Emotion recognition in VAD space during emotional events using CNN-GRU hybrid model on EEG signals

Majithia Tejas Vinodbhai [*,†,1] mit2020058@iiita.ac.in, Mohammad Asif [*,†,1] pse2017001@iiita.ac.in, Sudhakar Mishra [*,†,1] rs163@iiita.ac.in, Aditya Gupta [*1], and US Tiwary[1]

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

Abstract. Emotion recognition from brain signals is an emerging area of interest in the scientific community. We used EEG signals to classify emotional events on different combinations of valence(V), arousal(A) and dominance(D) dimensions and compared their results. DENS data is used for this purpose which is primarily recorded on the Indian population. STFT is used for feature extraction and used in the classification model consisting of CNN-GRU hybrid layers. Two classification models were evaluated to classify emotional feelings in valence-arousal-dominance space (eight classes) and valence-arousal space (four classes). The results show that VAD space's accuracy is 97.50% and VA space is 96.93%. We conclude that having precise information about emotional feelings improves the classification accuracy in comparison to long-duration EEG signals which might be contaminated by mind-wandering. In addition, our results suggest the importance of considering the dominance dimension during the emotion classification.

Keywords: Emotion Recognition · Affective Computing · Electroencephalography · EEG · Deep Learning · CNN · GRU · SFTP .

[*] These authors contributed equally to this work and author sequence is random.
[†] corresponding authors.

List of Abbreviations

CNN Convolution Neural Network
$DENS$ Dataset on Emotions with Naturalistic Stimuli
EEG Electroencephalography
GRU Gated recurrent unit
ICA Independent component analysis
RNN Recursive neural network
$STFT$ Short time fourier transform

1 Introduction

Emotion classification has been a challenging and emerging topic in AI and especially in affective computing. There are several methods for detecting emotions by the intelligent systems, e.g.- detecting emotions from a given image, video or from a text sequences. It is also possible to detect emotions from the brain signals and we can trust the results as these brain signals are infallible, so no manipulation there on this front. But brain signals come with their own challenges which mostly related to noise present in the signal. These signals consist of several artefacts including but not limited to line noise, noise from the muscle activities and eye movements, interference from other cognitive activities etc.

There are several methods for recording brain signals e.g., functional Magnetic Resonance Imaging(fMRI), EEG and Magnetoencephalography (MEG). For emotion recognition EEG is widely used as it is reliable, relatively less expensive and offers better temporal information. Some famous studies to recognize emotion from EEG data are [1] [2] [3] [4]. We have used our own data collected in our lab which follows a modified paradigm of collecting emotion information from t he participants which we call 'Emotional Event', more information in Method section.

After selecting which type of data to use, next comes an important decision of feature engineering and deep learning model architecture. As EEG data contains time series data, it is beneficial to convert it into cosine functions via a Fourier Transformation hence we used STFT for feature extraction. as these signals are converted into 3D spectrograms, CNNs are best suited to handle this type of data and RNNs are useful for handling the temporal information of the data. So we built a hybrid architecture for the emotion classification task.

Convolution Neural Networks are embodied with neurons capable of optimizing on their own through learning. CNNs are primarily used in images for pattern recognition, so we can extract key features from images to make the network stronger for accurate results [13]. The architecture of CNN consists of three dimension data as input height, width and depth. CNN can extract features, and CNN kernels are used to determine which part of input data we need to extract features [14]. The major advantage of using CNN is that its layer

is not necessarily connected to all the previous neurons. It majorly focuses on the part of data where useful information is recited. And also, in CNN many connections can use similar weights to reduce the complexity of the network.

GRU is the latest form of the recurrent neural network, which has the capability to resolve vanishing exploding problems of RNN similar to LSTM but is basically a lighter version of LSTM with similar or more efficiency [15]. The GRU controls the flow of information like the LSTM unit, but without using a memory unit. It just exposes the full hidden content without any control [16]. The major change in GRU is fusing the inner cell and the hidden state into a single one, and this collective data get passed on to the next GRU.

2 Method

2.1 Participants

Forty participants participated in the study, including 3 females (Mean age = 23.3, SD = 1.4). The institutional review board approved the study.

2.2 DENS Data

The EEG data which we have used in this analysis was collected from Indian samples for different emotion categories. EGI 128 dense array system was used for recording the brain activity during emotional activity. The complete experiment paradigm is available elsewhere [10]. Subjects are shown emotional videos selected from an affective film stimuli dataset validated on the Indian population [9]. Subjects were asked to mark the moments of emotional feelings by performing left mouse clicks while looking at the film stimulus. At the end of the stimulus watching, participants categorized their emotional feelings into suitable emotional categories. They were also asked to provide their feedback on valence, arousal, and dominance scales.

2.3 Data Preprocessing

The complete preprocessing details are available elsewhere [10]. The data is filtered using the fifth order Butterworth bandpass filter with passband 1 Hz to 40 Hz. Independent component analysis (ICA) was used to remove artefacts. After the preprocessing, a seven seconds segment of EEG signal around the subjective marking of emotional feeling is extracted. We will refer to this seven seconds segment as the emotional events. Likewise, 420 emotional events from 40 participants were collected.

2.4 Feature Extraction and Input to the Model

The Short-Time Fourier Transform, also called STFT, is a logical continuation of the Fourier transform that uses windows to do segmented analysis to deal

with non-stationarity of the signals [11]. The time-localized frequency information that is provided by the STFT is superior to the frequency information that is provided by the standard Fourier transform, which is an average of frequency information over the entirety of the signal's time interval [12]. Within the realms of signal processing and signal synthesis, STFT has been put to work in a wide variety of applications over the years. STFT is effective when applied to unimodal, univariate signals since there is no multiple component complexity, and signal artifacts and noise move relatively slowly still, STFT is an excellent option. Since the fixed window limits how many non-stationary features can be extracted, STFT features might need a threshold level when they are extracted and fitted. On the other hand, in comparison to the Fourier transform, the short-time Fourier transform (STFT) possesses superior temporal and frequency localization capabilities. The Space-Time Frequency Transform (STFT) divides the space of time and frequency into grids of the same size. Spectrogram is a two-dimensional image with time on the horizontal axis and frequency bins on the vertical axis. frequency bin count = (framesize / 2) + 1, while ((signal size - framesize) / hopsize) + 1 represents the number of time frames. The Gabor transform serves as the mathematical foundation for the spectrogram. Utilizing the Gabor transform, the spectrogram is computed. A special form of the short-time Fourier transform known as the Gabor transform is used to isolate the sinusoidal frequency and phase content of a signal in a certain region. Formally, we can extract the frequency components that make up any signal after applying a Fourier transform to it.

We have a total of 465 .mat files with emotional events in the DENS dataset. All 465 of the files have been chosen for the experiment. Each .mat files have the following format: (128, 1751), where 128 is the total number of channels and 1751 is the EEG data. After that, we changed the data from (465,128,1751) into (59520, 1751) so that we could extract features with a 0.5-second window and a 0.25-second overlap. Following the feature extraction, 59,520 spectrograms each have shape (63,26).

2.5 Architecture of the Model

Our model consists of 32 channels CNN followed by Dropout layer and Max Pooling layer. CNN architecture is a combination of different layers like convolution layer and pooling layer, connected layer and many more, and it works in a feed-forward fashion to execute various layers. The input shape of CNN is in 3 dimensions which compromise of height, width and features, usually 3 to represent R, G and B of the image.

Then we used two layers of GRU having 256 and 128 units, respectively, each followed by a dropout layer. The two important gates of GRU are updated and reset gates. The update gate basically determines how past information needs to be passed on future whereas the reset gate determines how much past information needs to be deleted (Fig. 1). The advantage of using GRU is that it requires less run time compared to LSTM, with quite a similar result.

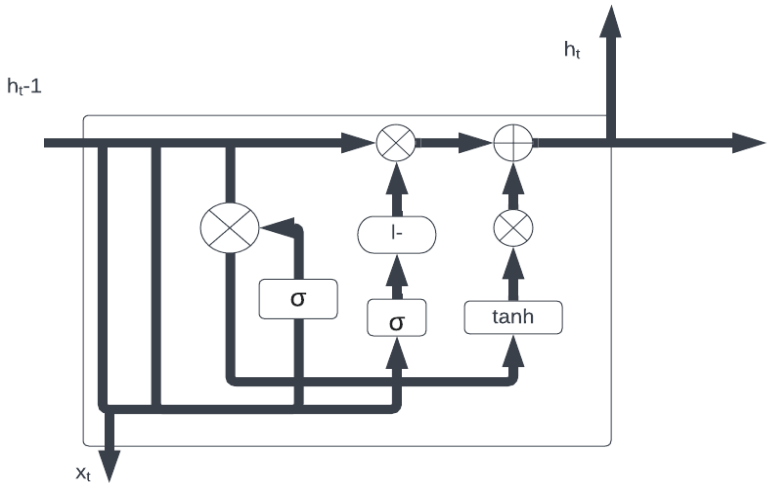


Fig. 1: GRU Structure

At last we used a 64 nodes fully connected layer followed by the output layer. The detailed architecture of the model is given in Fig.2

Formulation of GRU

$$r_t = \sigma(U^r x_t + W^r s_{t-1} + b^r)$$
$$z_t = \sigma(U^z x_t + W^z s_{t-1} + b^z)$$
$$s_t = Z_t o s_{t-1} + (1 - z_t) o(\sim s_t)$$
$$\sim s_t = g(U^s x^t + r_t o W^s s_{t-1}) + b^s$$

Where, t is the time instance; x_t is the input at time t; U^* &W^* are weight matrices; σ &g are sigmoid and tangent activation function, p^* &b^* are peephole connection and biases, and o means element wise product.

2.6 Classification Tasks

We have compared four sets of accuracies in this experiment. For VAD space label classification, Valence, Arousal, Dominance classes are used. In this setup each class is divided into high and low (threshold 5) and by combining these classes 8 labels like High Valence High Arousal High Dominance marked as 0, High Valence High Arousal low Dominance marked as 1, High Valence Low Arousal High Dominance marked as 2, High Valence Low Arousal Low Dominance marked as 3, Low Valence High Arousal High Dominance marked as 4, Low Valence High Arousal low Dominance marked as 5, Low Valence Low Arousal

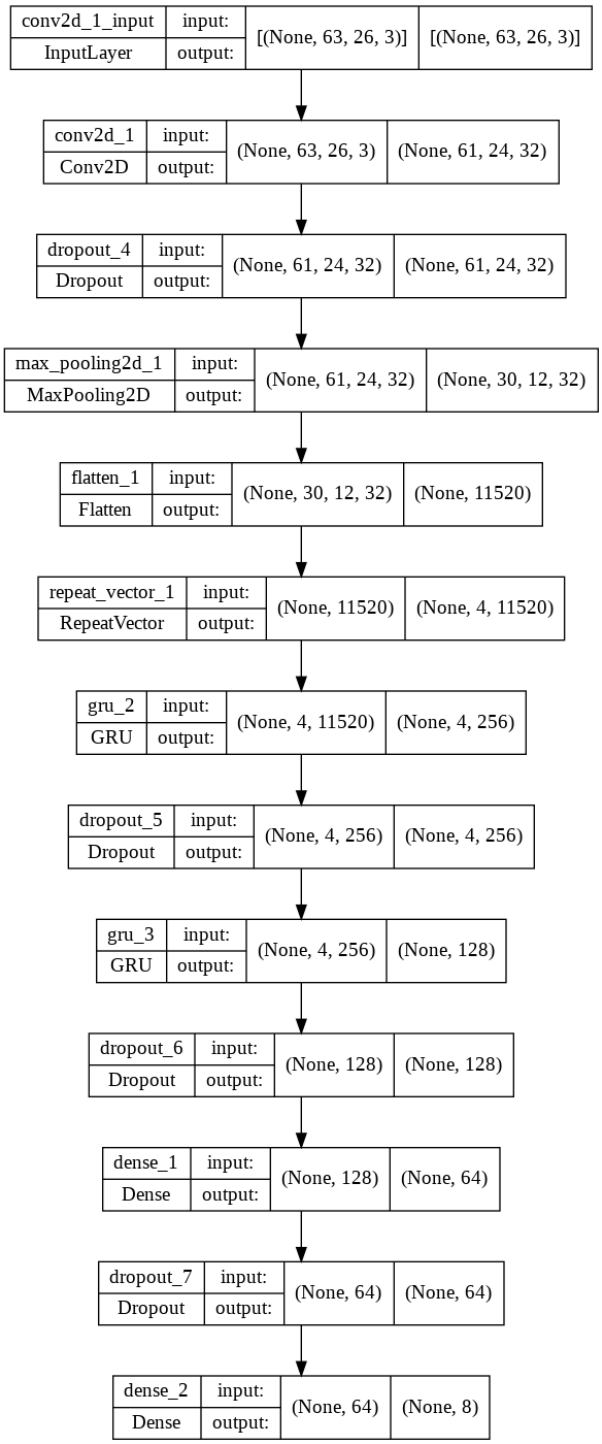


Fig. 2: Proposed Architecture of the Model

High Dominance marked as 6, and Low Valence Low Arousal Low Dominance marked as 7 are generated. VA space label classification we have used Valence and Arousal classes. Same as VAD space Valence and Arousal classes are divided into high and low with threshold value 5 and using Valence and Arousal combinations like High Valence High Arousal labelled as 0, High Valence Low Arousal labelled as 1, Low Valence High Arousal labelled as 2 and Low Valence Low Arousal labelled as 3. Regarding VD space Valence and Dominance classes are used and labels like High Valence High Dominance labelled as 0, High Valence Low Dominance labelled as 1, Low Valence High Dominance labelled as 2 and Low Valence Low Dominance labelled as 3 are generated with threshold value 5. For AD space Arousal and Dominance classes are used and labels like High Arousal High Dominance labelled as 0, High Arousal Low Dominance labelled as 1, Low Arousal High Dominance labelled as 2 and Low Arousal Low Dominance labelled as 3 are generated with threshold value 5. After mapping spectrogram features to particular sets of labels, a hybrid CNN_GRU classifier used to populate 25 accuracies with 5Repeated 5fold classification.

Since, unlike image data which have enormous amount of data to train on, our data is quite limited. Hence we used kfold method to evaluate our model and set k=5 with a repetition=5, a total of 25 times our model run on each classification tasks and we reported the mean accuracy achieved from it. Further, we compared the results of each classification tasks with each other based student's ttest.

3 Results

For Valence-Arousal based classification task our model achieved a mean accuracy of 96.93%. Also, for Valence-Arousal-Dominance based classification task, our model achieved a mean accuracy of 97.50%. These mean accuracies based on 25 different accuracies which model predicted on each 5-fold and 5-repeats on each type of task. Further F1-scores achieved for these tasks are 96.93% and 97.50% respectively.

For VAD and VA based classification task, confusion matrix shows that accuracies ranges between 96%-99% for each class which is similar to the overall mean accuracies (see Fig. 3a and Fig. 3b for confusion matrix for VAD and VA, respectively).

Epoch vs Accuracy graph and Epoch vs Loss graph shows a smooth convergence to an optimal point for both the classification tasks, which is desirable for a good fit. Please refer to the epoch vs accuracy graphs in figures 3c and 3d for VAD and VA classification tasks, respectively. In addition, the epoch vs loss graphs for VAD and VA classification tasks is shown in figures 3c and 3d, respectively.

Using t-test statistical testing, the 25 accuracies of VAD space with 8 labels (M = 97.50%, SD = 0.16%) compared with the 25 accuracies of VA space with 4 labels (M = 96.93%, SD = 0.38%), VAD space accuracies with 8 labels shows better results than VA space accuracies with 4 labels with t (32) = 6.68, $p <$ 0.0001, cohen's d=1.89 (large), 95 percent confidence interval=[1.20 2.57].

Table 1: Result Summary for two different classification analysis. VAD: classification in valence-arousal-dominance dimensions, VA: classification in valence-arousal dimension.

Metrics	**VAD**	**VA**
F1 Score	97.50%	96.93%
Accuracy	97.50%	96.93%

Result summary for different classification tasks based on all the three dimensions of Valence, Arousal and Dominance, is presented in table 1.

4 Discussion

In this analysis, we have performed emotion classification of emotional events recorded on Indian samples. The data we have used itself has a unique quality. For the first time, the precise information about emotional feelings was captured while subjects are provided with enough context to feel the naturalistic and real-life resembling emotions. This data is recorded in our lab. Hence, one of our aim in this analysis is to show that if the precise information about emotional feelings is captured, we don't need complicated models to perform emotion recognition. With the precise information, we mean that the emotional feelings are not mixed with the mind-wandering activity. We would have recorded the mind-wandering activity mixed with emotional feelings if we had taken the EEG recording for the whole stimulus length (which is a general trend in EEG based emotion experiments [1] [2] [3] [4])

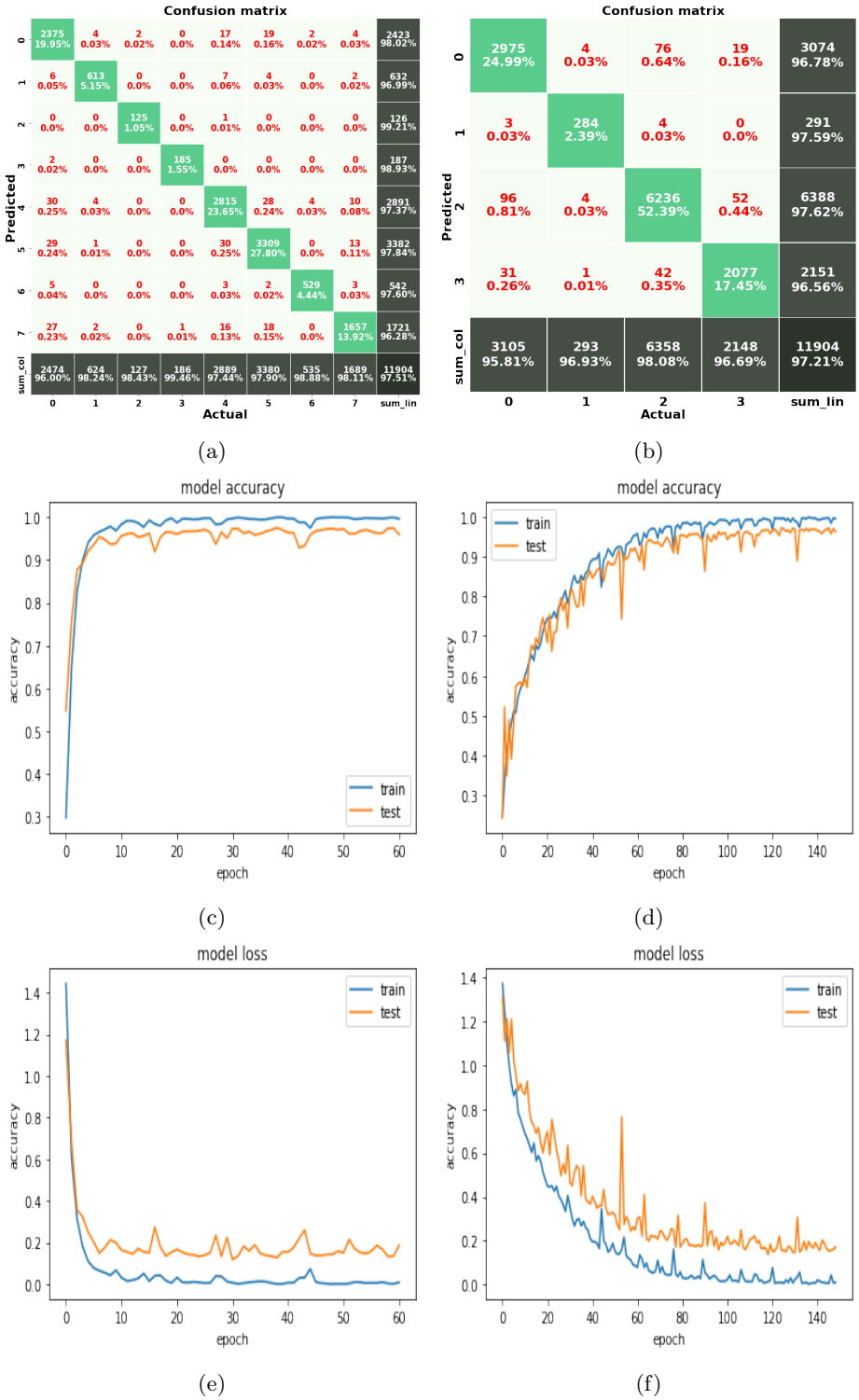
Another unique observation observed in our study is related to the importance of the dominance dimension. The literature is overwhelmingly filled with the classification of emotion in the quadrants taken in valence-arousal space. Our analysis found that including dominance dimensions provided significantly better accuracy than if only valence and arousal dimensions had been considered.

Generally, in literature other than valence and arousal, dominance is proposed as the third dimension [5] [6] but it is rarely considered in the classification tasks. The reason is that consideration of dominance as the dimension rarely explains more than 15% of the variance in subjective rating [5] hence it is still debatable to consider the dominance as the third dimension. However, it was acknowledged in the circumplex model of affect [7] that emotional episodes cannot be merely defined using core affect dimensions (valence and arousal) and there is a need to include third dimension for the attribution of cause and meta-cognitive judgement. For example, experientially, fear and anger are distinct emotional states. However, the two-dimensional model will project it as a high-arousal and low-valence state, which does not reflect the subjective experience of these emotions. In addition, time is also a contributing variable in the perception of emotion as emotions are dynamic in nature. The temporal dynamics of emotion perception is mainly dominated by subjective meta-cognitive evaluation of the situation [8].

Based on our results we believe that the essential information related to metacognitive evaluation in the emotional processing is contributed by the dominance dimension. We encourage future researchers to consider dominance an essential dimension while performing the emotion classification tasks.

References

1. Koelstra, S., Muhl, C., Soleymani, M., Lee, J., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A. & Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions On Affective Computing*. **3**, 18-31 (2011)
2. Zheng, W. & Lu, B. Investigating Critical Frequency Bands and Channels for EEG-based Emotion Recognition with Deep Neural Networks. *IEEE Transactions On Autonomous Mental Development*. **7**, 162-175 (2015)
3. Katsigiannis, S. & Ramzan, N. DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals From Wireless Low-cost Off-the-Shelf Devices. *IEEE Journal Of Biomedical And Health Informatics*. **22**, 98-107 (2018)
4. Miranda-Correa, J., Abadi, M., Sebe, N. & Patras, I. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions On Affective Computing*. **12**, 479-493 (2021)
5. Jerram, M., Lee, A., Negreira, A. & Gansler, D. The neural correlates of the dominance dimension of emotion. *Psychiatry Research: Neuroimaging*. **221**, 135-141 (2014)
6. Verma, G. & Tiwary, U. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage*. **102** pp. 162-172 (2014)
7. Barrett, L. & Russell, J. The structure of current affect: Controversies and emerging consensus. *Current Directions In Psychological Science*. **8**, 10-14 (1999)
8. Blascovich, J. & Mendes, W. Challenge and threat appraisals: The role of affective cues.. (Cambridge University Press,2000)
9. Mishra, S., Srinivasan, N. & Tiwary, U. Affective Film Dataset from India (AFDI): Creation and Validation with an Indian Sample. (PsyArXiv,2021)
10. Mishra, S., Srinivasan, N. & Tiwary, U. Cardiacndash;Brain Dynamics Depend on Context Familiarity and Their Interaction Predicts Experience of Emotional Arousal. *Brain Sciences*. **12** (2022), https://www.mdpi.com/2076-3425/12/6/702
11. Kehtarnavaz, N. CHAPTER 7 - Frequency Domain Processing. *Digital Signal Processing System Design (Second Edition)*. pp. 175-196 (2008), https://www.sciencedirect.com/science/article/pii/B9780123744906000076
12. Krishnan, S. 5 - Advanced analysis of biomedical signals. *Biomedical Signal Analysis For Connected Healthcare*. pp. 157-222 (2021), https://www.sciencedirect.com/science/article/pii/B9780128130865000037
13. Li, Z., Liu, F., Yang, W., Peng, S. & Zhou, J. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Transactions On Neural Networks And Learning Systems*. (2021)
14. O'Shea, K. & Nash, R. An introduction to convolutional neural networks. *ArXiv Preprint ArXiv:1511.08458*. (2015)
15. Golmohammadi, M., Ziyabari, S., Shah, V., Von Weltin, E., Campbell, C., Obeid, I. & Picone, J. Gated recurrent networks for seizure detection. *2017 IEEE Signal Processing In Medicine And Biology Symposium (SPMB)*. pp. 1-5 (2017)
16. Rana, R. Gated recurrent unit (GRU) for emotion classification from noisy speech. *ArXiv Preprint ArXiv:1612.07778*. (2016)



(a) (b) (c) (d) (e) (f)

Fig. 3: (a) Confusion Matrix for Valence-Arousal-Dominance Based Classification. (b) Confusion Matrix for Valence-Arousal Based Classification. (c) Accuracy graph for classification in valence-arousal-dominance dimension. (d) Accuracy graph for classification in valence-arousal dimension. (e) Model loss for classification in valence-arousal-dominance dimension. (f) Model loss for classification in valence-arousal dimension.