

Digital Object Identifier 10.1109/ACCESS.2023.3266804

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

Emotion Recognition Using Temporally Localized

Digital Object Identifier 10.1109/ACCESS.2023.3266804

Emotional Events in EEG With Naturalistic

Context: DENS# [ANONYMIZED]

MOHAMMAD ASIF

, (Graduate Student Member, IEEE), SUDHAKAR MISHRA

MAJITHIA TEJAS VINODBHAI, AND [ANONYMIZED], (Senior Member, IEEE)

Indian Institute of Information Technology Allahabad, Allahabad, Uttar Pradesh 211012, India

Corresponding authors: [ANONYMIZED] (rs163@iitaa.ac.in), [ANONYMIZED] (pse2017001@iitaa.ac.in), and (ust@iitaa.ac.in)

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

MOHAMMAD ASIF, (Graduate Student Member, IEEE), SUDHAKAR MISHRA, (Graduate Student Member, IEEE), MAJITHIA TEJAS VINODBHAI, AND UMA SHANKER TIWARY, (Senior Member, IEEE)

Indian Institute of Information Technology Allahabad, Allahabad, Uttar Pradesh 211012, India

Corresponding authors: Sudhakar Mishra (rs163@iitaa.ac.in), Mohammad Asif (pse2017001@iitaa.ac.in), and Uma Shanker Tiwary (ust@iitaa.ac.in)

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

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

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 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. This dataset can be used for detailed analysis of specific experienced emotions and related

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 par-

ticipants, 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, [ANONYMIZED] (EEG) is considered a

focus of the study was to develop a database that is

suitable and convenient method to record electrical activities

to measure brain activities as it is a non-invasive method, i.e.

there are no scalp 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

The associate editor coordinating the review of this manuscript and

approving it for publication was [ANONYMIZED]

Accordingly, Electroencephalography (EEG) is considered a

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

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.

For more information, see <https://creativecommons.org/licenses/by-nc-nd/4.0/>

VOLUME 11, 2023

39913

to be considered for the classification model, as there is no broader, e

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 feelings

To the best of our knowledge, there is no EEG affective datasets available for the Indian subcontinent population. Hence we tried to reduce this research gap in our work.

EEG measures the electrical signals from the scalp with In this work, we used a 128 channel EEG device to detect temporal details. Different EEG devices vary with the number of channels of EEG. Thirty-two or fewer EEG channels

Emotions are complex and challenging to understand as many theories exist about emotions, and there is a lack of a single consensus theory [18]. 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

dominance (controlling or feeling controlled). A few more dimensions are also considered, that make the spectrum of emotional states. Earlier, some basic emotions were used

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 data. 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 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.

ns, CNN is the most frequently used architecture for EEG
for disease 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
can help to enhance 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

[ANONYMIZED] et al.: Emotion Recognition Using Temporally Localized Emotional Events

FIGURE 1. [ANONYMIZED] of the Experiment.

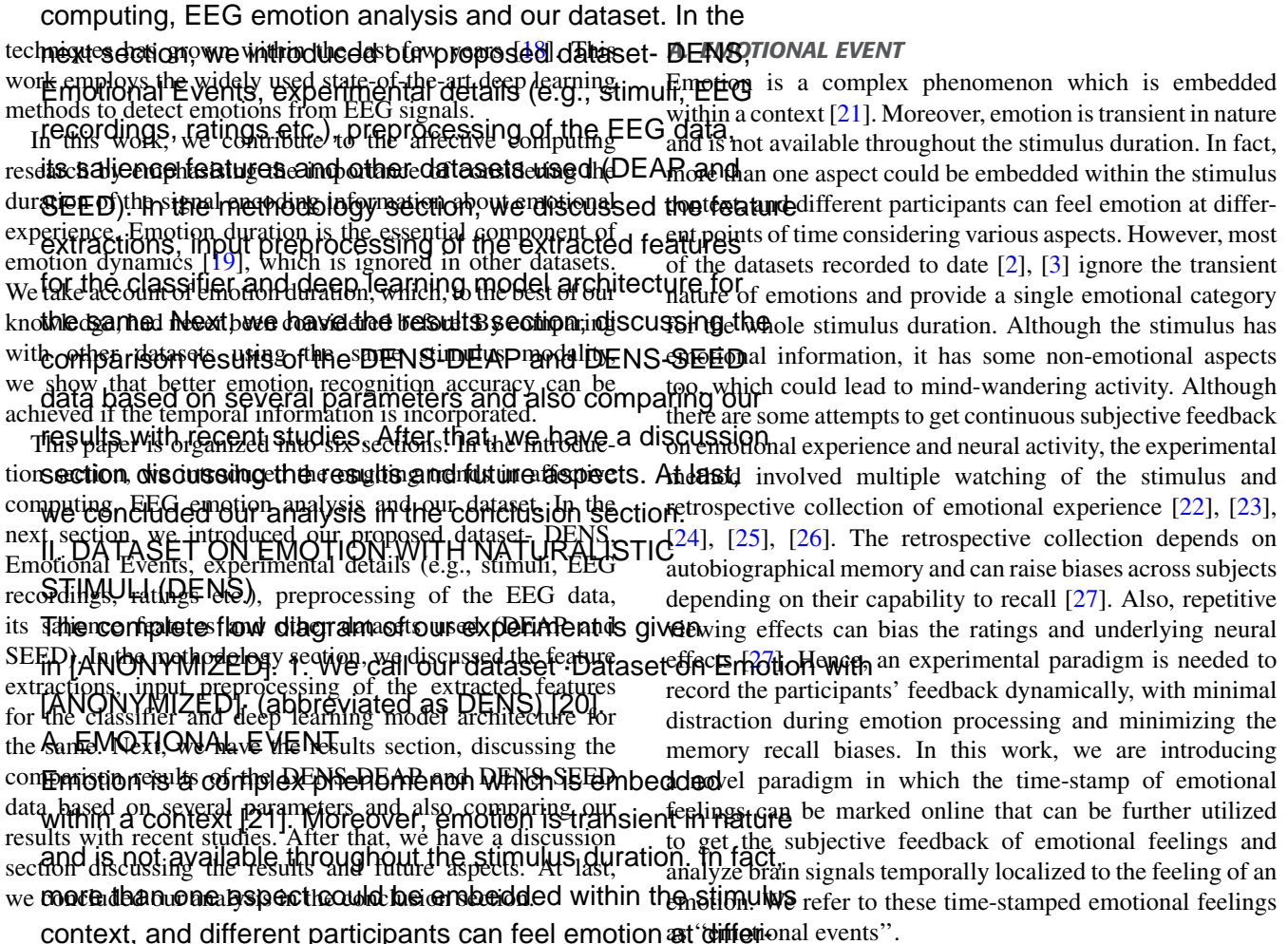
techniques has grown within the last few years [18]. This work employs the widely used state-of-the-art deep learning methods to detect emotions from EEG signals.

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, 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

FIGURE 1. Complete Flowgram of the Experiment.



next section, we introduced our proposed dataset- DENS, Emotional Events, experimental details (e.g., stimuli, EEG recordings, ratings etc.) preprocessing of the EEG data.

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, 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.

The complete flow diagram of our experiment is given in [ANONYMIZED] 1. We call our dataset- Dataset on Emotion with Naturalistic Stimuli (DENS), (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 design 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".

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 (DENS), (abbreviated as DENS [20]).

B. EXPERIMENTAL DETAILS

1) STIMULI

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

[ANONYMIZED] et al.: Emotion Recognition Using Temporally Localized Emotional Events

TABLE 1. Selected stimuli for EEG study from the stimuli dataset we created (23). The time duration of each stimulus is 60s. [ANONYMIZED] are given for references available in the open science framework repository.

Stimulus Name	Stimulus ID	Target Emotion
Ashtayen	199	Adventurous
Hunt For The Bambi	10	Frightened
Anacondas The Hunt For The Bambi	10	Alarmed
Lage Raho Munnabhai Only The Funny Scenes	98	Amused
Angst Behind of Robert Simm	108	Anxious
Divergent Kiss Scene Clip	54	Aroused
Best Horror Kills Ghost Ship Opening Scene	26	Disgust
Jar Jar Jar	182	Embarrassed
Sadda Haq	152	Excited
Mach	201	Happy
Cheerful Rang	201	Joyous
Crash Saddest Scene	3	Wretched
Has JBS	210	Miserable
Madani Movie Of Best Scene	113	Sad
Emotional Stimuli		Triumphant

- 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 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.

- 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:

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.

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.

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.

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.

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, EVLA, 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 60 seconds while the participant looked at the cross-hair 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.

[ANONYMIZED] et al.: Emotion Recognition Using Temporally Localized Emotional Events

FIGURE 2. Emotion Category Selection Screen for Emotional Event (Click): After the participants rated all the

Dominance, Liking, Familiarity and [ANONYMIZED], they are shown this screen for emotion category selection. The middle one belongs to the time of the click; the left one is 20 frames earlier, and the right one is 20 frames later (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 [ANONYMIZED]. The rating scales range from 1 to 9 for

Valence, Arousal and Dominance. For Liking, familiarity and

[ANONYMIZED], 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.

NEXT

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, while the left one is 20 frames earlier, and the right one is 20 frames later (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.

(1 second before the click and 6 seconds after the click) for

3) 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

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.

D. OTHER DATASETS USED

Although for each participant and each stimulus, EEG

recording is available, the 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.

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.

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.

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.

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

VOLUME 11 2023

[ANONYMIZED] et al.: Emotion Recognition Using Temporally Localized Emotional Events

FIGURE 3. Model Architecture: It is consisted of two 2D-convolution layers with 3×3 kernels and 32 filters a [ANONYMIZED] pooling layer followed by a dropout layer and flattening layer. A repeat vector layer of size 4 Two LSTM layers are used of sizes 256 units and 128 units respectively, each followed by a dropout layer. A 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 $X \in \mathbb{R}^{13950 \times 16000}$

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 $X \in \mathbb{R}^{13950 \times 16000}$

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, 1: 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

classifier. For the label, we used the same V-A space (HVHA, HVLA, LVLA, LVHA) (abbreviations- H: High, 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 $X \in \mathbb{R}^{63 \times 26 \times 3}$.

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 convolution operation using kernels that automatically capture 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 down 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

[ANONYMIZED] et al.: Emotion Recognition Using Temporally Localized Emotional Events

FIGURE 4. Comparison of Confusion matrices for DEAP and DENS datasets over Valence-Arousal space. T and assigned a label to it (0-HVHA, 1-HVLA, 2-LVHA and 3-LVLA). 4a: DEAP Dataset; 4b: DENS Dataset. A

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 [ANONYMIZED]. 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 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.

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. 1st LSTM layer consists of 256 cells with a

return sequence set to the True while 2nd 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 1st layer with 64 neurons and 2nd

layer with the number of classes are added for

further processing. As we have the multi-class classification,

the [ANONYMIZED] 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.

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 [ANONYMIZED]. 4 and [ANONYMIZED]. 5

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

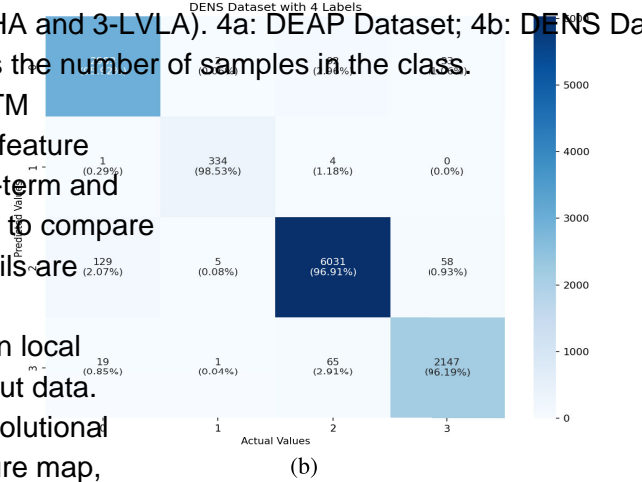


TABLE 2. Parameter Settings for the Model.

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.

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 the number of repeats = 5 so generated 5000 samples 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 over DEAP dataset ($t = 13.54$, $p < 0.0001$).

[ANONYMIZED] et al.: Emotion Recognition Using Temporally Localized Emotional Events

FIGURE 5. Comparison of Confusion matrices for SEED and DENS datasets over Valence space. This space dataset provided data with three classes, while DENS data is divided into three classes based on the valence participants) and assigned a label to it as follows: For [ANONYMIZED] for neutral, 1 for positive and 2 for negative 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.

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.

TABLE 5. SEED vs DENS with mean F1 scores.

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.

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

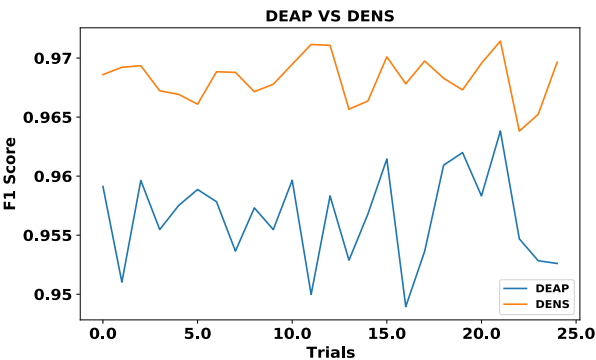


FIGURE 6. F1 scores of DEAP vs DENS for all the 25 trials.

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 (± 0.38)
DENS	96.82 (± 0.18)

TABLE 5. SEED vs DENS with mean F1 scores.

Dataset	Mean F1 scores (in %)
SEED	95.65 (± 0.37)
DENS	97.68 (± 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.

[ANONYMIZED] et al.: Emotion Recognition Using Temporally Localized Emotional Events

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

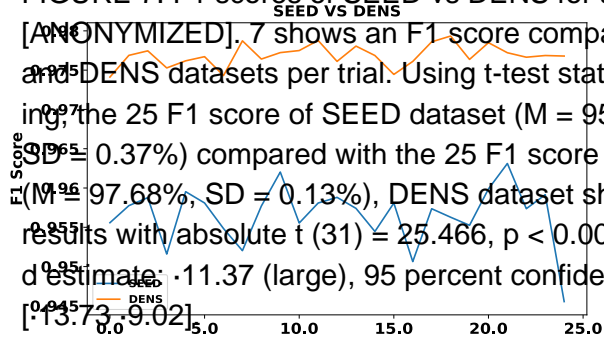


FIGURE 7. F1 scores of SEED vs DENS for all the 25 trials.

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 [9, 10].

V. DISCUSSION

C. COMPARISON WITH OTHER RECENT STUDIES

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 that is based on regional-to-global BiLSTM with Attention layer, Attention-based CNN-RNN Hybrid model (ACRNN), 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

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 naturalistic environment, we might achieve better results 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 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 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, empirical 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].

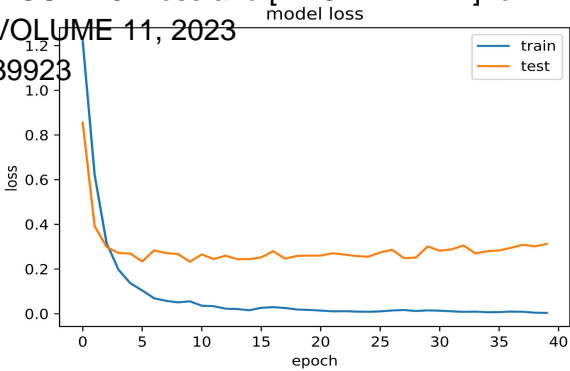
In [37], emotional states are classified by means of EEG-based functional connectivity patterns. Forty participants watched audio-visual film clips to evoke neutral, positive (one amusing and one surprising) or negative (one fear and one disgust) emotions. Correlation, coherence, and phase synchronization 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

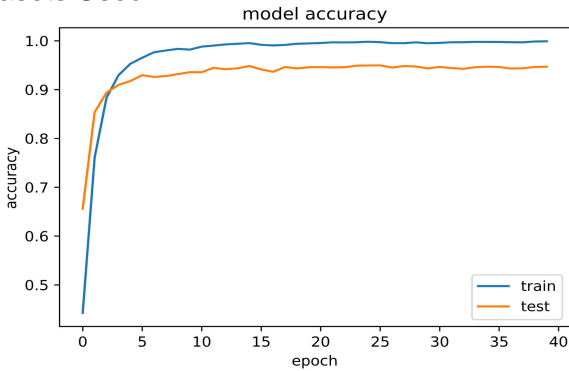
[ANONYMIZED] et al.: Emotion Recognition Using Temporally Localized Emotional Events

FIGURE 8. Loss and [ANONYMIZED] for All the datasets Used.

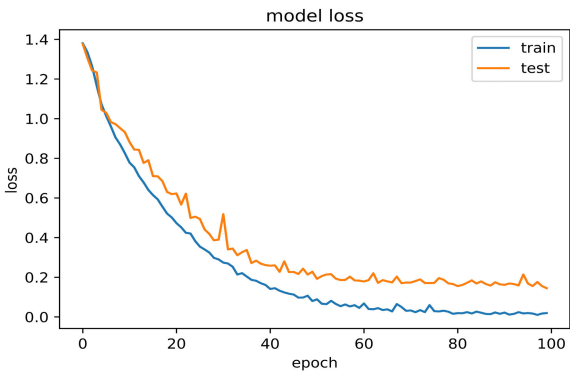
VOLUME 11, 2023
39923



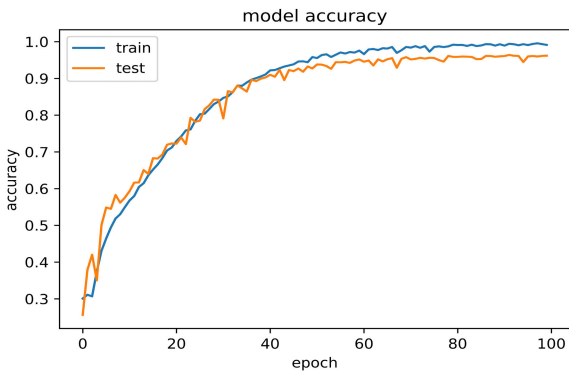
(a) Loss Graph for DEAP dataset



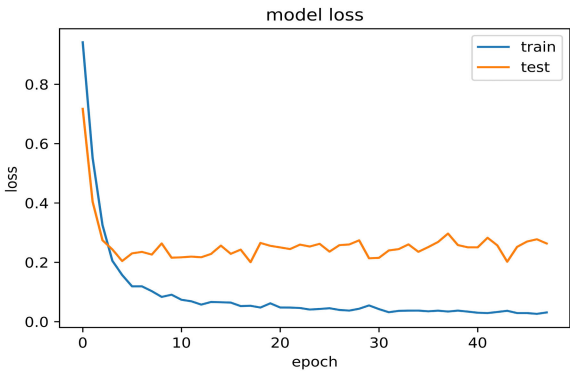
(b) Accuracy Graph for DEAP dataset



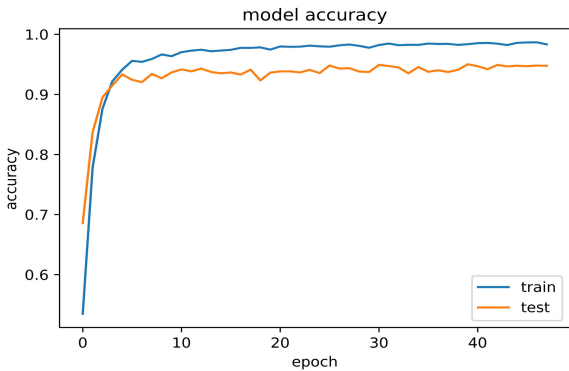
(c) Loss Graph for DENS (4 Classes) dataset



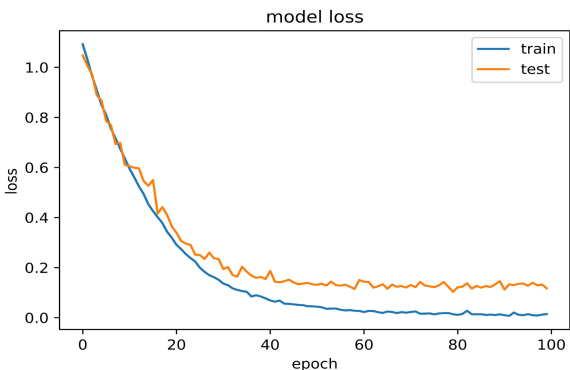
(d) Accuracy Graph for DENS (4 Classes) dataset



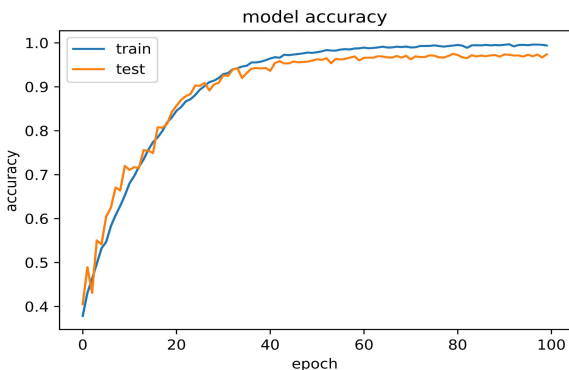
(e) Loss Graph for SEED dataset



(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

VI. CONCLUSION

VI. CONCLUSION The work presented in this article is based on the concept^[13] J. Dew

signals recorded during emotional stimulus watching might not contain emotional information for the whole duration.

compromising the ecological validity of the stimuli will [17] Y. Roy
we designed an EEG experiment which uniquely marks the J. Faub
contains more emotional information. To test the hypothesis thematic

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 recognizing and analyzing more complex emotions.

EEG can go beyond the limitations of fMRI and move towards
 at (<https://openneuro.org/datasets/ds003751>). (Mohammad

Dataset on Emotion with Naturalistic Stimuli Availability

Dataset on Emotion with Naturalistic Stimuli. Availability

- Apr. 2021. doi: 10.1109/TAFEC.2018.2884461

[33] [ANONYMIZED], [ANONYMIZED], [ANONYMIZED], [ANONYMIZED], [ANONYMIZED], and [ANONYMIZED]. D. Huang, S. Chen, C. Liu, L. Zheng, Z. Tian, and D. Jiang, "Differences first in asymmetric brain: A left-hemisphere discrepancy convolutional

- 

[illegible]

a Research Information
His research
He is also w
brain signals
detection us
architecture
SUDHAKAR
degree



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 had a position of Professor in 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.

interaction from the Indian Institute of Information Technology Allahabad, Prayagraj, India, where he is currently pursuing the [ANONYMIZED] 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.

...