IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 March 2023, accepted 6 April 2023, date of publication 13 April 2023, date of current version 26 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3266804

RESEARCH ARTICLE

Emotion Recognition Using Temporally Localized Emotional Events in EEG With Naturalistic Context: DENS# Dataset

, (Graduate Student Member, IEEE), ,
, AND , (Senior Member, IEEE)
Indian Institute of Information Technology Allahabad, Allahabad, Uttar Pradesh 211012, India
Corresponding authors: (rs163@iiita.ac.in), (pse2017001@iiita.ac.in), and (ust@iiita.ac.in)

This work was supported by the Ministry of Education, Government of India, funded by the acquisition of the EEG system.

ABSTRACT Emotion recognition using EEG signals is an emerging area of research due to its broad applicability in Brain-Computer Interfaces. Emotional feelings are hard to stimulate in the lab. Emotions don't last long, yet they need enough context to be perceived and felt. However, most EEG-related emotion databases either suffer from emotionally irrelevant details (due to prolonged duration stimulus) or have minimal context, which may not elicit enough emotion. We tried to overcome this problem by designing an experiment in which participants were free to report their emotional feelings while watching the emotional stimulus. We called these reported emotional feelings ''Emotional Events'' in our Dataset on Emotion with Naturalistic Stimuli (DENS), which has the recorded EEG signals during the emotional events. To compare our dataset, we classify emotional events on different combinations of Valence(V) and Arousal(A) dimensions and compared the results with benchmark datasets of DEAP and SEED. Short-Time Fourier Transform (STFT) is used for feature extraction and in the classification model consisting of CNN-LSTM hybrid layers. We achieved significantly higher accuracy with our data compared to DEAP and SEED data. We conclude that having precise information about emotional feelings improves the classification accuracy compared to long-duration recorded EEG signals which might be contaminated by mind-wandering. This dataset can be used for detailed analysis of specific experienced emotions and related brain dynamics.

INDEX TERMS Affective computing, CNN, DEAP, DENS, EEG, emotion dataset, emotion recognition, LSTM, SEED.

I. INTRODUCTION

Emotion recognition has been a challenging task in artificial intelligence. Several methods are available for measuring the participants' emotions. These methods include behavioural changes, subjective experiences self-reported by the participants, peripheral and central nervous system measures, etc [1]. Brain activities are among the most robust dimensions of detecting human affect, as it is difficult for the users to manipulate innate brain activity during the process. Accordingly, Electroencephalography (EEG) is considered a suitable and convenient method to record electrical activities to measure brain activities as it is a non-invasive method, i.e. there are no scalpel incisions.

Many studies have already been conducted to measure human affect with the help of EEG and other peripheral responses [2], [3], [4], [5]. In the previous studies, the focus of the study was to develop a database that is labelled and suitable for emotion detection by intelligent systems and has contributed to affective computing. There is a typical method in these studies to elicit emotion in the participants by presenting them with video clips as stimuli. In the process of emotion recognition and other classification tasks, all the EEG data for that stimulus are

The associate editor coordinating the review of this manuscript and approving it for publication was .

to be considered for the classification model, as there is no information about the precise temporal location at which a participant may experience the emotion. Models must consider all the data presented for that label, which is unnecessarily computationally expensive and decreases the system's efficiency by feeding not-so-essential data in the input.

In our approach, we have presented a novel method to overcome this issue by providing precise information about the emotion elicitation, self-reported by the participants. We call it an 'Emotional Event'. In this method, an additional task is given to the participants to mention precise temporal information by clicking on their computer screens while watching the emotional clips if they feel some emotion. Also, to the best of our knowledge, there are no EEG affective datasets available for the Indian subcontinent population. Hence we tried to reduce this research gap in our work. We have considered DEAP dataset [2] and SEED dataset [3] for comparison. We tried to follow a format similar to the benchmark datasets and compared our dataset's results with these datasets based on statistical significance.

EEG measures the electrical signals from the scalp with temporal details. Different EEG devices vary with the number of channels of EEG. Thirty-two or fewer EEG channels are especially notable in affective computing research [6]. A few studies are also available with up to 64 electrodes. In this work, we used a 128-channel EEG device to detect emotions. This EEG cap follows the International 10-10 system's standards [7].

Emotions are complex and challenging to understand as many theories exist about emotions, and there is a lack of a single consensus theory [8]. The study of emotions has been an emerging topic that combines multi-disciplines such as psychology, neuroscience, computer science and medicine, etc. There are different aspects involved in determining emotions, such as behavioural, psychological and physiological aspects, cognitive appraisals, facial expressions, vocal responses, subjective experiences, etc. This study focuses on physiological aspects of emotion, which are considered into account by the brain signals captured through EEG while watching emotional video clips. Further, this study tries to collect a comprehensive list of subjective experiences through a self-assessment rating at the end of each clip.

Many approaches could be used to assess the participants' emotional states. Earlier, some basic emotions were used that are universally recognised for study purposes [9]. Later, some theories explained some complex emotions that are a combination of basic emotions [10]. Multi-dimensional theories of emotions are the widely accepted theories for assessing core affect [11], [12]. According to these theories, emotions are considered a multi-dimensional array; one dimension is for valence (experiencing positive or negative) and the other for arousal (experiencing the intensity) or dominance (controlling or feeling controlled). A few more dimensions are also considered, that make the spectrum broader, e.g., relevance (how much the stimulus is relevant to the participant's emotional feelings), familiarity (how much the participant is familiar with the stimulus) and liking (how much the participant liked or disliked the stimulus). Asking participants to report these experiences on a continuous scale is common in similar studies. Some theories deal with the physiological responses of feeling emotions, e.g., body temperature and heartbeat change [13], [14]. It is obvious from the theories that emotion is not a one step process; instead, it is a combination of physiological responses and other information. Evidence shows that many brain regions are involved during emotion perception [15]. We have also collected ECG and EMG data of the participants along with EEG to consider these parameters.

Emotion recognition through EEG data follows a similar pattern as used in various EEG signal analyses. First, the data is acquired, and some preprocessing is applied to the signal. These preprocessing steps involve removing artefacts such as ocular activity, muscle activity, and powerline interference. Also, downsampling of the signal and band-pass filtering are used to make data more useful. Various dimensionality reduction techniques, such as ICA and PCA, are also used to prune the data to make it feature-rich. After preprocessing, features are extracted from the signal to feed into the model for the classification task. Different kinds of features are extracted such as time-domain (e.g., event-related potential (ERP), high-order crossing (HOC), etc.), frequency-domain (e.g., power spectral density (PSD), etc.); and time-frequency domain (e.g., STFT, wavelet analysis, etc.) features.

EEG records multi-frequency non-stationary brain signals from various electrodes. Analyzing these signals is challenging because of the complex and irregular nature of EEG signals. The time-frequency domain analysis has the benefits of both the time and frequency domains, e.g., better spatial and temporal information from EEG signals. One basic time-frequency domain feature extraction method is Short-Time Fourier Transform (STFT). STFT is a time-ordered sequence of spectral estimates and is one of the powerful and general-purpose signal processing techniques. It has been used in the field of spectral analysis of a signal. The STFT is used to compute spectrograms which are used extensively for signal processing. Spectrograms are visual representations of the spectrum of frequencies of a signal with varying times [16].

CNN is the most frequently used architecture for EEG analysis and classification tasks, and DBN and RNN follow it [17]. Hence we have used a combination of the CNN and LSTM model. It also helped to compare our dataset with the benchmark datasets in terms of maximum classification accuracy. Using artificial intelligence for affective computing provides better learning capabilities to intelligent systems. With the advancement of computing power and the development of effective and advanced neural network research, the trend of using various machine learning and deep learning

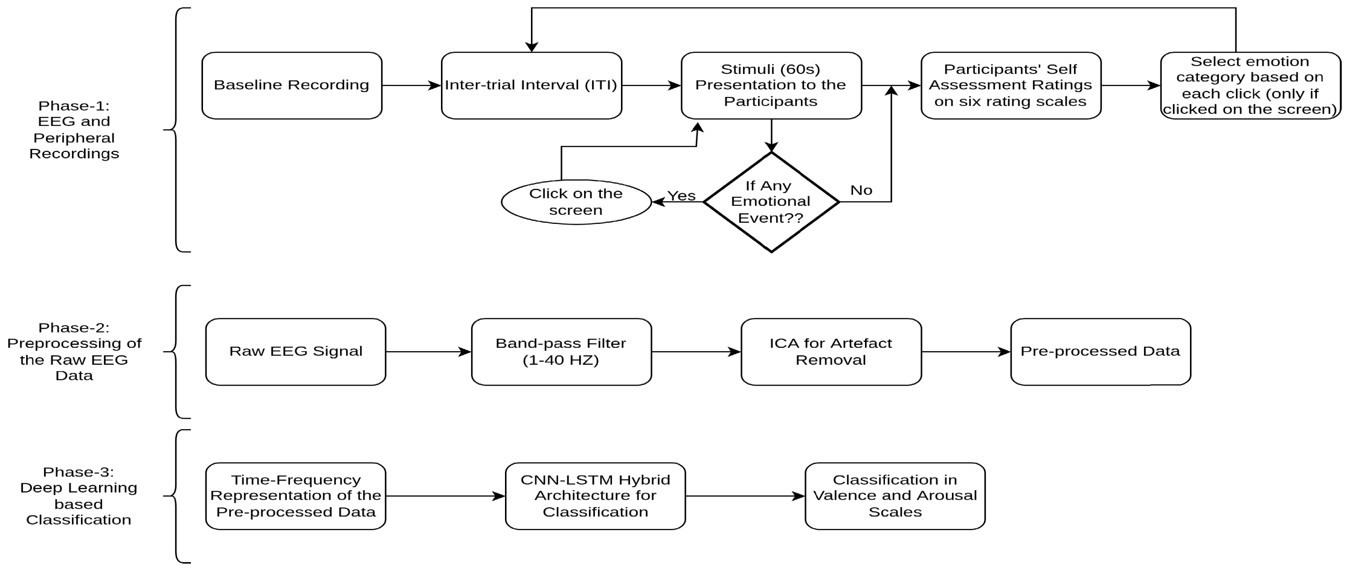


FIGURE 1. **Complete Flowgram of the Experiment.**

techniques has grown within the last few years [18]. This work employs the widely used state-of-the-art deep learning methods to detect emotions from EEG signals.

In this work, we contribute to the affective computing research by emphasising the importance of considering the duration of the signal encoding information about emotional experience. Emotion duration is the essential component of emotion dynamics [19], which is ignored in other datasets. We take account of emotion duration, which, to the best of our knowledge, had never been considered before. By comparing with other datasets using the same stimulus modality, we show that better emotion recognition accuracy can be achieved if the temporal information is incorporated.

This paper is organized into six sections. In the introduction section, we introduced the ongoing trends in affective computing, EEG emotion analysis and our dataset. In the next section, we introduced our proposed dataset- DENS, Emotional Events, experimental details (e.g., stimuli, EEG recordings, ratings etc.), preprocessing of the EEG data, its salience features and other datasets used (DEAP and SEED). In the methodology section, we discussed the feature extractions, input preprocessing of the extracted features for the classifier and deep learning model architecture for the same. Next, we have the results section, discussing the comparison results of the DENS-DEAP and DENS-SEED data based on several parameters and also comparing our results with recent studies. After that, we have a discussion section discussing the results and future aspects. At last, we concluded our analysis in the conclusion section.

II. DATASET ON EMOTION WITH NATURALISTIC STIMULI (DENS)

The complete flow diagram of our experiment is given in Fig. 1. We call our dataset 'Dataset on Emotion with Naturalistic Stimuli' (abbreviated as DENS) [20].

A. EMOTIONAL EVENT

Emotion is a complex phenomenon which is embedded within a context [21]. Moreover, emotion is transient in nature and is not available throughout the stimulus duration. In fact, more than one aspect could be embedded within the stimulus context, and different participants can feel emotion at different points of time considering various aspects. However, most of the datasets recorded to date [2], [3] ignore the transient nature of emotions and provide a single emotional category for the whole stimulus duration. Although the stimulus has emotional information, it has some non-emotional aspects too, which could lead to mind-wandering activity. Although there are some attempts to get continuous subjective feedback on emotional experience and neural activity, the experimental method involved multiple watching of the stimulus and retrospective collection of emotional experience [22], [23], [24], [25], [26]. The retrospective collection depends on autobiographical memory and can raise biases across subjects depending on their capability to recall [27]. Also, repetitive viewing effects can bias the ratings and underlying neural effects [27]. Hence, an experimental paradigm is needed to record the participants' feedback dynamically, with minimal distraction during emotion processing and minimizing the memory recall biases. In this work, we are introducing a novel paradigm in which the time-stamp of emotional feelings can be marked online that can be further utilized to get the subjective feedback of emotional feelings and analyze brain signals temporally localized to the feeling of an emotion. We refer to these time-stamped emotional feelings as "emotional events".

B. EXPERIMENTAL DETAILS

1) STIMULI

The selection of stimuli to induce participants' emotions also plays a vital role in emotion recognition. A careful selection

TABLE 1. Selected stimuli for EEG study from the stimuli dataset we created [28]). The time duration of each stimulus is 60s. Stimulus Ids are given for references available in the open science framework repository.

Stimulus Name	Stimulus ID	Target Emotion
Ashayen	199	Adventorous
Horror	214	Afraid
Anacondas The Hunt For The Blood Orchid Clip	10	Alarmed
Lage Raho Munnabhai Only The Funny Scenes	98	Amused
Anger Legend of Baghat Singh	198	Angry
Divergent Kiss Scene Clip	54	Aroused
Butterfly Nets	40	Calm
Best Horror Kills Ghost Ship Opening Scene	26	Disgust
Jai Ho	92	Enthusiastic
Sadda Haq	152	Excited
Masoom	109	Happy
Cheerful Rang	201	Joyous
Crash Saddest Scene	51	Melancholic
Hate LBS	210	Miserable
Madari Movie Of Best Scene	113	Sad
Final Race of Milkha Singh Career	67	Triumphant

of stimuli is critical, and for that, technical validation of the video clips is crucial to assess if the intended emotional experience is elicited by the stimuli. We have used naturalistic stimuli to elicit emotions in the participants. Naturalistic stimuli are dynamic emotional scenes in which multi-sensory perception is applied. It resembles more to the real-life scenario as compared to static and simple stimuli. In our previous work, we have validated a set of multimedia stimuli and created an affective stimuli database [28]. We selected 16 emotional stimuli from this database to perform our EEG experiment. The selection criteria for these 16 emotional stimuli are based on three factors:

1) A high probability of eliciting target emotions (calculated on the basis of ratings available).
2) Few stimuli must be available for each emotion category.
3) Since this experiment was done on the Indian population, more emphasis was given to Indian clips.

Besides these 16 emotional stimuli, we have validated 2 non-emotional stimuli separately. These clips were rated around 5 mean valence and arousal values (on a scale of 1 to 9). These non-emotional clips included the world's longest road routes or animated history of the Babylonian era, which may not contribute to eliciting emotions. The inclusion of non-emotional stimuli was to validate the participants' responses and avoid the long accumulation of the affects during the experiment.

For each participant, nine (9) emotional stimuli were selected randomly from the 16 selected emotional stimuli and two (2) non-emotional stimuli. Each stimulus was of 60 seconds.

Table 1 shows the list of 16 emotional stimuli with the target emotions assigned during the stimuli validation.

2) EEG RECORDING

We recorded the EEG activity of forty participants (23.3 ± 1.25, F=3) while they were watching emotional film stimuli. Following are some critical pieces of information regarding the experiment:

- Each participant saw nine (9) emotional stimuli randomly extracted from the set of 16 emotional stimuli and two (2) non-emotional stimuli as described in the previous subsection.
- While watching the emotional film stimuli, participants were instructed to perform a mouse click the moment they felt any emotion. We call it an Emotional Event.
- At the end of each video stimuli, participants are provided six self-assessment scales, including valence, arousal, dominance, liking, familiarity, and relevance.
- For each click, participants were supposed to select one emotion from the provided list of emotions pooled into four quadrants of V-A space (HVHA, LVHA, LVLA, HVLA) (abbreviations- V: Valence, A: Arousal, H: High, L: Low) in the drop-down menu. Participants were also given a choice to enter the emotional category which suits their emotional experience but is unavailable in the provided emotion list. For more details see Fig.2.

Before the main experiment begins, participants go through the training phase. In the training phase, participants were given instructions about the experiment procedure, rating scales were properly explained by giving them a small quiz, and also they were trained to mouse-click when they felt emotion during the stimulus.

The main experiment consists of the following steps for each participant:

1) Baseline Recording: EEG signal was recorded for 80 seconds while the participant looked at the cross-mark on the screen and performed no task.
2) After baseline recording, one stimulus of 60s was presented to the participant. Participants were told to click on the screen when they felt the emotion during the stimulus. Participants may click more than once if they felt so but were instructed to refrain from multiple clicks for the same emotion. EEG signals were recorded during this phase.
3) After the stimulus ends, participants go through self-assessment ratings of valence, arousal, dominance, liking, familiarity, and relevance. These scales are explained in detail in the next subsection.
4) At last, participants were supposed to select one emotion category for each click (emotional event). To help the participants in recalling about the click, they were presented with three frames around the click.
5) After this, an inter-stimulus interval comes with no time limit. During this interval, participants were given a quick and easy mathematical calculation (e.g., 2+5*2=?). It helps participants to flush their previous emotional state.
6) After that, the next stimulus is presented to the participant, and steps 1 to 5 are followed similarly for each stimulus. A total of 11 stimuli (9 emotional and 2 non-emotional) were presented to each participant.

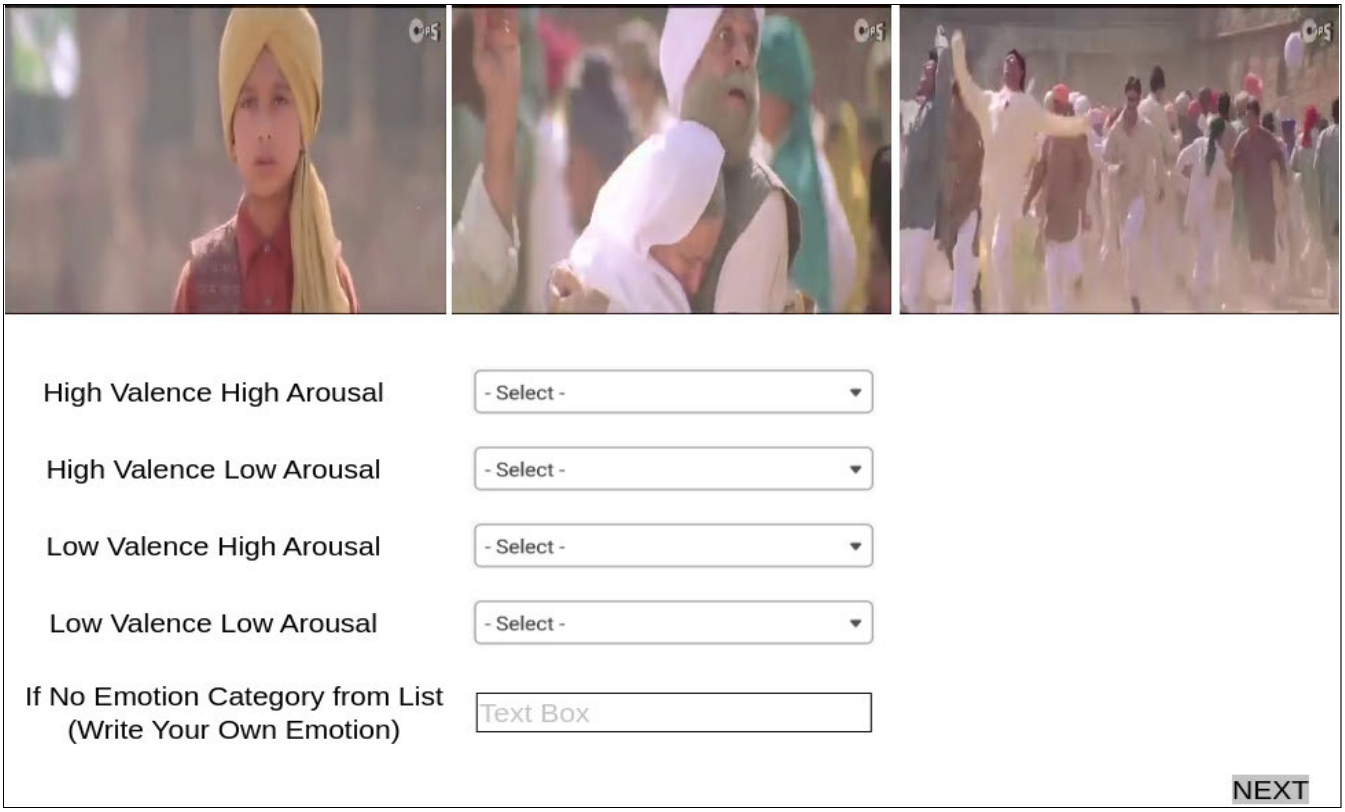


FIGURE 2. Emotion Category Selection Screen for Emotional Event (Click): After the participants rated all the six rating scales of Valence, Arousal, Dominance, Liking, Familiarity and Relevance, they are shown this screen for emotion category selection. On this screen, three image frames were shown. The middle one belongs to the time of the click; the left one is 20 frames earlier, and the right one is 20 frames after the click (Please note that the stimulus clips were shown in 30 frames per second). It helps participants to recall easily. They only have to select one emotion category. If the experienced emotion is not present in the list, they were free to write their own.

3) RATINGS

Subjective ratings are one of the well-known methods to evaluate the personal emotional experience of the participants. Emotional pictures/videos or audio clips are presented to the participants, and they are asked to rate these clips on different scales based on their personal experiences. These scales include Valence, Arousal, Dominance, Liking, Familiarity and Relevance. The rating scales range from 1 to 9 for Valence, Arousal and Dominance. For Liking, familiarity and Relevance, it ranges from 1 to 5. Although, in this analysis, we considered only valence and arousal scales.

4) SUMMARY OF THE EEG SIGNALS

As explained above, 465 emotional events were extracted from the forty participants in this experiment. All the participants clicked at least one time (average **1.29** times) during the stimulus.

Although for each participant and each stimulus, EEG recording is available for the whole stimulus (i.e., for the 60s), we have considered the signal for 7 seconds duration (1 second before the click and 6 seconds after the click) for each emotional event. We have tested for other time durations (e.g., 8s, 9s, up to 10s) but found better results with 7s duration. The recording has a sampling rate of 250 Hz.

C. PREPROCESSING AND ARTIFACT REMOVAL OF THE EEG DATA

The procedure followed to perform the preprocessing is described elsewhere [29]. The critical step which should be described here includes filtering and artifact removal. We had 128-channel EEG raw data with a sampling rate of 250 Hz. The raw signal is filtered using a Butterworth fifth-order bandpass filter with the passband 1-40 Hz. Independent component analysis (ICA) is used to remove artifacts, including heart rate, muscle movement, and eye blink-related artifacts.

D. OTHER DATASETS USED

We have used DEAP dataset [2] (a dataset for emotion analysis using EEG, physiological and video signals) and SEED dataset [3] (A dataset collection for various purposes using EEG signals) for comparing the results with our dataset (DENS).

The DEAP dataset consisted of 40 videos/trials, and for each trial, there are 40 channels of EEG, including peripheral signals, are available, and data is given for each channel. We have used only 32 channels (i.e., discarded peripheral signals) for the experiment as we only want to use data from the brain only. This data was already preprocessed as 128 Hz downsampled, bandpass frequency of 4-45 Hz and EOG removed. For each trial, there are 4 labels available- Valence (V), Arousal (A), Dominance and Linking. We have used only V-A space for the experiment purpose.

The SEED dataset was recorded for 15 participants, and emotions were presented to the participants into three categories- positive, negative and neutral emotions (i.e., only valence (V) values were used). We have used only V-space in the DENS dataset to match the number of classes for both the datasets. The data was recorded using 62 channels.

E. SOME SALIENT FEATURES OF THE DENS

To sum up, we are highlighting some key points of our dataset-

- To the best of our knowledge, the first time, we created a dataset on Emotion with Naturalistic Stimuli (DENS) and recorded EEG signals from participants in the Indian subcontinent.
- Stimuli that are used to record EEG data of the participants are pre-validated on a different set of participants for the selected emotion categories.
- Participants were free to select any emotion category, whatever they felt for the stimuli from the given list.
- We used 128-channel high-density EEG recording for higher spatial resolution.
- **Emotional Event:** Temporal markers are available for each emotion category when participants feel the emotion, resulting in higher temporal resolution.

III. METHODOLOGY

A. FEATURE EXTRACTION

EEG Signals are non-stationary, meaning the signal's statistical characteristics change over time. If these signals are transformed to the frequency domain using Fourier Transform, it provides the frequency information, which is averaged over the entire EEG signal. So, information on different frequency events is not analyzed properly. If a signal is cut into minor segments such that it could be considered as stationary and focus on signal properties at a particular section which is called a windowing section and apply Fourier transform on it, it is called as Short-Time Fourier Transform (STFT). It will move to the entire signal length and apply Fourier transform to find the spectral content of that section and display the coefficient as a function of both time and frequency. It provides insight into the nature of the time-varying spectral characteristics of the signal. Before STFT, let's look at the discrete Fourier transform. Consider x: [0: L-1] = $\{0, 1, \ldots, L-1\}$ → R be a discrete-time signal where L is the signal length which is acquired by equidistance sample points with respect to the fixed sampling frequency. Mathematically DFT equation is,

$$\widehat{x}(k) = \sum_{n=0}^{N-1} x(n).e^{-i2\pi nk/N} \tag{1}$$

where k ϵ [0, K] and K is the frequency index with respect to frequency. N is the duration of the section. The equation returns the complex Fourier coefficients for the kth. These coefficients provide two parameters: phase and magnitude. For STFT, consider the additional parameter hop size (H), which is the step size of the window to be shifted. ω be a sampling window function which is ω: [0, N-1] → R. STFT can be defined as,

$$S(m, k) := \sum_{n=0}^{N-1} x(n + mH)\omega(n)e^{(-i2\pi kn/N)} \tag{2}$$

where m ϵ [0,M] and M is the maximum frame index mathematically M = $\left\lfloor \frac{L-N}{H} \right\rfloor$. The Short-Time Fourier Transform is not only a function of k but also m which is a proxy time representation. Here, the function returns Fourier coefficients for the kth proxy frequency at the mth temporal bin.

Spectrograms are nothing but the squared magnitude of STFT of the signal.

$$\chi(m, k) := |S(m, k)|^2 \tag{3}$$

It is a 2D image where the horizontal axis represents time, and the vertical axis represents frequency bins. The number of frequency bins is (framesize / 2) + 1 and the number of time frames is ((size of signal – framesize) / hopsize) + 1. $\chi(m, k)$ represents intensity or color at (m, k).

B. INPUT PREPROCESSING TO FEED DATA INTO THE CLASSIFIER

It is essential to convert the data into a meaningful format that can be fed into our classifier model. As all three datasets are available in different formats, we have provided information on the input preprocessing for each dataset as follows:

1) FOR DEAP DATA

Each subject in the DEAP dataset is given by a tensor that is in the form of X $\in \mathbb{R}^{40\times40\times8064}$, representing 40 videos, 40 EEG-channels (including peripheral channels), and 8064 EEG data samples for each channel. For labels, DEAP data provided a matrix in the form of X $\in \mathbb{R}^{40\times4}$; i.e., for each subject, there are 40 videos and 4 scales. From the DEAP dataset, the first 15 subjects are picked. For the label, we used the Valence-Arousal space and divided it into four classes- HVHA, HVLA, LVLA, LVHA (abbreviations- H: High, L: Low, V: Valence, A: Arousal). The ratings for valence and arousal range from 1 to 9. Hence, we considered ratings from 1 to 5 as 'Low' and 5 to 9 as 'High' and divided the V-A space into 4 quadrants accordingly.

We converted 15 subjects' data tensor into a matrix of X $\in \mathbb{R}^{19200\times8064}$ ($i.e.$, 15 $subjects$ × 40 $videos$ × 32 $channels$, 8064 $samples$). Moreover, this data was processed for feature extraction using STFT with a window size of 0.5s and an overlap of 0.25s of data samples. Using STFT, we have converted every 8064 sizes of EEG data samples into a spectrogram image size of (33,251), as mentioned in the feature extraction section. Then, a hybrid CNN-LSTM classifier was implemented for multi-class classification with an input tensor of X $\in \mathbb{R}^{33\times251\times3}$.

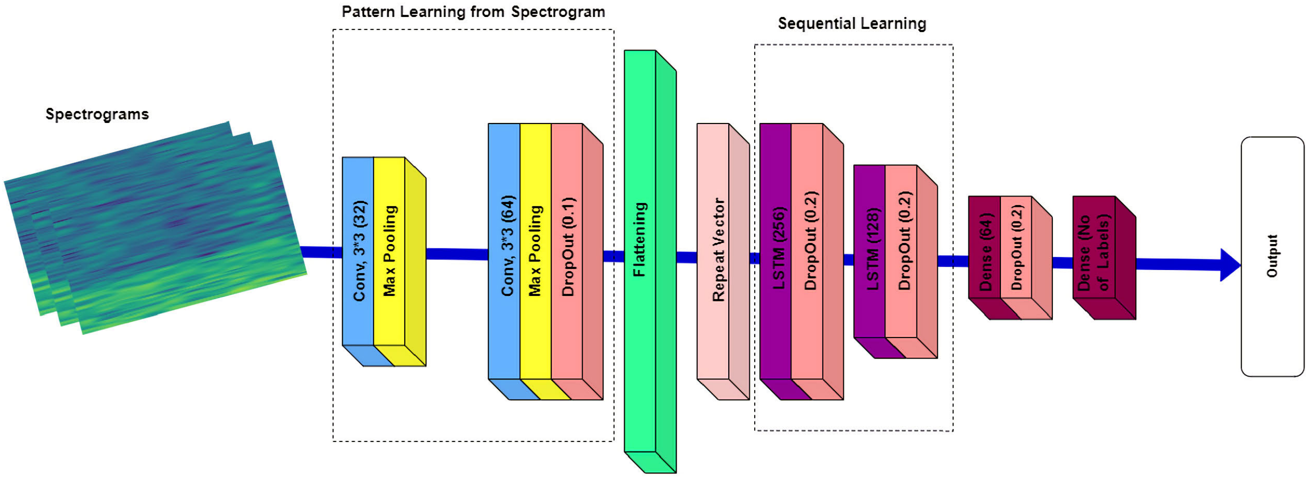


FIGURE 3. Model Architecture: It is consisted of two 2D-convolution layers with 3 × 3 kernels and 32 filters and 64 filters respectively, followed by a max pooling layer followed by a dropout layer and flattening layer. A repeat vector layer of size 4 is used before sending the data to the LSTM layers. Two LSTM layers are used of sizes 256 units and 128 units respectively, each followed by a dropout layer. At the end, two dense layers are used of sizes 64 (followed by a dropout layer) and 4 or 3 (equals the number of the output classes).

2) FOR SEED DATA

SEED dataset contains 45.mat files for 15 subjects for each subject with 3 trials. The label file contains 3 emotional labels −1 for negative, 0 for neutral, and 1 for positive on the valence scale. After renaming, the labels become 0 for neutral, 1 for positive, and 2 for negative. For classification, we have considered 15.mat files, one trial per subject. Due to the different sizes of data length in each channel, the first 16000 sample for each data which is the first 80s of data, is considered for further processing. EEG cap includes 62 channels according to the 10-20 international system. So, 15 subjects, 15 trials, 62-channels, and 16000 EEG data are converted into a tensor of $X \in \mathbb{R}^{13950\times 16000}$ ($i.e.$, $15\ subjects \times 15\ trials \times 62\ channels$, $16000\ samples$)for feature extraction. As mentioned in the DEAP dataset experiment, using STFT with a window size of 0.5s and overlap of 0.25s, each 16000 EEG data is converted into a spectrogram with the shape of (51,319). Then, a hybrid CNN-LSTM classifier was implemented for multi-class classification with input tensor shape $X \in \mathbb{R}^{51\times 319\times 3}$.

3) FOR DENS DATA

For the DENS dataset, we have 465.mat files which contain emotional events. All 465 files are picked for the experiment. Each.mat file is a matrix of $X \in \mathbb{R}^{128\times 1751}$, where 128 is the number of EEG channels and 1751 is the sample data for each channel. Then we have converted the data tensor of $X \in \mathbb{R}^{465\times 128\times 1751}$ into the form of $\mathbb{R}^{59520\times 1751}$ ($i.e.$, $465\ emotional\ events \times 128\ channels$, $1751\ samples$) for feature extraction with window size 0.5s and overlap is 0.25s. After feature extraction, we have 59520 spectrograms, and each spectrogram is in the shape of (63, 26).

To compare with the DEAP dataset, the DENS dataset with 4-label classification is performed with a hybrid CNN-LSTM classifier. For the label, we used the same V-A space (HVHA, HVLA, LVLA, LVHA) (abbreviations- H: High, L: Low, V: Valence, A: Arousal) as it was used with the DEAP dataset. The ratings for valence and arousal range from 1 to 9. Hence, we considered ratings from 1 to 5 as 'Low' and 5 to 9 as 'High' and divided the V-A space into 4 quadrants accordingly. The dimension of input tensor is of $X \in \mathbb{R}^{63\times 26\times 3}$.

To compare with the SEED dataset, we have used a 3-label classification since there are only three classes available in SEED dataset. For the DENS dataset, on the valence scale, ratings below 4.5 are marked as negative (0 labelled), and ratings above 5.5 are marked as positive (2 labelled). For neutral labels, in the DENS dataset, we have non-emotional files; we have marked neutral (1 labelled) for those files' data. Then with the classifier, the input tensor of $X \in \mathbb{R}^{63\times 26\times 3}$ is used for classification.

C. MODEL ARCHITECTURE FOR THE CLASSIFICATION TASK

Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) are one of the most widely used deep learning techniques. CNNs are used to extract meaningful patterns and features from the data. The key element in CNN is the convolution operation using kernels that automatically learn the local patterns from data. These local features are then combined into more complex features when multiple CNN layers are stacked. Filters (i.e, weights trained) in this process are also known as feature detector matrices. Input data will be convoluted with a filter map by sliding the kernel window. At the same time, LSTM networks can capture the sequential pattern as LSTMs are best suited for time-series data. LSTMs are designed to work for temporal correlations. Therefore, to exploit the benefits of both CNN and LSTM, a hybrid CNN-LSTM architecture is used for

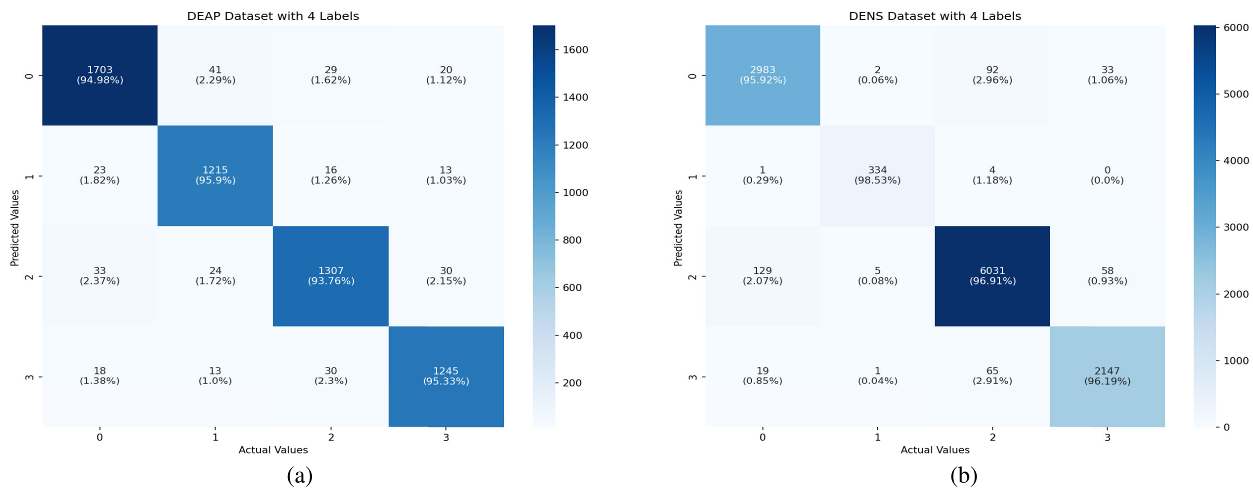


FIGURE 4. Comparison of Confusion matrices for DEAP and DENS datasets over Valence-Arousal space. This space is divided into four classes and assigned a label to it (0-HVHA, 1-HVLA, 2-LVHA and 3-LVLA). 4a: DEAP Dataset; 4b: DENS Dataset. Abbreviations of the terms- V:Valence; A:Arousal; L:Low; H:High. The color bar represents the number of samples in the class.

the classification of emotions. The hybrid CNN-LSTM model utilizes the ability of convolutional layers for feature extraction from data, and LSTM layers are for long-term and short-term dependencies. The same model is used to compare all three datasets. The model classifier and its details are shown in Fig. 3.

CNN is often placed in the initial layers as it helps in local pattern learning from spectrogram or in general input data. The Pattern learning block consists of two 2D-convolutional blocks, each with a kernel size of (3 × 3). The feature map, which is the output of convolutional layers, keeps track of the location of the features in the input. A max-Pooling layer is added in between two consecutive convolutional layers. A pooling layer is added after the convolutional layer to reduce the feature-map dimension; hence it reduces the computational cost, and the activation function is applied to enhance the capability of the model. Rectified Linear Unit (ReLU) activation function which has been widely used to resilient vanishing gradient problem. In between, the dropout layer is used in some places to avoid the overfitting problem. The flattening layer transforms these feature maps into one-dimensional vectors. The repeat vector gives extra dimension for the LSTM layer. The sequential learning block consists of 2 LSTM layers which capture the long-term temporal dependencies from the feature map extracted by CNN layers. 1^{st} LSTM layer consists of 256 cells with a return sequence set to 'True' while 2^{nd} LSTM consists of 128 cells and as it is the last LSTM layer return sequence is 'False'. Between LSTM layers, dropout layers with rate = 0.2 are added to avoid overfitting issues. Finally, two fully-connected layers where 1^{st} layer with 64 neurons and 2^{nd} layer with the number of classes as neurons are added for further processing. As we have the multi-class classification, the SoftMax activation function is used in the output layer as it outputs a vector representing the probability distributions of a list of potential classes.

TABLE 2. Parameter Settings for the Model.

Parameter	Setting
Optimizer	Adam
Loss function	Categorical Cross-entropy
Learning rate	0.001
Adjustment	Early Stopping criteria: monitor - 'val_loss'; patience = 30 Model Checkpoint: monitor - 'val_accuracy'
Batch size	256
Epochs	100

The parameter setting for the developed deep learning model is mentioned in Table 2.

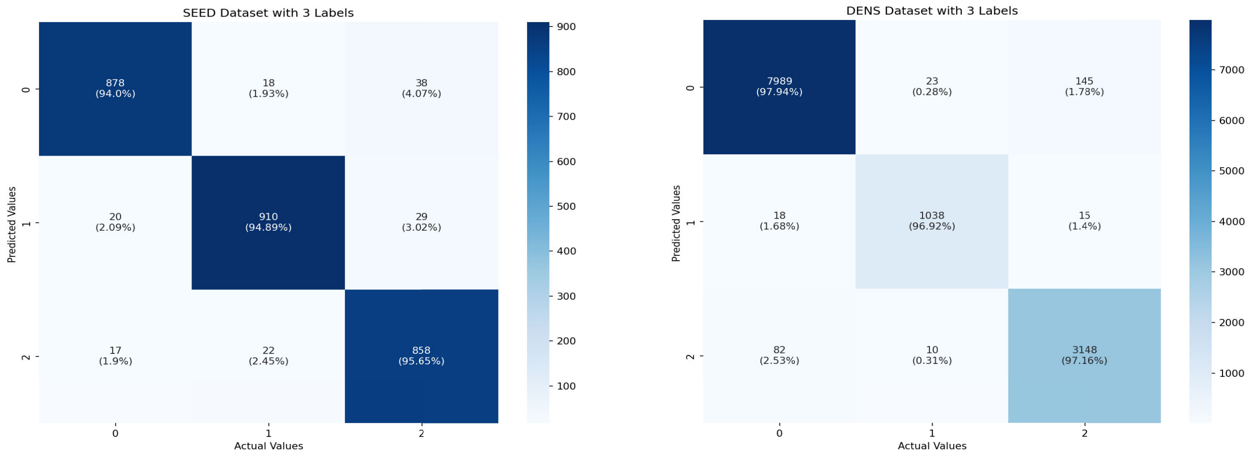
IV. RESULTS

The confusion matrix for DEAP, SEED and DENS datasets are shown in Fig. 4 and Fig. 5. In the confusion matrix shown, each cell contains data on the number of population. The X-axis represents actual labels and the Y-axis represents predicted labels by the classifier. The diagonal of the matrix represents the correctly identified label. The color bar represents the number of samples in the class.

A. COMPARISON BETWEEN DEAP AND DENS

We have used repeated K-Fold cross-validation with K = 5 and the number of repeats = 5 so generated 25 accuracies for DENS and DEAP. For label classification, we have used V-A space (HVHA, HVLA, LVLA, LVHA). Comparison between DEAP and DENS is mentioned in Table 4. The loss and accuracy graphs are mentioned in Fig. 8.

Fig. 6 shows an F1 score comparison between DEAP and DENS datasets per trial. Using t-test statistical testing, the 25 F1 scores of DEAP dataset (M = 95.65%, SD = 0.38%) compared with the 25 F1 scores of DENS dataset (M = 96.82%, SD = 0.18%), DENS dataset shows better results with absolute t (35) = 13.54, $p < 0.0001$,



(a) (b)

FIGURE 5. **Comparison of Confusion matrices for SEED and DENS datasets over Valence space. This space is divided into three classes (SEED dataset provided data with three classes, while DENS data is divided into three classes based on the valence ratings provided by the participants) and assigned a label to it as follows: For SEED:0 for neutral, 1 for positive and 2 for negative. For DENS: 0 for low-valence (valence ratings range from 1-4.5), 1 for non-emotional data (valence ratings range from 4.5-5.5, as well as neutral categories stimuli) and 2 for high-valence (valence ratings ranges from 5.5-9). 5a: SEED Dataset; 5b: DENS Dataset. The color bar represents the number of samples in the class.**

TABLE 3. **Comparison Table with Other Recent Studies.**

Method	Dataset	Subject Dependency	Emotion Classes	Result Accuracy (%)
CNN-RNN Hybrid Model [30]	DEAP	Subject Dependent	2	Valence: 72.06 Arousal: 74.12
R2G-STNN Model (region to global BiLSTM with Attention Layer) [31]	SEED	Both	3	Sub. Dependent: 93.38 Sub. Independent: 84.16
ACRNN (Attention Based C-RNN Model) [32]	DEAP	Subject Dependent	2	Valence: 93.72 Arousal: 93.38
BiDCNN (Bi-hemisphere Discrepancy CNN model) [33]	DEAP	Both	2	Sub. Dependent: Valence- 94.38, Arousal- 94.72 Sub. Independent: Valence- 68.14, Arousal- 63.94
ECLGCNN (A fusion model of GCNN + LSTM) [34]	DEAP	Both	2	Sub. Dependent: Valence- 90.45, Arousal- 90.60 Sub. Independent: Valence- 84.81, Arousal- 85.27
Our Work (CNN-RNN Hybrid Model using STFT)	DENS DEAP SEED	Subject Dependent	3 and 4	Valence (3 Classes): DENS- 97.68, SEED- 95.65 V-A Space (4 Classes): DENS- 96.82, DEAP- 95.65

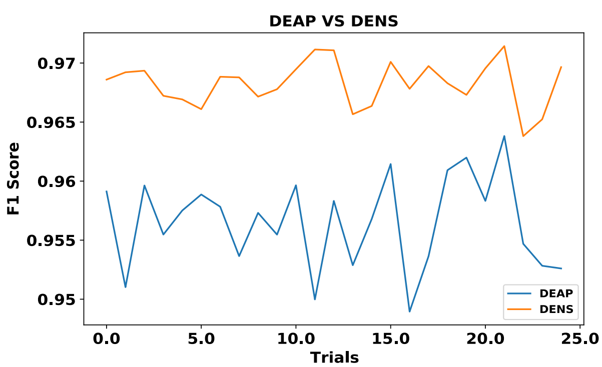


FIGURE 6. **F1 scores of DEAP vs DENS for all the 25 trials.**

d estimate: -13.70 (large), 95 percent confidence interval: $[-16.51\ -10.89]$.

TABLE 4. **DEAP vs DENS with mean F1 scores.**

Dataset	Mean F1 score (in %)
DEAP	95.65 ($\pm$ 0.38)
DENS	96.82 ($\pm$ 0.18)

TABLE 5. **SEED vs DENS with mean F1 scores.**

Dataset	Mean F1 scores (in %)
SEED	95.65 ($\pm$ 0.37)
DENS	97.68 ($\pm$ 0.13)

B. COMPARISON BETWEEN SEED AND DENS

For SEED vs DENS comparison, label classification we have used 3 labels on the valence scale. Comparison between SEED and DENS results is mentioned in Table 5. The loss and accuracy graphs are mentioned in Fig. 8.

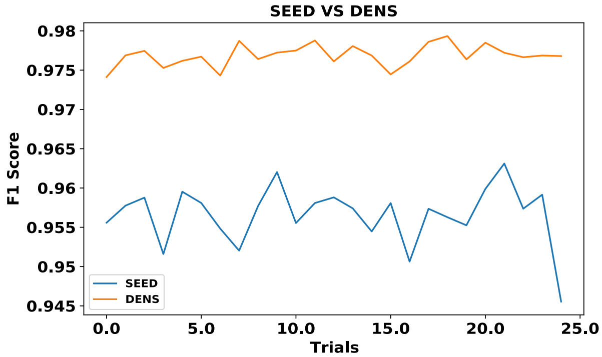


FIGURE 7. **F1 scores of SEED vs DENS for all the 25 trials.**

Fig. 7 shows an F1 score comparison between SEED and DENS datasets per trial. Using t-test statistical testing, the 25 F1 score of SEED dataset (M = 95.65%, SD = 0.37%) compared with the 25 F1 score of DENS dataset (M = 97.68%, SD = 0.13%), DENS dataset shows better results with absolute t (31) = 25.466, $p < 0.0001$, 's d estimate: −11.37 (large), 95 percent confidence interval: [−13.73 −9.02].

C. COMPARISON WITH OTHER RECENT STUDIES

We have included some other recent studies and given a comparative table for their results in Table 3. The studies consist of CNN-RNN Hybrid models, R2G-STNN model that is based on regional to global BiLSTM with Attention layer, Attention-based CNN-RNN Hybrid model (ACRNN), BiDCNN that is Bi-hemisphere Discrepancy CNN model and ECLGCNN that is a fusion model of Graph CNN and LSTM model.

V. DISCUSSION

In this work, we captured emotional experiences within the ecologically valid naturalistic environment with a precise temporal marker than any study to date. As per recent theories, emotional experience is a constructing phenomenon which involves networks of the brain, including the default mode network, salience network, and fronto-parietal network. These networks are not specific to emotional experiences. In fact, these networks are domain-general networks which are involved in perception (in general). Though, the connectivity among these networks might not be the same in different perceptions which is apparently shown in our previous work [35]. In addition, different from normal perception, emotional experiences involve changes in body physiology [29]. Putting together the above-mentioned ideas from recent results hints that the emotional experiences can be easily confused with other perceptions, which might not be an emotional experience.

One of the major concerns is the mind-wandering activity while using the film stimuli. In the previous research, the whole stimulus is considered to elicit a single emotional experience. And the duration of the stimulus varied from seconds to minutes. Research shows that averaging the participant's feedback for the whole duration of the stimulus might not be correctly capturing emotional experience (in particular) [36]. Hence, it is important to know the duration of the emotional experience without compromising the ecological validity of the stimuli.

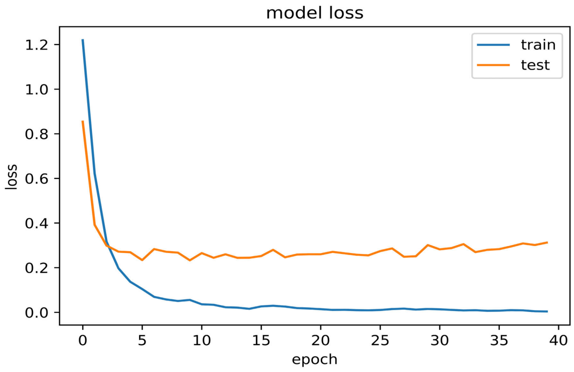
The main idea behind this work is that if we can capture the temporal marker of emotional experience within a real-life resembling environment, we might achieve better accuracy than the accuracy achieved to date with other datasets lacking information about time. Although, due to the limited number of subjects, we didn't go for the subject-independent classification for now. Though, in future, we will be collecting more data to mitigate this limitation.

In our results, we observed that the same hybrid deep learning model on our dataset not only outperformed other datasets, including benchmark datasets like DEAP and SEED but also achieved a better result when comparing with other noteworthy relevant studies (see Table 3). Classification of DEAP data into four labels, including HVHA, LVHA, LVLA, and HVLA, resulted in 95.65% mean accuracy. At the same time, the classification of DENS data into four labels resulted in 96.82% mean accuracy. Similarly, the classification of SEED data into three labels resulted in 95.65% mean accuracy while DENS data resulted in 97.68% mean accuracy. The significance testing showed that even with multiple iterations, the classification accuracy was significantly higher for our data.

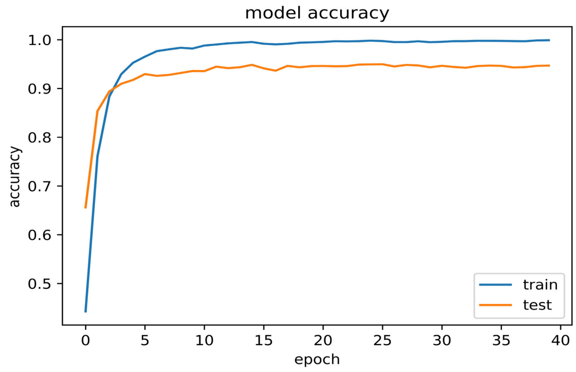
To the date, most of the work on emotion recognition applied different shallow machine learning and deep learning techniques using many different configurations of input data including, spectrogram, raw signals, statistical features, variational mode decomposition (VMD), empirical mode decomposition (EMD), functional connectivity based features, fractal features and so on. However, still, the recognition of emotion from EEG stands as a problem. Most of the works on emotion recognition have used some benchmark datasets, including DEAP, SEED, AMIGO, MAHNOB-HCI and so on. Though, most of the emotion classification works revolve around DEAP and SEED datasets [2].

In [37], emotional states are classified by means of EEG-based functional connectivity patterns. Forty participants viewed audio-visual film clips to evoke neutral, positive (one amusing and one surprising) or negative (one fear and one disgust) emotions. Correlation, coherence, and phase synchronization are used for estimating the connectivity indices. They stated significant differences among emotional states. A maximum classification rate of 82% was reported when the phase synchronization index was used for connectivity measure.

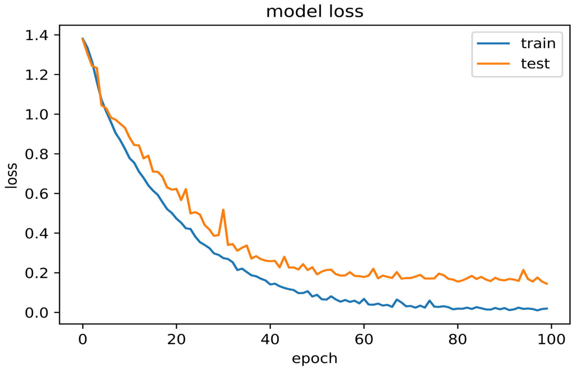
The classes considered in the study are elementary. We suspect that with the increasing number of emotional classes, which includes not only basic classes but complex emotions as well, taking the long-duration signal without a temporal marker may not be able to categorize emotional classes. The reason is that there are fewer chances for a movie stimulus to have a positive as well as a negative emotional



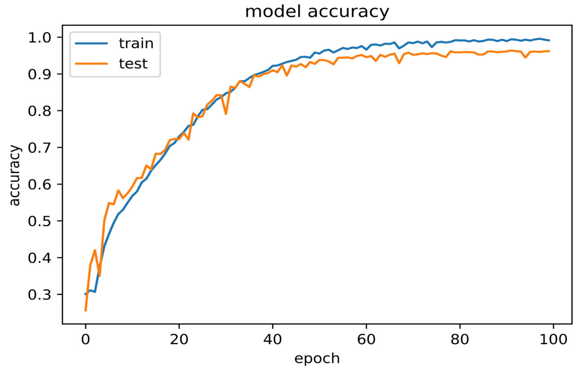
(a) Loss Graph for DEAP dataset



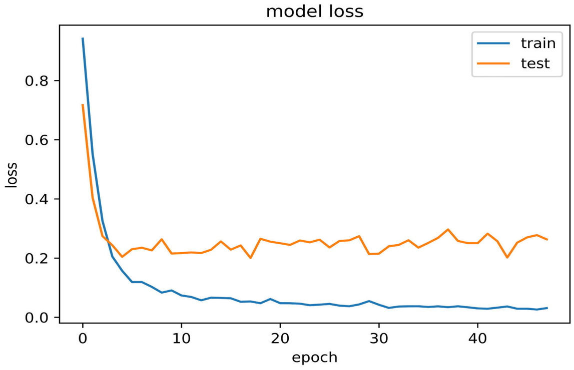
(b) Accuracy Graph for DEAP dataset



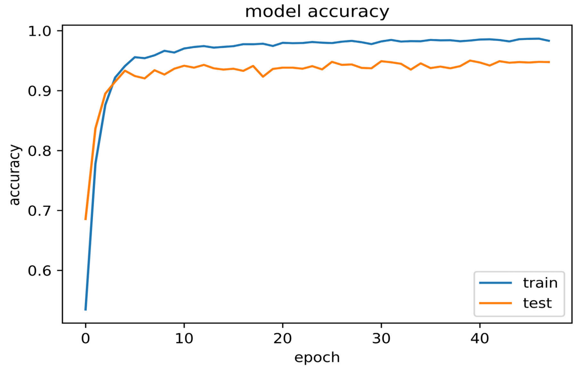
(c) Loss Graph for DENS (4 Classes) dataset



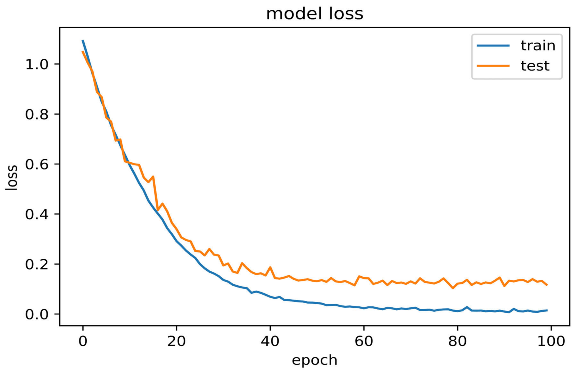
(d) Accuracy Graph for DENS (4 Classes) dataset



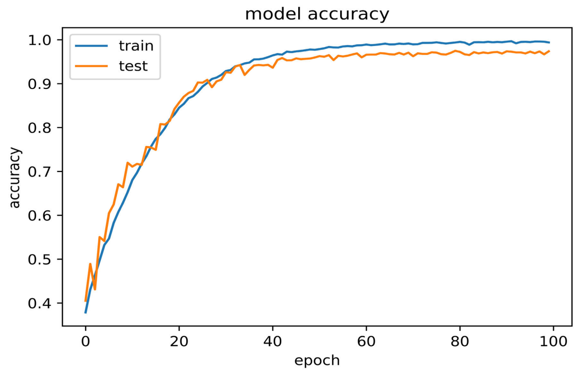
(e) Loss Graph for SEED dataset



(f) Accuracy Graph for SEED dataset



(g) Loss Graph for DENS (3 Classes) dataset



(h) Accuracy Graph for DENS (3 Classes) dataset

FIGURE 8. Loss and Accuracy Graphs for All the datasets Used.

experience in the same stimuli, but it is certainly possible that it can have more than one positive or more than one negative feeling in the movie.

VI. CONCLUSION

The work presented in this article is based on the concept that emotion is a short-lived phenomenon which might last for very few seconds. Hence, using long-duration EEG signals recorded during emotional stimulus watching might not contain emotional information for the whole duration. Therefore, we hypothesized that using only the duration of the signal where an emotional event is reported without compromising the ecological validity of the stimuli will contain more emotional information. To test the hypothesis, we designed an EEG experiment which uniquely marks the duration of the emotional event in the continuous recording of brain waves using EEG. We performed deep learning analysis using a hybrid CNN and LSTM model and found results that significantly favoured our hypothesis. In this work, we saw the problem with a different aspect which has not attracted the attention of the researcher. We suggest that future research on emotion recognition should adapt our approach to collect more such kinds of data so that emotion recognition using EEG can go beyond the emotions only and move towards recognizing and analyzing more complex emotions.

ACKNOWLEDGMENT

# Dataset on Emotion with Naturalistic Stimuli. Availability at (https://openneuro.org/datasets/ds003751). *(and contributed equally to this work.)*

REFERENCES

[1] and , “Measures of emotion: A review,” *Cognition Emotion*, vol. 23, no. 2, pp. 209–237, Feb. 2009.

[2] , , , , , , , and , “DEAP: A database for emotion analysis; using physiological signals,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan./Mar. 2012.

[3] and , “Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks,” *IEEE Trans. Auto. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015, doi: 10.1109/TAMD.2015.2431497.

[4] and , “DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices,” *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 98–107, Jan. 2018, doi: 10.1109/JBHI.2017.2688239.

[5] , , , and , “AMIGOS: A dataset for affect, personality and mood research on individuals and groups,” *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Apr. 2021, doi: 10.1109/TAFFC.2018.2884461.

[6] , , , , , and , “Recognition of human emotions using EEG signals: A review,” *Comput. Biol. Med.*, vol. 136, Sep. 2021, Art. no. 104696.

[7] and , “Determination of the hydrocel geodesic sensor nets’ average electrode positions and their 10–10 international equivalents,” Electrical Geodesics, Inc., Eugene, OR, USA, Tech. Note, 2005, pp. 1–11.

[8] , *Human Emotions*. New York, NY, USA: Springer, 2013.

[9] , “Basic emotions, natural kinds, emotion schemas, and a new paradigm,” *Perspect. Psychol. Sci.*, vol. 2, no. 3, pp. 260–280, Sep. 2007.

[10] , “A general psychoevolutionary theory of emotion,” in *Theories of Emotion*. Amsterdam, The Netherlands: Elsevier, 1980, pp. 3–33.

[11] , “Core affect and the psychological construction of emotion,” *Psychol. Rev.*, vol. 110, no. 1, p. 145, 2003.

[12] and “Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals,” *NeuroImage*, vol. 102, pp. 162–172, Nov. 2014.

[13] “The theory of emotion: I: Emotional attitudes,” *Psychol. Rev.*, vol. 1, no. 6, pp. 553–569, Nov. 1894.

[14] , “The James–Lange theory of emotions: A critical examination and an alternative theory,” *Amer. J. Psychol.*, vol. 39, nos. 1–4, p. 106, Dec. 1927.

[15] , “The emotional brain,” *Nature Rev. Neurosci.*, vol. 5, pp. 583–589, Jul. 2004.

[16] , “Applications of the short time Fourier transform to speech processing and spectral analysis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 1982, pp. 1012–1015, doi: 10.1109/ICASSP.1982.1171703.

[17] , , , and , “Deep learning-based electroencephalography analysis: A systematic review,” *J. Neural Eng.*, vol. 16, no. 5, Oct. 2019, Art. no. 051001.

[18] , *Deep Learning With Python*. New York, NY, USA: Simon and Schuster, 2021.

[19] , “Comment: Affective chronometry has come of age,” *Emotion Rev.*, vol. 7, no. 4, pp. 368–370, Oct. 2015.

[20] , , , and “Dataset on emotion with naturalistic stimuli (DENS) on Indian samples,” *bioRxiv*, pp. 1–11, Dec. 2022. [Online]. Available: https://www.biorxiv.org/content/early/2022/12/31/2021.08.04.455041, doi: 10.1101/2021.08.04.455041.

[21] , , and , “Context in emotion perception,” *Current Directions Psychol. Sci.*, vol. 20, no. 5, pp. 286–290, Oct. 2011, doi: 10.1177/0963721411422522.

[22] , , , , , , , , and , “Neural, electrophysiological and anatomical basis of brain-network variability and its characteristic changes in mental disorders,” *Brain*, vol. 139, no. 8, pp. 2307–2321, Aug. 2016.

[23] , , , , , and , “Dynamic shifts in large-scale brain network balance as a function of arousal,” *J. Neurosci.*, vol. 37, no. 2, pp. 281–290, Jan. 2017.

[24] , , , , , , , , , , and , “Functional connectivity dynamics during film viewing reveal common networks for different emotional experiences,” *Cogn., Affect., Behav. Neurosci.*, vol. 16, pp. 709–723, May 2016.

[25] , , , and , “Dynamic intersubject neural synchronization reflects affective responses to sad music,” *NeuroImage*, vol. 218, Sep. 2020, Art. no. 116512.

[26] , , , , , , , , and , “Default and control network connectivity dynamics track the stream of affect at multiple timescales,” *Social Cogn. Affect. Neurosci.*, vol. 17, no. 5, pp. 461–469, May 2022.

[27] , , and , “Repeated movie viewings produce similar local activity patterns but different network configurations,” *NeuroImage*, vol. 142, pp. 613–627, Nov. 2016.

[28] , , , and (Nov. 2021). *Affective Film Dataset From India (AFDI): Creation and Validation With an Indian Sample*. [Online]. Available: https://psyarxiv.com/yajsk

[29] , , and “Cardiac–brain dynamics depend on context familiarity and their interaction predicts experience of emotional arousal,” *Brain Sci.* vol. 12, no. 12, p. 702, 2022.

[30] , , , , and , “Emotion recognition from multi-channel EEG data through convolutional recurrent neural network,” in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 352–359.

[31] , , , , and , “From regional to global brain: A novel hierarchical spatial–temporal neural network model for EEG emotion recognition,” *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 568–578, Apr. 2022.

[32] , , , , , , and , “EEG-based emotion recognition via channel-wise attention and self attention,” *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 382–393, Jan. 2023.

[33] , , , , and , "Differences first in asymmetric brain: A bi-hemisphere discrepancy convolutional neural network for EEG emotion recognition," *Neurocomputing*, vol. 448, pp. 140–151, Aug. 2021.

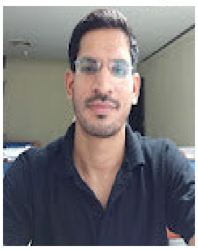
[34] , , , and , "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Appl. Soft Comput.*, vol. 100, Mar. 2021, Art. no. 106954.

[35] , , and "Dynamic functional connectivity of emotion processing in beta band with naturalistic emotion stimuli," *Brain Sci.*, vol. 12, no. 8, p. 1106, Aug. 2022.

[36] , "Naturalistic stimuli in affective neuroimaging: A review," *Frontiers Hum. Neurosci.*, vol. 15, p. 318, Jun. 2021.

[37] and , "Classifying different emotional states by means of EEG-based functional connectivity patterns," *PLoS ONE*, vol. 9, no. 4, Apr. 2014, Art. no. e95415.

(Graduate Student Member, IEEE) received the bachelor's degree in computer science and the master's degree in cognitive science and in information technology (specializing in software engineering). He is currently a Research Scholar with the Indian Institute of Information Technology Allahabad, Allahabad. His research interest includes affective computing. He is also working on emotion recognition using brain signals. He is using EEG for emotion detection using validated stimuli. He is also working on deep learning architectures.



received the master's degree in human–computer interaction from the Indian Institute of Information Technology Allahabad, Prayagraj, India, where he is currently pursuing the Graduate degree. He is also doing research on spatio-temporal dynamics of emotions. He has conducted two important experiments on Indian samples, which results in the availability of stimuli dataset (validated on an Indian sample) and the availability of EEG dataset with unique information about the time of emotional experience during watching the naturalistic multimedia stimuli. He is a member of the Society for Neuroscience.

was born in Jamnagar, Gujarat, India, in June 1995. He is currently pursuing the M.Tech. degree in IT with a specialization in machine learning and intelligent systems with the Indian Institute of Information Technology Allahabad, Allahabad. His research interests include machine learning, deep learning, and its application in cognitive science. He has two years of work experience as a Software Engineer with Tech Mahindra Ltd., Pune, India.

(Senior Member, IEEE) received the Ph.D. degree from the Department of Electronics Engineering, Institute of Technology, Banaras Hindu University, Varanasi, India, in 1991. He was a Lecturer with the Department of Electronics and Communication, J. K. Institute of Applied Physics and Technology, University of Allahabad, from September 1988 to March 1992. From March 1992 to June 2002, he was a Reader in computer science with the J. K. Institute of Applied Physics and Technology, University of Allahabad. He was also a Visiting Scientist with the Department of Computer Science and Engineering, IIT Kanpur, from December 1995 to July 1996. He was an Associate Professor with the Indian Institute of Information Technology Allahabad, Allahabad, India, from July 2002 to December 2006, where he has been a Professor with the Department of Information Technology, since December 2006. He is holding research and teaching experience for more than 30 years, in which he is very much involved in image processing, computer vision, medical image processing, pattern recognition and script analysis, digital signal processing, speech and language processing, wavelet transforms, soft computing and fuzzy logic, neurocomputing and soft-computers, speech-driven computers, natural language processing, brain simulation, cognitive science, and affective computing.