Fine-Grained Emotion Recognition from EEG Signal Using Fast Fourier Transformation and CNN

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address

***Abstract*—Emotions are mental states originating in the human brain, and this is closely related to the activities of the nervous system. Electroencephalogram (EEG) is a well-established approach to record neuron activities which is reliable for emotion recognition compared to the non-physiological clues. So far, there have been reports of various researches searching for active patterns involving different emotions. However, most of the previously published system could only classify 4 human emotions using the technique of binary classification but humans have more and complex emotions which couldn't be captured with only 4 classes. So, we proposed a fine-grained emotion classification technique which can classify 64 emotions including all the complex emotions. Hence, this paper presents convolutional neural network (CNN) models working on the DEAP dataset, and it contains emotional states which are arousal, valence, dominance and liking. Our binary models achieved 96.63% and 96.18% accuracy respectively for valence and arousal. Only four emotions are found with binary classification whereas 8-class classification can precisely recognize 64 emotions. The 8-class classification achieves a promising accuracy of 93.83 % and 93.79% respectively for valence and arousal. For both cases, Fast Fourier Transformation (FFT) has been used as the feature extraction method and all the four classification models are created under 1D-CNN using the same architecture.**

***Contribution*—This is an original paper successfully establishing an effective method to recognize 64 human emotions using FFT and CNN.**

***Keywords*—EEG; emotion recognition; CNN; DEAP; FFT**

I. Introduction

Emotions are a significant factor of human knowledge, behavior and communication. It is a neural impulse that moves an organism to action, prompting automatic reactive behavior that has been adapted through evolution as a survival mechanism to meet a survival need [1]. Emotion recognition is the process of identifying human emotions. Emotions can be divided into two categories: Positive and Negative [2]. Positive emotions are necessary for optimal health; however, bad emotions can lead to mental health issues like depression, stress and anxiety [3]. Emotions are known to arise from the central and peripheral nervous systems and cause temporal movement due to the synchronized execution of neurons [4]. They are expressed by internal signals as well as external expressions like facial expression, speech, body posture, eye blinking and skin response. If only external expressions are used for emotion measurement, incorrect results may be obtained, because in many cases external expressions can be controlled. That's why internal signals get priority. Electroencephalogram (EEG), Temperature (T), Electrocardiogram (ECG), Electromyogram (EMG), Galvanic Skin Response (GSR), Respiration (RSP) are examples of such internal signals. Excluding others, EEG is selected for its non-invasive, fast, and low-cost characteristics, as compared with other physiological signals.

Neuropsychological measurement of electrical activity in the brain is known as Electroencephalography (EEG) and it is recorded by electrodes. These are normally placed on the scalp, or in special cases, subdurally and in the cerebral cortex. EEG estimates voltage fluctuations that result due to ionic flows inside the neurons of the cerebrum. EEG provides an excellent temporal resolution, even though it has a poor spatial resolution and requires many sensors placed on the scalp. Pure EEG signal is a composition of sub-bands: Theta (3 - 7 Hz), Alpha (8 - 13 Hz), Beta (14 - 29 Hz) and Gamma (30 - 47 Hz) [17]. Individual sub-bands are associated with individual relevant physical activities. For instance, Theta wave refers to REM sleep, deep and raw emotions as well as cognitive processing. A drowsy state indicates the Alpha waves. It also becomes the cause of relaxation and calmness. Beta points to the conscious state during the thought process. The Gamma waves are available when trying to perceive two different senses at the same time as sound and sight [7].

Yet now, there comes out of extensive research using machine learning to identify states of emotion with EEG. Machine learning-based theoretical methods are often effectively used for emotion classification, while the disadvantage of such methods is that researchers have to spend a lot of effort to detect and design different emotion-related features from the resulting noisy signals and these features are time-consuming calculations. Various methods of EEG feature extraction have been explored in recent years, although the Fast Fourier Transformation (FFT), Short-time Fourier Transform (STFT), and Discrete Wavelet Transformation (DWT) etc are most effectively used. FFT is used to decompose this work's EEG signal data. Then deep learning models are applied to the extracted features so that they are trained for recognizing emotions. Support Vector Machine (SVM), Linear and nonlinear regression, Decision trees, and K-nearest neighbor (KNN) are examples of mostly used machine learning model architecture. Convolution Neural Networks (CNN), Long-Short Term Memory (LSTM), Convolution Long-Short Term Memory (CLSTM) are popular model architectures from deep learning for this field. There are numerous EEG datasets that are publically available and some researchers use their own dataset. Some famous publically available datasets are DEAP (A Database for Emotion Analysis using Physiological Signals), SEED (SJTU Emotion EEG Dataset) and MAHNOB (MAHNOB-HCI-Tagging database) etc. Among these, the DEAP dataset has been used for this research work.

From the EEG signal, using main data and only two emotion labels (arousal and valence), it is possible to recognize human emotion properly. Each emotion label is divided into two equal parts, and a total of four emotions are created by classifying the binary-class using both emotion labels. It is possible to increase the accuracy here more than in the previous work. We have tried and succeeded in overcoming this deficit through our experiment. Although most of the research is done with binary-class classification, these four classes are insufficient to accurately distinguish emotions. To recognize emotion more precisely, we have worked with sixty-four emotion spaces where many real-life emotions exist.
Significant contributions of this paper:

- We have brought the best accuracy using the binary classification which is the conventional emotion recognition method and for this, we have used a unique combination of FFT and 1D CNN model.
- Also we have proposed the 8-class emotion classification method which is able to recognize emotion much more accurately and we have also found a satisfactory classification accuracy for it.

The residual part of this study is highlighted as follow:

An overview of literature review in Section II. A brief description of the DEAP dataset in Section III. The research methodology of this paperwork is described including EEG data preprocessing, feature extraction, labeling and normalization as well as CNN model structure in Section IV. Experiment and result comparison with previous DEAP dataset-related work in Section V. An indication for our future work and the conclusion of this paper are given in Section VI.

II. Literature Review

There has been a lot of research on publicly available datasets (DEAP, SEED, MAHNOB, LUMED) for emotion recognition. However, some people have attempted to recognize emotions using their own datasets. Since the DEAP dataset is used in this paper, it would be better to review the contributions and the recent studies using the DEAP dataset. Table I attempts to represent literature reviewed papers in a concise manner.

Rahul Sharmaa Et al. [4] achieved an accuracy of 82.01% with a ten-fold cross-validation technique using Long short-term memory (LSTM) by decomposing with DWT. They worked on only two dimensions, namely arousal and valence, and with four quadrants respectively LaLv (low arousal low valence), HaLv (high arousal low valence), LaHv (low arousal high valence) and HaHv (high arousal high valence). Anubhav Et al. [7] earned a handsome accuracy when classifying emotions using valence and arousal labels, of 94.69% and 93.13% respectively. Although they tested KNN, SVM, Decision Tree, and Random Forest for classifying emotion, their best accuracy was achieved with the use of LSTM.

Zhongke Gao Et al. [9] proposed a model named Channel-fused dense convolutional network (CDCN), consisting of a 1D convolution layer and 1D dense layer. For pre-extracting, they used differential entropy (DE) and worked on four emotions. Their model applied on the SEED dataset and the DEAP dataset and obtained an accuracy of 90.60% and 92.58% respectively.

Yuling Luo Et al. [10] demonstrated their best performance with Spiking Neural Networks (SNNs) using three pre-extracting methods: DWT, Variance and FFT. They gained their best results with the use of variance, both on SEED and DEAP datasets. The emotion states of arousal, valence, dominance and liking were classified with accuracies of 74%, 78%, 80% and 86.27% for the DEAP dataset, as well as an overall accuracy of 96.67% for the SEED dataset.

Eman A. Abdel-Ghaffar Et al. [13] proposed a two-dimensional emotion model named Log-Euclidean Riemannian Metric (LERM) using Symmetric Positive Definite manifold (SPD). They received an accuracy of 88.30% for HVHA, 84.38% for LVHA, 79.30% for LVLA, and 78.40% for HVLA. The average accuracy for valence was 74.60% ± 3.9, and 72.60% ± 6.70 for arousal.

Fei Wang Et al. [5] used the Electrode-frequency distribution maps (EFDMs) model for classifying and Short-Time Fourier Transform (STFT) for feature extraction. Gradient weighted class Activation mapping (Grad-CAM) was used in their research to obtain a better understanding of their selected features. When they applied their model on the SEED dataset, they obtained 90.59% for accuracy, and on the DEAP dataset, they obtained an accuracy of 82.84%. However, they only worked with valence labels with three classes: negative, neutral, positive.

Kit Hwa Cheah Et al. [8] used two types of CNN models: single-path CNN and two-path CNN model using 4 folds of

TABLE I. Overview of Literature Review

No	Research	Year	Feature extraction	Modeling technique	No. of class for each emotional state	Working label	Performance
1	Rahul Sharmaa Et al. [4]	2020	DWT	LSTM	2-Class	Arousal, Valance	82.01%
2	Divya Acharya Et al. [3]	2020	FFT	LSTM	2-Class	Arousal, Valance	89.83%
3	Zhongke Gao Et al. [9]	2020	DE	CDCN	2-Class	Arousal, Valance	92.58%
4	Yulong Luo Et al. [10]	2020	Variance	SNN	2-Class	Arousal, Valance	Valence:78% Arousal:74%
5	Yucel Cimtay Et al. [11]	2020	Raw Data	CNN	2-Class	Arousal, Valance	72.81%
6	Eman A. Et al. [13]	2020	SPD	LERM	2-Class	Arousal, Valance	Valence: 74.6% ± 3.9, Arousal: 72.6% ± 6.7
7	Xiaolong Zhong Et al. [14]	2020	DE	CNN	2-Class	Arousal, Valance	Valence:66.23% Arousal:68.50%
8	Yucel Cimtay Et al. [15]	2020	Raw Data	CNN	2-Class	Arousal, Valance	91.5%
9	Guolu Cao Et al. [12]	2019	PCA	CNN	2-Class	Arousal, Valance	Valance: 81.2±3.0% Arousal: 84.3±4.0%
10	Soheil Rayatdoost Et a1. [1]	2018	HOC,PSD, DE, HOS	RF	2-Class	Arousal, Valance	Valence:60.86% Arousal:58.08%
11	Ningjie Liu Et a1. [16]	2018	LFCC	KNN, ResNets	2-Class	Arousal, Valance	Valence:90.39% Arousal:89.06%
12	Abeer Al-Nafjan Et a1. [22]	2017	PSD	DNN	2-Class	Arousal, Valance	Valence:82% Arousal:82%
13	Samarth Tripathi Et a1. [24]	2017	GD	CNN	2-Class	Arousal, Valance	Valence:81.41% Arousal:73.36%
14	Xiang Li Et a1. [21]	2016	CWT	C-RNN	2-Class	Arousal, Valance	Valence:72.06% Arousal:74.12%
15	Wei-Long Zheng Et a1. [25]	2016	DE	GELM	2-Class	Valence, Arousal	69.67%
16	Wei Liu Et a1. [23]	2016	PSD, DE	BDAE	2-Class	Valence, Arousal	Valence:85.20% Arousal:80.50%
17	Hyun Joong Yoon Et al. [18]	2013	FFT	Bayes classifier	2-Class	Valence, Arousal	Valence:70.9%, Arousal:70.1%
18	Viktor Rozgić Et al. [19]	2013	PCA	SVM	2-Class	Valence, Arousal	Valence:76.9% Arousal:68.4%
19	Xiaowei Zhang Et al. [20]	2013	Sliding 4-second windows with a 2-second overlap	Ontology Reasoning BIO-EMOTION	2-Class	Valence, Arousal	Valence:75.19% Arousal:81.74%

cross-validation. However, they did not use any manual pre-extraction methods, and instead, worked with only valence and arousal. Each emotional state was divided into three classes. Single-path CNN received an accuracy of 97.59% and 98.4% for 3-class valence and arousal, while two-path CNN received 98.75% and 97.58% for 3-class valence and arousal. Yucel Cimta Et al. [11] did not use any manual pre-extraction methods, and instead, depended on the CNN Deep Learning method. Three datasets were used in their research work: DEAP, SEED and LUMED. In the SEED dataset when studying two classes of emotions, an accuracy of 86.56% was obtained, and 78.34% was obtained for three classes of emotions. When applied to the DEAP dataset, they received 72.81% accuracy for two emotion states: valence and arousal. Guolu Cao Et al. [12] created a CNN model, utilizing Principal Component Analysis (PCA) as the pre-extracting technique. They worked with two classes for arousal and valence with an accuracy of 84.3±4.0% and 81.2±3.0% respectively. Xiaolong Zhong Et al. [14] concentrated on the physiological forms of brain waves. Their method was efficient in recognizing emotions, especially in beta and gamma waves. 2D SE_CNN was applied to the DEAP and the MAHNOB-HCI datasets. In their study, the DEAP dataset received an accuracy of 66.23% for valence and 68.50% for arousal, while the MAHNOB-HCI dataset obtained 70.25% for valence and 73.27% for arousal. Yucel Cimtay Et al. [15] used the InceptionResnetV2 CNN model, by following a hybrid fusion strategy on the DEAP dataset as well as LUMED-2 datasets with facial expressions and galvanic skin response (GSR). They achieved maximum of 91.5% accuracy on the DEAP dataset with arousal and valence.

It can be seen from the preceding discussion that practically almost everyone has worked on binary classification. However, the current testing accuracy is insufficient, and there is a need for improvement. Through our efforts, we have been able to attain the highest level of accuracy. With a binary classification that fails to distinguish real-life emotions, only four emotional zones can be found. To address this limitation, we have introduced an 8-class classification technique in our experiments, which can recognize a wide range of emotions and has also achieved satisfactory accuracy. In addition, Divya Acharya Et al. [3] used FFT and LSTM and Hyun Joong Yoon Et al. [18] used FFT and Bayes classifiers to build their models. But in this paper, FFT and CNN are applied to build a new model and this combination of approaches has not been done previously. So, this work is unique in two ways. One for classifying 64 emotions instead of 4 emotions and using a new combination of pre-processing technique and machine learning model.

III. Dataset

The DEAP dataset is a publicly available multimodal dataset [17] that includes electroencephalogram (EEG) signals and used for detecting human emotional states. This data was gathered by a specialized team of researchers from the Queen Mary University of London (UK), University of Twente (Netherlands), University of Geneva (Switzerland) and EPFL (Switzerland). The overview of the DEAP dataset can be found in Table II.

TABLE II. Synopsis of The DEAP Dataset

Types of dataset	Multimodal dataset
No. of participant	32
No. of EEG channel	32
Data collection method	Showing one-minute long excerpts of music videos
No. of used data collection resource	40 music videos
Sampling rate	128Hz
Rating values	Continuous scale 1-9
Rating scales	Arousal and Valence

The DEAP dataset is available in two parts, with the first part containing an online self-assessment of 14-16 volunteers based on arousal, valence, and dominance for 120 one-minute music videos. The second part contains the participant ratings, physiological recordings, and face video of an experiment where 32 volunteers watched 40 music videos which are the subset of the previously mentioned 120 music videos. Physiological signals and EEG signals were recorded where each participant also rated the videos following the above procedure. Facial expression at the time of watching videos was also recorded from 22 participants. Individual online self-assessment ratings, a list of the used YouTube music videos, the ratings that the participants gave for the videos, all the answers for the questionnaire of the participants before the experiment and participant's frontal face video recordings as well as raw physiological data recordings in BioSemi .bdf format are in the official dataset. Forty experiments for each of the 32 participants are found in the dataset. For each of the 40 experiments, the label array for each participant contained ratings of arousal, valence, dominance, and liking. For each participant, 8064 physiological / EEG signals data were collected with 40 different channels for each experiment and put into the data array. Among 40 channels 32 are EEG channels. Brain data is collected using electrode caps, EEG signal is collected via 512 Hz sampling frequency. After watching the videos all participants rated those according to a 1-9 scale based on valence, arousal and dominance. The duration of every sampled data is 63s. For experiments normally pre-processed data is used where 128Hz downsampling, electrooculogram (EOG) removal, filtering, segmentation and so on have been used. Two versions of preprocessed data are found on the official website. One of them is the data_preprocessed_mathlab folder processed with Matlab where files are in .mat format, and another is the data_preprocessed_python folder processed with Python (numpy) where files are in .dat format. The preprocessed data folder contains 32 files and each file holds the individual data of each of the 32 subjects. The data format is shown in Table III.

TABLE III. Data Orientation of Each Subject

Name of array	Shape of array	Contents of array
data	40 x 40 x 8064	video/trial x channel x data
labels	40 x 4	video/trial x label (valence, arousal, dominance, liking)

Both .dat files and .mat files contain data field with shape 40*40*8064 and label field with shape 40*4, where data field shape 40*40*8064 stands for 40 trials, 40 channels and 8064 refers to 63*128. Here sampling time is 63 seconds and the sampling frequency is 128Hz. In label field shape 40*4 indicates 40 experiments and 4 dimensions respectively for valence, arousal, dominance and liking. Signals were recorded according to the international 10-20 system. From the

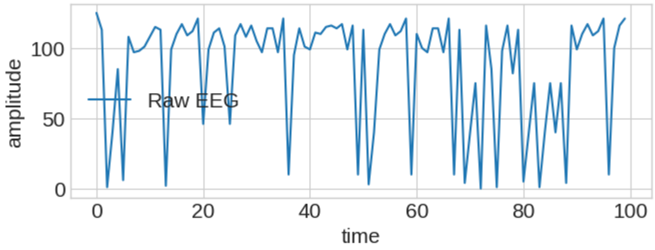


Figure 1. Raw EEG signal for one subject.

data_preprocessed_python folder, some parts of a file out of 32 preprocessed files are plotted and shown in Fig. 1.

IV. Methodology

The procedure for EEG Data Analysis is shown below through Fig. 2. Firstly, the dataset is collected then raw data is cleaned using the various pre-processing techniques. Next, this cleaned raw data is segmented in the feature extraction step. Then those extracted features are used to train a model for getting a better classification result.

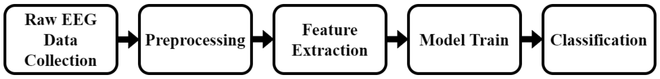


Figure 2. Workflow of EEG data analysis method.

A. Prepossessing

The DEAP EEG signal has been recorded with a good instrument and as a result, the chance of artifacts are minimized. The DEAP dataset's EEG signals are downsampled to 128 Hz first, so that the data content is collected in a good way between 0-48 Hz. Then the electromyogram (EMG) and electrooculogram (EOG) has been removed from the downsampled data. A bandpass filter has been applied to separate the delta waves from the analysis process. A blind source separation technique has been used removing eye artifacts. Using Common Average Reference (CAR) the data has been averaged. Each recording data has been partitioned into 60 seconds segments and a pre-trial baseline of 3 seconds has been removed.

B. Feature Extraction

In the research field, emotion classification potentiality depends on two factors: feature extraction and classification. Feature extraction reduces the initial dataset by identifying key features of data and later these features are used for classification. Distinguishing property, recognizable measurement, and functional components obtained from a section of a pattern are represented by features. A better classification accuracy comes if extracting features from a dataset are used instead of the original dataset. By Feature extraction, various advantages could be found like minimizing the loss of important signal, decreasing the risk of over-fitting, improving the visualization of data, and reducing the implementation complexity.

Three types of features were found [7]:

1) Time-domain features: used for statistical features.
2) Frequency-domain features: decomposition of preprocessed signal data into sub-bands.
3) Time-Frequency domain features: used for non-stationary waveform signals [16].

Frequency-domain features are used for this work. Different methods are used for EEG feature extraction, including FFT, Wavelet Transform (WT), Time-Frequency Distribution (TFD), Equivocator methods (EM), Auto-Regressive methods (ARM), etc. From these methods, FFT is ultimately applied to the preprocessed EEG datasets by using a python module named PyEEG, which is formatted in NumPy.

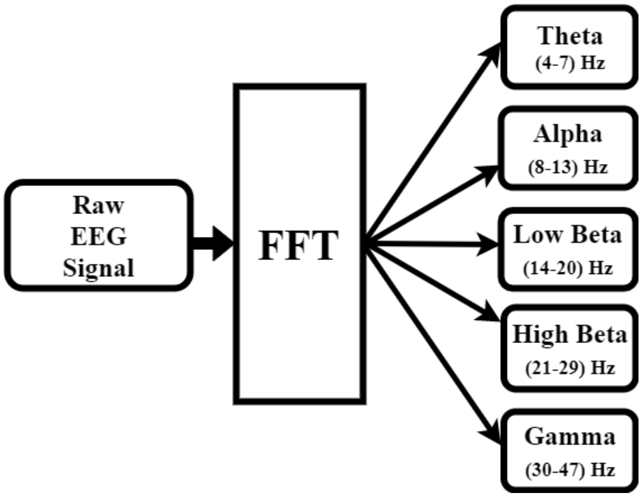


Figure 3. Feature extraction method.

The EEG signal is a complex and real wave which consists of different frequencies. At first, decompose the raw EEG signal into different sub-bands based on the frequencies like Alpha, High-beta, Low-beta, Theta, Gamma and this concept is exhibited in Fig. 3. Fourier analysis is commonly used for signal processing to convert time-domain signals into frequency domain signals. To compute the frequency components, the most popular algorithm is FFT, which computes the Discrete Fourier Transform (DFT) of a sequence [3].

$$X_k = \sum_{i=0}^{N-1} x_i(n) e^{\frac{-j2\pi ik}{N}} \tag{1}$$

In equation (1), $k = 0, 1, 2....N-1$ and X_k is the coefficient of discrete Fourier, length available data is N and $x_i(n)$ is the time domain input signal.

As the purpose is to extract key features from pre-processed data and use them to classify emotions through a CNN-based deep learning model. The Fast Fourier Transform method is used when employing mathematical tools to extract the EEG features. By using power spectral density (PSD) estimation, the characteristics of the EEG signal are computed. The EEG spectrum of wave characteristics is then divided into four frequency bands. Through the approximate auto-correlation sequence of the Fourier converter, PSD can be counted accurately.

When extracting features by using FFT for EEG, there are mainly two types of techniques: the Periodogram Method and Welch's method.

The most straightforward method is to use a periodogram to calculate PSD. Frequency decomposition is included by Periodogram, and the modulus squared of the Fourier transform of the signal is expressed as:

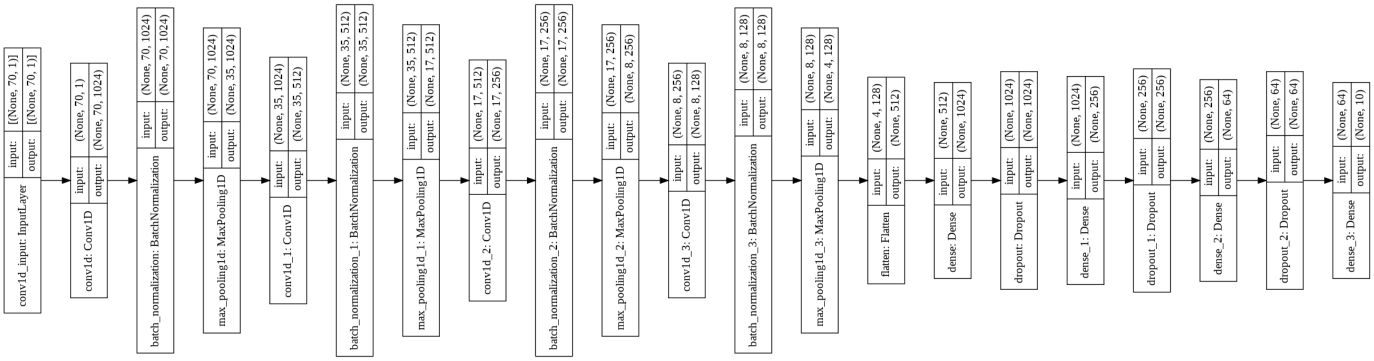


Figure 4. CNN model architecture.

$$\tilde{P}_{xx}(f) \;=\frac{\Delta t}{N}\left|\sum_{i=0}^{N-1} x_i(n)e^{\frac{-j2\pi ik}{N}}\right|^2 \tag{2}$$

In equation (2), Δt refers to space between the samples, available data length is *'N'*, $x_i(n)$ is the time domain input signal and $\tilde{\mathrm{P}}_{xx}(f)$ denotes PSD for $x_i(n)$.

Welch's method is another PSD assessment strategy that is used to improve the modified periodogram accuracy. This method is established through the use of signals in overlapping windows where for each window a periodogram is calculated and then to calculate the PSD those periodograms are constructed.

Supposing signals $x(n)$ have finite length, then the relationship with power spectral density is estimated as:

$$\begin{gathered} x_i(n) = x(n+iD), \quad n = 0,\ 1,\ 2...M-1 \\ While\ i = 0,\ 1,\ 2...L-1 \end{gathered} \tag{3}$$

In equation (3), *'iD'* represents the start of the i^{th} sequence and *'L'* is the length of the formatted data segment.

The subsequent outcome periodograms give:

$$\tilde{\tilde{P}}_{xx}(f) = \frac{1}{MU}\left|\sum_{n=0}^{N-1} x_i(n)\ w(n)\ e^{-j2\pi fn}\right|^2 \tag{4}$$

In equation (4), *'U'* represents the normalization factor of the power in the window function and is expressed such that

$$U = \frac{1}{M}\sum_{n=0}^{M-1} w^2(n) \tag{5}$$

In equation (5), $w(n)$ indicates the window function. The mean of these modified periodograms presents Welch's power spectrum which is considered as:

$$P_{xx}^{W} = \frac{1}{L}\sum_{i=0}^{L-1} \tilde{\tilde{P}}_{xx}^{(i)}(f) \tag{6}$$

After decomposition, the CNN filters are used to extract the deep features. Decomposition is required so that CNN filters can pick up more effective deep features.

C. *Labeling and Normalization*

The DEAP dataset contains raw data but after some preprocessing procedures, it becomes suitable to fed the classification model. To get the frequency domain of these data FFT has been applied. Then all the data are split into training and testing segments following the 7:1 ratio. Encoding is subsequently applied for labeling to avoid over-fitting. This dataset contains four label columns: arousal, valence, dominance and liking. But only arousal and valence labels are used for categorizing with the help of categorical function. Then normalization has been used to bring the different ranges of data between 0 and 1. Normalization can sometimes help to improve the accuracy of models. Standard Scalar is one of the techniques for normalization. Although the DEAP dataset contains 2D arrays at first, the utilized models require 3D data as input. That's why 2D data is converted into 3D using reshaping.

D. *CNN model structure*

Initially, all preprocessing parts are completed before a model is created, so that models can correctly learn various complicated features. A 1D-CNN model architecture is used, with a hidden layer that may be adjusted to improve accuracy. A segment is created with a 1D Convolution layer, a batch normalization layer as well as a 1D max-pooling layer and this segment has been found a total of four times at the starting of the model architecture. The fourth segment's result is converted into a 1D array with the help of flattening. Then three connected layers (dense layers) are applied and after every connected layer a drop-out layer is also given to avoid the overfitting problem. Then the output layer has been constructed including the number of classes and the softmax activation function. The entire model architecture has been demonstrated in Fig. 4.

V. Experiment and Result Analysis

To accomplish our experiment, many important parameters are used. Among them, there are numerous hyper-parameters also, and through adjusting their values the experiment's performance has been improved. The relevant hyper-parameters include window size, step size, sample rate, and batch size, among others. Here,

- **Window size:** The length of a cutout (sliding) of a time sequence of data is known as the window size.
- **Step size:** During the training period amounts of weights are needed to update, that is called step Size.
- **Sample rate:** From a non-digital or continuous signal to create a digital or discrete signal how many samples are taken per second, this number is known as the sampling rate.
- **Batch size:** The amount of training examples used in one iteration is referred to as a batch size.

Hyper-parameter optimization has been used with the following set of parameters listed in Table IV to achieve the best accuracy.

TABLE IV. Hyper Parameters

Hyper-parameters	**Hyper-parameter values**
Window size	32, 64, 128, 256, 512
Step size	8, 16, 32, 64
Sample rate	16, 32, 64, 128, 256, 512
Batch size	16, 32, 64, 128, 256, 512
Optimizer	Adam, SGD, RMSprop, Adadelta
Loss function	Categorical cross entropy, Sparse categorical cross entropy

We have determined that window size = 256, step size = 16, and sample rate = 128 are the optimum hyper-parameters for our experiment after doing hyper-parameter tuning.

The extracted data contains the main data as well as the emotion state label. This research strives to better identify emotions with two different approaches. One of them is the conventional binary classification method and another is our proposed 8-class classification method. However, the same 1D-CNN model architecture is employed in both classification approaches, which is designed by us. Total four models are used for the entire experiment. Two models for the binary classification and the other two models for the 8-class classification, here all the model works on arousal and valence emotional states. For both valence and arousal, the label array has floating values ranging from 1 to 9. But among all the values, the amount of 9 is very poor. For the convenience of reducing calculation and space-complexity, we converted the 9 into 8.99. In this case, the difference between 9 and 8.99 is very negligible and has no bearing on the outcome.

A. Binary-Class

In a two-dimensional emotion recognition system binary class classification is found as a conventional method. All the values of the label array are divided into two classes where 1-4.99 for one class and 5-8.99 for other classes. Then the binary arousal classification model and the binary valence classification model have been trained on all the data. Binary classification divides the entire emotion space into four classes and each of the four classes is the combination of multiple real-life known emotions. Four emotions are expressed by HaHv, LaHv, LaLv and HaLv. These four class ideas are demonstrated in Fig. 5.

The binary classifiers get 96.18%accuracy for arousal and 96.63% accuracy for valence. The binary arousal classifier achieved 99.65%training accuracy and 96.18% test accuracy in 131 epochs and at epoch 126, it provides the best accuracy. And the binary valence classifier provides 96.63% test accuracy when the training accuracy is 99.73% after 163 epochs but it shows the best test accuracy at 150th epoch. The confusion matrixes for the binary classification are shown in Fig. 6. For binary classification, all the train data are split following an 80:20 ratio and then five-fold cross-validation has been performed. All of the accuracies are similar, hence the average has been calculated. The classification reports are demonstrated in Table V, and Fig. 7 depicts the validation accuracy as well as validation loss graph for binary classification.

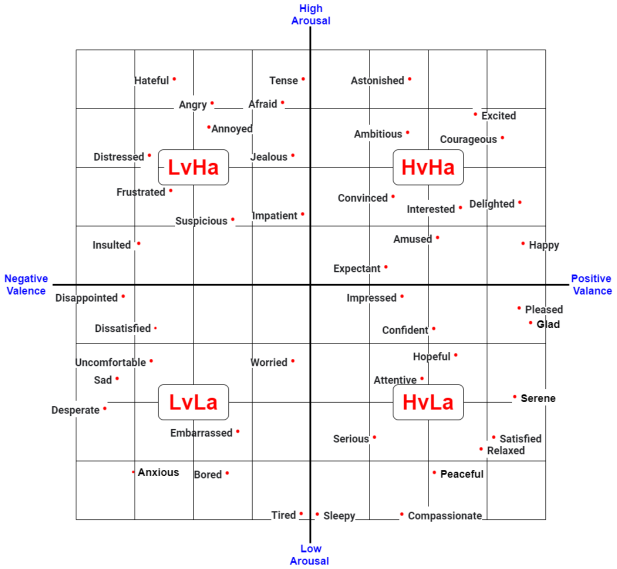


Figure 5. Circumplex model for emotions.

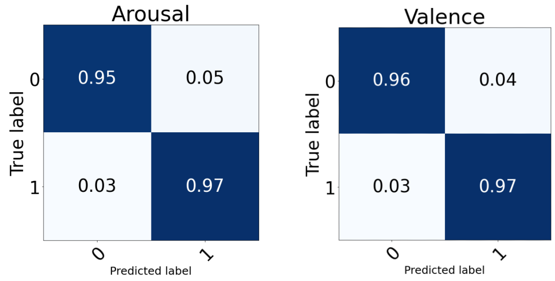


Figure 6. Binary-class classification confusion matrix.

TABLE V. Binary-Class Classification Report

Arousal				**Valence**			
Class	***Precision***	***Recall***	***F1 score***	***Class***	***Precision***	***Recall***	***F1 score***
0	0.96	0.95	0.96	0	0.96	0.96	0.96
1	0.96	0.97	0.97	1	0.97	0.97	0.97

B. 8-Class

This technique works precisely to recognize emotion properly using the DEAP datasets values. Many real-life emotions can be recognized using the 8-class classification technique where the binary class classification is able to find out only

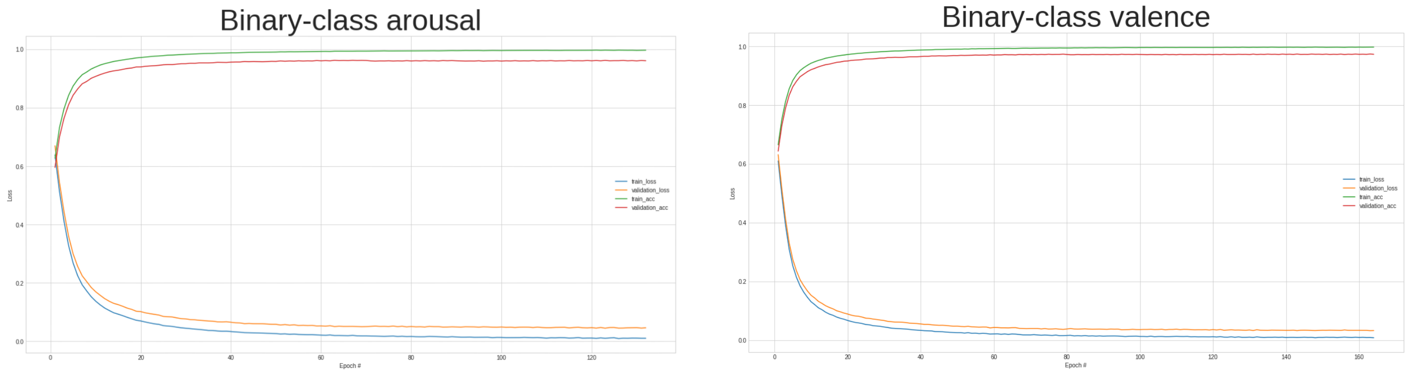


Figure 7. Validation accuracy and validation loss graph for binary-class.

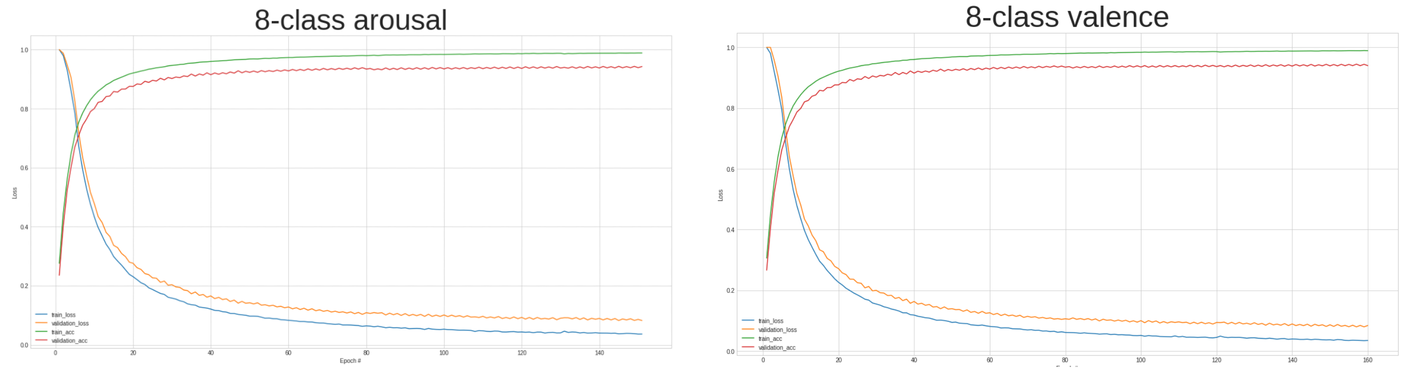
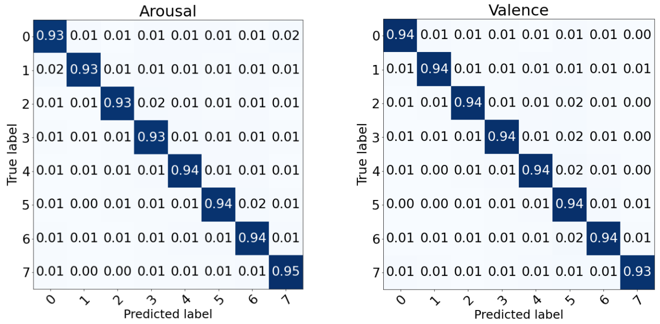


Figure 8. Validation accuracy and validation loss graph for 8-class.

four compound emotions, and each of these compound emotions consists of multiple real-life emotions. Using 8-class classifications those emotions are possible to recognize, a small number of them are mentioned in Fig. 5.


Arousal

True label \ Predicted label	0	1	2	3	4	5	6	7
0	0.93	0.01	0.01	0.01	0.01	0.01	0.01	0.02
1	0.02	0.93	0.01	0.01	0.01	0.01	0.01	0.01
2	0.01	0.01	0.93	0.02	0.01	0.01	0.01	0.01
3	0.01	0.01	0.01	0.93	0.01	0.01	0.01	0.01
4	0.01	0.01	0.01	0.01	0.94	0.01	0.01	0.01
5	0.01	0.00	0.01	0.01	0.01	0.94	0.02	0.01
6	0.01	0.01	0.01	0.01	0.01	0.01	0.94	0.01
7	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.95

Valence

True label \ Predicted label	0	1	2	3	4	5	6	7
0	0.94	0.01	0.01	0.01	0.01	0.01	0.01	0.00
1	0.01	0.94	0.01	0.01	0.01	0.01	0.01	0.01
2	0.01	0.01	0.94	0.01	0.01	0.02	0.01	0.00
3	0.01	0.01	0.01	0.94	0.01	0.02	0.01	0.00
4	0.01	0.00	0.01	0.01	0.94	0.02	0.01	0.00
5	0.00	0.00	0.01	0.01	0.01	0.94	0.01	0.01
6	0.01	0.01	0.01	0.01	0.01	0.02	0.94	0.01
7	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.93

Figure 9. 8-class classification confusion matrix.

To be successful in this classification technique, arousal and valence states label array values are needed to divide into eight segments. Eight label segmentation have been created maintaining 1 - 1.99, 2 - 2.99 , 3 - 3.99, 4 - 4.99, 5 - 5.99, 6 - 6.99 , 7 - 7.99 ,8 - 8.99 procedure.Then the 8-class arousal classification model, and the 8-class valence classification model have been applied to all the data for training.

The 8-class arousal classifier gets 98.83% training accuracy after running 151 epochs and gets 93.79% best test accuracy at 126th epoch. The 8-class valence classifier shows the best test accuracy of 93.83% at 120th epoch where it takes 160 epochs to get 98.92% train accuracy. Fig. 9 depicts the confusion matrixes for the 8-class classification. Five-fold cross-validation has been applied in 8-class classification, and the entire train dataset is split into an 80:20 ratio. All the accuracies are similar, and the average accuracy has been taken. Table VI depicts the classification reports for 8-class classification. The validation accuracy and validation loss graphs are displayed in Fig. 8.

The best results from all of our experiments are summarized in Table VII.

TABLE VI. 8-CLASS CLASSIFICATION REPORT

Arousal				**Valence**			
Class	***Precision***	***Recall***	***F1 score***	***Class***	***Precision***	***Recall***	***F1 score***
0	0.93	0.92	0.92	0	0.94	0.94	0.94
1	0.95	0.93	0.94	1	0.94	0.93	0.94
2	0.95	0.93	0.94	2	0.95	0.94	0.94
3	0.95	0.93	0.94	3	0.94	0.94	0.94
4	0.94	0.94	0.94	4	0.95	0.93	0.94
5	0.94	0.94	0.94	5	0.94	0.94	0.94
6	0.95	0.94	0.94	6	0.94	0.94	0.94
7	0.94	0.95	0.94	7	0.94	0.93	0.93

TABLE VII. RESULT SUMMARY

Type	**Arousal**	**Valence**
Binary-class	96.18%	96.63%
8-Class	93.79%	93.83%

TABLE VIII. Result Comparison For Binary-Class

No.	Modeling technique	No. of class	Accuracy
1	CNN [15]	2-Class	91.5%
2	LSTM [3]	2-Class	89.83%
3	CDCN [9]	2-Class	92.58%
4	KNN, ResNets [16]	2-Class	Valence:90.39% Arousal:89.06%
5	**Our Model: 1D-CNN**	**2-Class**	**Valence:96.63%** **Arousal:96.18%**

Using the DEAP dataset those model architectures are on the top list for better emotion recognition accuracy, the best accuracy models from each of them are demonstrated in Table VIII . Yucel Cimtay Et al. [15] have worked with raw data for binary classification , and get the highest accuracy among the reviewed CNN models with an accuracy of 91.5% . Many researchers have worked with the famous LSTM model for binary classification , and also got good results. According to the literature review, Divya Acharya Et al. [3] have got 89.83% accuracy , and this is the best accuracy for LSTM. To get this best LSTM binary classification accuracy have used FFT as feature extraction. Analyzing the results of our reviewed papers, CDCN is also found in the top category for accuracy of binary class emotion recognition and Zhongke Gao Et al. [9] have worked with this model. They have got 92.58% accuracy where DE is the extraction method. Ningjie Liu Et al. [16] have applied their KNN, ResNets on DEAP dataset to get valence: 90.39% and arousal: 89.06% accuracy. But in the same field using FFT with the help of two 1D-CNN models we got 96.63% accuracy for valence and 96.18% accuracy for arousal which is the highest compared to all other models of Table VIII.

VI. Conclusion and Future Work

This study investigates the efficiency of the convolution neural network based on previous research in the field of emotion recognition. Our proposed CNN models are more effective for emotion recognition and outperform previous research in terms of accuracy. These are able to effectively classify preprocessed EEG data. For arousal-valence binary classification accuracy exceeds the same benchmark activities that shows a noticeable difference, and introduces a much more precise 8-class classification approach which provides a satisfactory result also.

For future work, we would like to work with our methodology on real-time data so that the emotions of mentally challenged and autistic people can be expressed easily. We will also focus on how to make our recognizing models more efficient and portable.

References

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

[2] D. Acharya, S. Goel, R. Asthana, and A. Bhardwaj, ”Anovel fitness function in genetic programming to handle unbalanced emotion recognition data,” Pattern Recognition Letters (PRL), vol. 133, pp. 272-279, 2020.

[3] D. Acharya, S. Goel, H. Bhardwaj, A. Sakalle and A. Bhardwaj, ”A Long Short Term Memory Deep Learning Network for the Classification of Negative Emotions Using EEG Signals,” 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-8.

[4] R. Sharma, R. B. Pachori, and P. Sircar, ”Automatic Emotion recognition based on higher order statistics and deep learn-ing algorithm,” Biomedical Signal Processing and Control (BSPC), vol. 58, pp. 101867, 2020.

[5] F. Wang, S. Wu, W. Zhang, Z. Xu, Y. Zhang, C. Wu, and S. Coleman, ”Emotion recognition with convolutional neural network and EEG-based EFDMs,” Neuropsychologia, vol. 146, pp. 107506, 2020.

[6] R. Hassan, S. Hasan, M. J. Hasan, M. R. Jamader, D. Eisenberg and T. Pias, ”Human Attention Recognition with Machine Learning from Brain-EEG Signals,” 2020 IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), 2020, pp. 16-19.

[7] Anubhav, D. Nath, M. Singh, D. Sethia, D. Kalra and S. Indu, ”An Efficient Approach to EEG-Based Emotion Recognition using LSTM Network,” 2020 16th IEEE International Colloquium on Signal Processing Its Applications (CSPA), 2020, pp. 88-92.

[8] K. H. Cheah, H. Nisar, V. V. Yap and C. Lee, ”Short-time-span EEG-based personalized emotion recognition with deep convolutional neural network,” 2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2019, pp. 78-83.

[9] Z. Gao, X. Wang, Y. Yang, Y. Li, K. Ma and G. Chen, ”A Channel-fused Dense Convolutional Network for EEG-based Emotion Recognition,” in IEEE Transactions on Cognitive and Developmental Systems.

[10] Y. Luo et al., ”EEG-Based Emotion Classification Using Spiking Neural Networks,” in IEEE Access, vol. 8, pp. 46007-46016, 2020.

[11] Y. Cimtay, and E. Ekmekcioglu, ”Investigating the use of pre-trained convolutional neural network on cross-subject and cross-dataset EEG emotion recognition,” Sensors, vol. 20, no. 7, pp. 2034, 2020.

[12] G. Cao, Y. Ma, X. Meng, Y. Gao and M. Meng, ”Emotion Recognition Based On CNN,” 2019 Chinese Control Conference (CCC), 2019, pp. 8627-8630.

[13] E. A. Abdel-Ghaffar and M. Daoudi, ”Emotion Recognition from Multidimensional Electroencephalographic Signals on the Manifold of Symmetric Positive Definite Matrices,” 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2020, pp. 354-359.

[14] X. Zhong, Z. Yin and J. Zhang, ”Cross-Subject emotion recognition from EEG using Convolutional Neural Networks,” 2020 39th Chinese Control Conference (CCC), 2020, pp. 7516-7521.

[15] Y. Cimtay, E. Ekmekcioglu and S. Caglar-Ozhan, ”Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion,” in IEEE Access, vol. 8, pp. 168865-168878.

[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), 2018, pp. 896-900.

[17] 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.

[18] H.Y. Joong, and S.Y. Chung, ”EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm,” Computers in biology and medicine (CBM), vol. 43, no. 12, pp. 2230-2237, 2013.

[19] 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, 2013, pp. 1286-1290.

[20] X. Zhang, B. Hu, J. Chen, and P. Moore, ”Ontology-based context modeling for emotion recognition in an intelligent web,” World Wide Web 16 (WWW), 2013, pp. 497-513.

[21] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou and B. Hu, ”Emotion recognition from multi-channel EEG data through Convolutional Recurrent Neural Network,” 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016, pp. 352-359.

[22] A. Al-Nafjan, M. Hosny, A. Al-Wabil, and Y. Al-Ohali, ”Classification of human emotions from electroencephalogram (EEG) signal using deep neural network,” Int. J. Adv. Comput. Sci. Appl, vol. 9, no.8, pp. 419-425, 2017.

[23] W. Liu, W. L. Zheng, and B. L. Lu, ”Emotion recognition using multimodal deep learning,” in International conference on neural information processing (ICNIP), 2016, pp. 521-529.

[24] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya, ”Using Deep and Convolutional Neural Networks for Accurate Emotion

Classification on DEAP Dataset," in Twenty-ninth IAAI conference, 2017.

[25] W. Zheng, J. Zhu and B. Lu, "Identifying Stable Patterns over Time for Emotion Recognition from EEG," in IEEE Transactions on Affective Computing, vol. 10, no. 3, pp. 417-429, 1 July-Sept. 2019.

VII. Review

A. Reviewer #1

Comments to Authors: Other works have applied CNNs to the problem of emotion recognition from EEGs. Where does this sit in the literature? This is unclear as many important works are not cited. E.g. Tripathi, Samarth, et al. "Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset." Twenty-ninth IAAI conference. 2017.
However, it seems the focus of the current work is on using manual feature extraction and applying CNN afterward.

Answer: There have already been many research works on the DEAP dataset where the researcher used different types of pre-processing techniques with various machine learning algorithms. However, we found a gap in the literature. Most of the researchers used binary classification which can classify only 4 emotions. Humans, on the other hand, have much more complex emotions. So, we proposed an 8 class classification method by which 64 different emotions can be classified with great accuracy. In addition, Divya Acharya Et al. [3] used FFT and LSTM and Hyun Joong Yoon Et al. [18] used FFT and the Bayes classifier to build their models. So, we used FFT and CNN to build our models and this combination of approaches has not been done previously. So, our research is unique in two ways. One for classifying 64 emotions instead of 4 emotions and using a new combination of pre-processing technique and machine learning model.
We tried to highlight most of the important and relevant publications on emotion recognition related to the DEAP dataset in the Literature Review section (Section II) of our paper. While selecting publications for literature review we gave more priority to the most important and most recent publications. Being an important publication in this field, the mentioned paper was already added in our first submission. That publication is in reference 24 and also analyzed in the literature review in row 13 of Table I . However, as per the reviewer's recommendation, we have made another search for relevant publications but we couldn't find any to cite. So, we think our literature review is rigorous and complete.

B. Reviewer #2

Comments to Authors: In this work, the authors have proposed the 8-class emotion classification method to recognize emotion accurately, They have summarized the literature review in Table I in a very compact form. Their methodologies are well explained and the results show satisfactory over the other researchers' works as in Table V.
However, they need to correct their English language accurately in their manuscripts. The abstract of the paper is too lengthy. Only the works related to this research should be summarized here. They need to maintain the IEEE format while writing references.

- **Correction 1**: However, they need to correct their English language accurately in their manuscripts.
  Answer: Thank you so much for the suggestion. We have made necessary corrections to make the manuscript more sound.
- **Correction 2**: The abstract of the paper is too lengthy. Only the works related to this research should be summarized here.
  Answer: We have shortened the abstract by summarizing only the works related to this research.
- **Correction 2**: They need to maintain the IEEE format while writing references.
  Answer: All references have been corrected by maintaining the IEEE format.

C. Reviewer #3

Comments to Authors: Some comments to improve the paper:

- **Correction 1**: According to the workflow presented in Fig. 2 , after the feature extraction, a classifier should be used to classify the emotions. you have extracted the features from the FFT signals. My question is why you are using the CNN model to extract the features again. We know that CNN is used to extract automatic features rather than using a features descriptor.
  Answer: The EEG is a complex and real wave which consists of different frequencies. At first, we wanted to decompose the EEG signal into different sub-bands based on the frequency. So, we used FFT to decompose the complex EEG signal into Alpha, High beta, Low beta, Theta, Gamma. Then after decomposition we used CNN filters to extract the deep features. In other words, the features are being extracted into two phases. The first phase is required so that CNN filters can pick up more effective deep features. Previously, we used only 1D-CNN (without FFT) directly on the EEG signal but it didn't achieve good accuracy. And we added clarification about this in Subsection B of Section IV
- **Correction 2**: The abstract is not focused and too long. It should be more precise.
  Answer: Thank you for the suggestion. The abstract of our paper has been made more precise and concise.
- **Correction 3**: You should present the class wise precise and recall.
  Answer: The class-wise precision and recall are added in Table V and Table VI.
- **Correction 4**: In conclusion, you claim that your model is simple and lightweight . It needs justification why the model is lightweight.
  Answer: We have checked the reference papers and found out that our 1D-CNN model lighter than some 3D-CNN from the literature review. However, our model is not the lightest. So we are not claiming that our model is

lightweight now and we have removed the sentences from our paper.

- **Correction 5**: You should include the validation loss and accuracy curves.

  Answer: We have included the validation loss and accuracy curves in Fig. 7 and Fig. 8.
- **Correction 6**: You should use cross-validation to justify the accuracy.

  Answer: For binary-class and 8-class classification, five-fold cross-validation is applied and the entire train dataset is split following an 80:20 ratio. All the accuracies are similar and the average accuracy has been taken. We have added about this information in Subsections A and Subsections B of Section V.

  Answer: What classifier did you use?

  Answer: 1D CNN classifier which consists of CNN filters for extracting deep features and artificial neural network (dense layer) for classification. And this classifier architecture was demonstrated in Fig. 4 and described in Subsection D of Section IV.
- **Correction 8**: In the title you have used ”Fine grained”. Why are you calling your recognizer ”fine gained”? You did not do any kind of hyper-parameters optimization.

  Answer: Most of the previous work can classify all the emotions into 4 classes but our model classifies the same emotions into 64 classes. As a result, emotions can be recognized more precisely by our model. That’s why in the title we have used “Fine-Grained”. So, the fine-grained is related to the number of emotions that our model can classify and not related to the hyper-parameter optimization.

  But we have used proper hyper-parameter optimization to achieve the highest accuracy. We have demonstrated this with the help of Table IV of Section V.
- **Correction 9**: The figures should be clearly understandable.

  Answer: We have added high resolution graphs and pictures and added more illustrative descriptions of the figures so that the figures can be easily and clearly understandable.
- **Correction 10**: The writing should be improved. There are lots of grammar issues.

  Answer: We have checked and corrected all the grammatical issues and try to improve the manuscript.