.28	67.03	67.04

The results are expressed as a percent (%).

Table 6.3 Average results for phase III. *

Types of Features	Types of Models	Experiment 1		Experiment 2	
		Valence	*Arousal*	*Valence*	*Arousal*
Left Frontal	***1D CNN-LSTM***	**62.05**	**65.23**	**58.57**	**58.93**
	SVM	58.28	59.00	-	-
	KNN	60.93	60.22	57.04	60.60
	RF	60.15	60.28	55.89	59.28
Left Frontal + Peripheral	***1D CNN-LSTM***	**93.67**	**94.15**	**72.87**	**71.39**
	SVM	73.44	73.64	-	-
	KNN	88.61	88.47	65.98	66.64
	RF	87.58	87.14	65.75	66.59
Right Frontal	***1D CNN-LSTM***	**61.96**	**65.22**	**58.57**	**58.93**
	SVM	60.28	61.53	-	-
	KNN	60.64	65.06	57.03	60.94
	RF	59.87	64.14	56.07	59.40
Right Frontal + Peripheral	***1D CNN-LSTM***	**93.69**	**93.99**	**71.20**	**71.03**
	SVM	73.44	73.65	-	-
	KNN	88.51	88.33	65.53	66.19
	RF	87.59	87.15	65.72	66.45
Left Parietal-Temporal-Occipital	***1D CNN-LSTM***	**61.94**	**65.38**	**56.57**	**58.92**
	SVM	60.27	61.52	-	-
	KNN	60.76	65.00	56.84	60.19
	RF	60.24	64.24	55.84	59.09
Left Parietal-Temporal-Occipital + Peripheral	***1D CNN-LSTM***	**93.65**	**93.70**	**69.67**	**69.41**
	SVM	73.44	73.67	-	-
	KNN	88.37	88.47	64.84	64.61
	RF	86.59	87.13	64.67	65.52
Right Parietal-Temporal-Occipital	***1D CNN-LSTM***	**61.08**	**65.42**	**56.57**	**58.99**
	SVM	60.29	61.57	-	-
	KNN	60.62	64.97	57.14	60.53
	RF	59.94	64.11	55.82	59.21
Right Parietal-Temporal-Occipital + Peripheral	***1D CNN-LSTM***	**93.57**	**94.00**	**69.99**	**69.30**
	SVM	73.43	74.10	-	-
	KNN	89.38	88.50	65.73	65.46
	RF	87.59	88.13	64.62	64.11

The results are expressed as a percent (%).

6.4 Discussion

In this research, a deep 1D Convolutional-LSTM neural network is proposed for the classification of emotions into valence and arousal dimension using the physiological signals of the human body. Previously, hybrid models of 1D CNN-LSTM have been used for speech emotion recognition [13], [14] or for recognizing emotions through auditory and visual modalities [12], but, this is the first study which has used 1D CNN-LSTM architecture for recognizing emotions based on physiological signals of the human body. The results show that the proposed model when used with EEG + peripheral data, predicts with high classification accuracy.

Upon observing Table 6.1, we can conclude that multimodal data is better than unimodal data for classifying emotions as the combination of EEG and peripheral data gave higher accuracy than when EEG and peripheral data were used separately. In a multimodal dataset, each of the modalities has its own distinct properties and combining them allows us to learn more useful representations of data.

The peripheral data performed better than EEG data. The poor EEG performance can be attributed to low spatial resolution. EEG method records brain signals from the top of the scalp. These brain signals originate from neurons and pass through several layers of skin before they reach the electrodes, therefore, they are of low quality. Moreover, EEG is contaminated with artefacts like heart rate (ECG) and eye movements (EOG). Even though EOG signals can be separated from EEG, ECG signals are very hard to remove. The presence of artefacts further degrades the quality of EEG data. The results with EEG data can be improved by using a more efficient recording device or by using a better artefact removal algorithm. Another reason why EEG data did not perform well as compared to peripheral data is that, for the same participant and

the same feature, EEG values varied a lot over time while the peripheral data values remained almost constant.

The performances of 1D CNN-LSTM, K-Nearest Neighbor, Support Vector Machine and Random Forest models were compared, and out of the four models, 1D CNN-LSTM gave the best results, whereas, SVM was least accurate. It is because deep neural networks can extract complex features and learn intricate non-linear interactions in the data in a more robust manner. Another reason why the proposed model performed better is due to the high temporal resolution of physiological data. Deep neural networks are more suitable for large datasets and because of high temporal resolution of physiological signals there was a lot of data to work upon.

In almost all the experiments, KNN's performance was better than SVM and RF. It is because KNN uses the similarity between features as a basis for classification and the similarity between multichannel physiological data generated by the same stimulus and for the same person is higher than for different persons [16]. This accounts for the high average accuracy of KNN in experiment 1 when training and testing were done on each participant individually.

As seen in Tables 6.1, 6.2 and 6.3, the proposed deep neural model gave high classification accuracy when used with a combination of EEG and peripheral data, but could not perform well when used with EEG data alone. This shows that more EEG data is required to get high classification results with the proposed model.

To evaluate generalization, data of different participants were used for training and testing. While the results were encouraging for individual subjects, they were not promising across different participants. As can be seen in experiment 2 section of Table 6.1, EEG + peripheral data produced the best results as compared to the other two modalities. Again, in this case, EEG performed poorly. This can be attributed to the poor generalizability of EEG features across

subjects [71]. The information contained in EEG data varies from person to person and it is very hard to learn and analyze one person's EEG data and then use it to predict the emotional state of an unknown person. DEAP dataset has only 32 participants and therefore, it is difficult to generalize results with such a small number of subjects. One solution is to build larger datasets with more number of participants. Another reason why all the models failed to generalize across subjects is that people do not self-assess their emotional state in a similar manner.

A lot of previous studies have reported emotion classification accuracies based on entire EEG data of the DEAP dataset. However, it should be noted that when a person expresses emotion, then the related brain activity becomes dominant in only a few regions of the brain. The four lobes or the two hemispheres of the brain are responsible for different functionalities and they generate different types of emotional brain activity. Therefore, for emotion recognition, a separate analysis of these regions makes more sense. Thus, this study analyzed the classification performance of different brain lobes and hemispheres. Tables 6.2 and 6.3, show the arousal and valence classification accuracies for the different brain lobes and the different regions of the two brain hemispheres.

Previous studies have shown that the frontal and temporal lobes of the brain and left and right frontal regions of the brain showcase higher emotional activity [34], [35], [36], [37]. The results of this study also support this fact. As seen in Table 6.2, the frontal and temporal lobe data gave higher classification results and similarly, in Table 6.3, the left and right frontal regions performed better than other regions of the brain.

Chapter 7: Conclusion and Future Work

The emerging fields of Affective Brain-Computer Interface (Affective-BCI) and Affective Computing (AC) focuses on developing technologies that enhance the computer's ability to perceive user's emotional state during human-computer interactions and has a potential for a wide range of applications. The focus of this thesis lies at the intersection of Affective-BCI and Affective Computing and it tries to introduce a new method to recognize emotions with better performance. The classification was done using deep neural networks which have been currently, recognized as one of the most powerful and robust techniques available for classification. The experiments were conducted on the DEAP dataset which consists of physiological data of 32 subjects.

The 1D CNN-LSTM architecture built in this study achieved a high classification accuracy for both arousal and valence dimensions of data. The proposed network's performance was compared with machine learning classifiers and the deep neural model achieved high classification results. The results also show that the multimodal approach that combines different physiological signals of the body, is better than the unimodal approach for classifying emotions. Overall, the left frontal region data when combined with peripheral data gave the highest accuracy. Another interesting result that can be observed is that for almost all the experiments, accuracy for the arousal dimension was more than the valence dimension.

One of the challenges that occurred during this study was that the model was unable to analyze the data taken from a single EEG channel. It happened because of the lack of data for a single channel in the DEAP database. Therefore, a new database is needed which has enough data

for every channel so that performance analysis can be done per channel. Another solution is to build a new network that can work with fewer data and can perform analysis on each channel.

Further studies can be done to analyze the performance of the proposed 1D CNN-LSTM model on a different dataset. This will help to validate the performance of this model. Data from other sources like facial images and speech can also be combined with physiological data to find out how the added modalities affect the performance. Another future goal is to identify the effect of gender and age in emotion classification. This thesis adopted a binary classification approach by classifying emotions into low/high valence and low/high arousal. Further studies can be done to classify emotions into multiple classes such as happy, sad, joy, angry, etc.

The performance of the 1D CNN-LSTM model can also be analyzed on a different dataset that has more EEG data. We wish to get a higher emotion classification accuracy with the proposed model and the EEG data as opposed to using a combination of EEG and peripheral data so that the 1D CNN-LSTM model can be for future BCI applications.

References

[1] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cognition and emotion*, vol. 23, no. 2, pp. 209-237, 2009.

[2] B. Reeves and C. I. Nass, "The media equation: How people treat computers, television, and new media like real people and places," Cambridge university press, 1996.

[3] T. S. Polzin and A. Waibel, "Emotion-sensitive human-computer interfaces," *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[4] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, pp. 2074, 2018.

[5] S. Jerritta, M. Murugappan, R. Nagarajan and K. Wan, "Physiological signals based human emotion Recognition: a review," *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, Penang, 2011, pp. 410-415.

[6] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," 2016, *arXiv:1603.07285*.

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

[8] S. Koelstra *et al*., "DEAP: A Database for Emotion Analysis; Using Physiological Signals," in *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18-31, Jan.-March 2012.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273-297, 1993.

[10] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, pp. 1883, 2009.

[11] T. K. Ho, "Random decision forests," *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, Quebec, Canada, vol. 1, pp. 278-282, 1995.

[12] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks," in *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301-1309, Dec. 2017.

[13] N. Kurpukdee, T. Koriyama, T. Kobayashi, S. Kasuriya, C. Wutiwiwatchai and P. Lamsrichan, "Speech emotion recognition using convolutional long short-term memory neural network and support vector machines," *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Kuala Lumpur, 2017, pp. 1744-1749.

[14] J. Zhao, X. Mao and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312-323, 2019.

[15] S. Alhagry, A. Aly, and R. A., "Emotion Recognition based on EEG using LSTM Recurrent Neural Network," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, 2017.

[16] N. Liu, Y. Fang, L. Li, L. Hou, F. Yang and Y. Guo, "Multiple Feature Fusion for Automatic Emotion Recognition Using EEG Signals," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, 2018, pp. 896-900.

[17] W. Liu, J. Qiu, W. Zheng, and B. Lu, "Multimodal Emotion Recognition Using Deep Canonical Correlation Analysis," 2019, *arXiv:1908.05349*.

[18] W. Liu, W. Zheng, and B. Lu, "Emotion Recognition Using Multimodal Deep Learning," *Neural Information Processing Lecture Notes in Computer Science*, pp. 521–529, 2016.

[19] W. Liu, W. Zheng and B. Lu, "Multimodal emotion recognition using multimodal deep learning," *2016, arXiv:1602.08225*.

[20] A. Ben Said, A. Mohamed, T. Elfouly, K. Harras and Z. J. Wang, "Multimodal Deep Learning Approach for Joint EEG-EMG Data Compression and Classification," *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, San Francisco, CA, 2017, pp. 1-6.

[21] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion Assessment: Arousal Evaluation Using EEG's and Peripheral Physiological Signals," *Multimedia Content Representation, Classification and Security Lecture Notes in Computer Science*, pp. 530–537, 2006.

[22] W. Zheng, B. Dong and B. Lu, "Multimodal emotion recognition using EEG and eye tracking data," *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, IL, 2014, pp. 5040-5043.

[23] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage*, vol. 102, pp. 162–172, 2014.

[24] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," in *IEEE Transactions on Information Theory*, vol. 36, no. 5, pp. 961-1005, Sept. 1990.

[25] H. Ranganathan, S. Chakraborty and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Placid, NY, 2016, pp. 1-9.

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

[27] W. Lin, C. Li, and S. Sun, "Deep Convolutional Neural Network for Emotion Recognition Using EEG and Peripheral Physiological Signal," *Lecture Notes in Computer Science Image and Graphics*, pp. 385–394, 2017.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[29] L. Du, W. Liu, W. Zheng and B. Lu, "Detecting driving fatigue with multimodal deep learning," *2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*, Shanghai, 2017, pp. 74-77.

[30] S. Rayatdoost and M. Soleymani, "CROSS-CORPUS EEG-BASED EMOTION RECOGNITION," *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, Aalborg, 2018, pp. 1-6.

[31] J. A. Russell, "A circumplex model of affect.," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[32] J. Posner, J. A. Russell and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715-734, 2005.

[33] B. Farnsworth. "EEG (Electroencephalography): The Complete Pocket Guide". Imotions.com.
https://imotions.com/blog/eeg/ (accessed Feb. 15, 2020).

[34] S. M. Alarcão and M. J. Fonseca, "Emotions Recognition Using EEG Signals: A Survey," in *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374-393, 1 July-Sept. 2019.

[35] R. J. Davidson, "Anterior cerebral asymmetry and the nature of emotion," *Brain and cognition*, vol. 20, no. 1, pp. 125-151, 1992.

[36] N. A. Fox, "If it's not left, it's right: Electroencephalograph asymmetry and the development of emotion," *American psychologist*, vol. 46, no. 8, pp. 863, 1991.

[37] G. Gainotti, "The role of the right hemisphere in emotional and behavioural disorders of patients with Fronto-temporal degeneration: An updated review," *Frontiers in aging neuroscience*, vol. 11, pp. 55, 2019.

[38] G. H. Klem, H. O. Lüders, H. H. Jasper, and C. Elger, "The ten-twenty electrode system of the International Federation," *Electroencephalogr Clin Neurophysiol*, vol. 52, no. 3, pp. 3-6, 1999.

[39] M. Teplan, "Fundamentals of EEG measurement," *Measurement science review*, vol. 2, no. 2, pp. 1-11, 2002.

[40] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067-2083, Dec. 2008.

[41] E. Maria, L. Matthias and H. Sten, "Emotion recognition from physiological signal analysis: a review," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35-55, 2019.

[42] D. Kulic and E. A. Croft, "Affective State Estimation for Human–Robot Interaction," in *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991-1000, Oct. 2007.

[43] L. Li and J. Chen, "Emotion recognition using physiological signals," *International Conference on Artificial Reality and Telexistence*, Berlin, Heidelberg, pp. 437-446, 2006.

[44] M. W. Johns, A. Tucker, R. Chapman, K. Crowley and N. Michael, "Monitoring eye and eyelid movements by infrared reflectance oculography to measure drowsiness in drivers," *Somnologie-Schlafforschung und Schlafmedizin*, vol. 11, no. 4, pp. 234-242, 2007.

[45] E. Gu and N. I. Badler, "Visual attention and eye gaze during multiparty conversations with distractions," *International workshop on intelligent virtual agents*, Berlin, Heidelberg, pp. 193-204, 2006.

[46] L. Schulze, B. Renneberg and J. S. Lobmaier, "Gaze perception in social anxiety and social anxiety disorder," *Frontiers in human neuroscience*, vol. 7, pp. 872, 2013.

[47] M. M. Bradley and P. J. Lang, “Measuring emotion: The self-assessment manikin and the semantic differential,” *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[48] P. Trebuňa, J. Halčinová, M. Fil'o and J. Markovič, "The importance of normalization and standardization in the process of clustering," *2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, Herl'any, 2014, pp. 381-385.

[49] M. Mohammadi, F. Al-Azab, B. Raahemi, G. Richards, N. Jaworska, D. Smith, S. D. L. Salle, P. Blier, and V. Knott, “Data mining EEG signals in depression for their diagnostic value,” *BMC Medical Informatics and Decision Making*, vol. 15, no. 1, 2015.

[50] X. Zhang, L. Yao, D. Zhang, X. Wang, Q. Z. Sheng, and T. Gu, “Multi-Person Brain Activity Recognition via Comprehensive EEG Signal Analysis,” *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing Networking and Services*, pp. 28-37, 2017.

[51] 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.

[52] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807-814, 2010.

[53] C. Szegedy *et al.*, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 1-9.

[54] G. E. Dahl, T. N. Sainath and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, pp. 8609-8613.

[55] C. Nwankpa, I. Winifred, A. Gachagan and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," 2018, *arXiv:1811.03378*.

[56] H. Wu and X. Gu, “Max-Pooling Dropout for Regularization of Convolutional Neural Networks,” *International Conference on Neural Information Processing*, pp. 46–54, 2015.

[57] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929-1958, 2014.

[58] P. Sibi, S. A. Jones and P. Siddarth, "Analysis of different activation functions using back propagation neural networks," *Journal of Theoretical and Applied Information Technology*, vol. 47, no. 3, pp. 1264-1268, 2013.

[59] T. N. Sainath, O. Vinyals, A. Senior and H. Sak, "Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, QLD, 2015, pp. 4580-4584.

[60] B. Karlik and A. Vehbi, "Performance analysis of various activation functions in generalized MLP architectures of neural networks," *International Journal of Artificial Intelligence and Expert Systems*, vol. 1, no. 4, pp. 111-122, 2011.

[61] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26-31, 2012.

[62] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.

[63] N. Ketkar, "Introduction to keras," In *Deep learning with Python*, pp. 97-111, Apress, Berkeley, CA, 2017.

[64] Y. Kim, H. Lee and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, pp. 3687-3691.

[65] M. Li, H. Xu, X. Liu and S. Lu, "Emotion recognition from multichannel EEG signals using K-nearest neighbor classification," *Technology and Health Care*, vol. 26, no. S1, pp. 509-519, 2018.

[66] D. Ayata, Y. Yaslan and M. Kamaşak, "Emotion recognition via random forest and galvanic skin response: Comparison of time-based feature sets, window sizes and wavelet approaches," *2016 Medical Technologies National Congress (TIPTEKNO)*, Antalya, 2016, pp. 1-4.

[67] L. Kessous, G. Castellano and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *Journal on Multimodal User Interfaces*, vol. 3, no. 1-2, pp. 33-48, 2010.

[68] X. Li, P. Zhang, D. Song, G. Yu, Y. Hou and B. Hu, "EEG based emotion identification using unsupervised deep feature learning," 2015.

[69] Z. Yin, Y. Wang, L. Liu, W. Zhang and J. Zhang, "Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination," *Frontiers in neurorobotics*, vol. 11, pp. 19, 2017.

[70] V. Rozgić, S. N. Vitaladevuni and R. Prasad, "Robust EEG emotion classification using segment level decision fusion," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, 2013, pp. 1286-1290.

[71] X. Li, D. Song, P. Zhang, Y. Zhang, Y. Hou and B. Hu, "Exploring EEG features in cross-subject emotion recognition," *Frontiers in neuroscience*, vol. 12, pp. 162, 2018.

[72] D. Fabiano and S. Canavan, "Emotion Recognition Using Fused Physiological Signals," *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Cambridge, United Kingdom, 2019, pp. 42-48.