Biomedical Signal Processing and Control 72 (2022) 103361

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Capsule neural networks on spatio-temporal EEG frames for cross-subject emotion recognition

Gopal Chandra Jana *, Anshuman Sabath, Anupam Agrawal

Interactive Technologies & Multimedia Research Lab, Department of Information Technology, Indian Institute of Information Technology – Allahabad, Prayagraj 211015, UP, India

ARTICLE INFO

Keywords:
Spatio-temporal representation
Spatio-temporal EEG
CapsNet
EEG Emotion Recognition

ABSTRACT

Scalp EEG plots are plots of scalp potentials against time, and hence, capture spatial information, owing to the placement of electrodes on the scalp, as well as, temporal information from variations in brain waves. In this paper we propose a novel method to make a combined representation of spatial and temporal information, by incorporating the signals into a sparse spatio-temporal frame, such that it can be easily processed by deep learning algorithms in the computer vision domain. Familiarities of a model to the test data in the setting of emotion recognition from EEG, is also defined, and a form of data splitting such that the model has to perform on a set with which it has the minimum degree of familiarity is introduced. A CapsNet architecture is trained on DEAP dataset to perform on a cross-subject binary classification task, and tuning of the hyperparameters using Bayesian Optimization is analyzed. The proposed model reports a best-case accuracy of 0.85396 and average case accuracy of 0.57165 for LOO subject, and a best case of 1.0 and average case of 0.51071 for unseen-subject-unseen-record classification, when averaged across all the classes (i.e., valence, dominance, arousal, and liking), which is comparable to that reported by other works.

1. Introduction

Emotion Recognition using EEG signals is one of the rudimentary applications of Human Computer Interfaces (HCI). Emotion Recognition could be performed using a variety of sensing mechanisms, like, electroencephalogram (EEG), magnetoencephalogram (MEG), electrocardiogram (ECG), electroocoulogram (EOG), galvanic skin response (GSR), heart rate variability (HRV), respiratory rate (RR), skin temperature (SKT), etc. [1]. EEG signals could be collected either using invasive electrodes placed surgically on the brain to observe its activity, or may be non-invasively placed on the scalp. This type of positioning renders, to the data, both a spatial dimension owing to the arrangement of electrodes, and a temporal dimension owing to the time-variation of electrodes potentials. Moreover, EEG signals are highly non-linear [2]. Successful distinction of different emotions based on EEG signals from non-invasive sensors could enable application of HCI using EEG to a more diverse field in medical, technological and entertainment domains. Moreover, EEG signals recordings are widely studied in diagnosis of various other ailments clinically and the methods and tools developed could have a cross-domain application. One reason to focus on emotion recognition for developing such tools and methods could be the easy availability of experimental data compared to clinical data, which is often recorded for the diagnosis of a certain ailment, hence, in turn, could be confidential. This provides a strong motivation to pursue automated emotion recognition using EEG Signals.

Clear distinction of emotions using automatic emotion recognition is a complicated problem owing to the ambiguous boundaries between multiple emotions [1], when mapped to the measured signal domain. Hence, majority of studies using EEG for Emotion Recognition focus on a dimensional model for emotions, either considering Valence-Arousal (two-dimensional model) [3], or considering Valence-Dominance-Arousal (three-dimensional model) [4]. Depending on the emotional model chosen for the study, the different classes for classification are termed as High Valence-Low Valence (HV/LV), High Arousal-Low Arousal (HA/LA), and High Dominance-Low Dominance (HD/LD). The classifiers could be trained on binary classification task when considering the individual labels separately or multi-class classification considering all the labels at once.

Further, depending on how the classifiers are trained and tested, the classification problem could be intra-subject or an inter-subject. When a classification model is trained on a portion of data of a single subject treated as training set and tested on the data held out from the same

* Corresponding author.
E-mail addresses: go.gopal.ch.jana@gmail.com, rsi2018508@iiita.ac.in (G.C. Jana).

https://doi.org/10.1016/j.bspc.2021.103361
Received 11 June 2021; Received in revised form 8 October 2021; Accepted 7 November 2021
Available online 18 November 2021

person, it could be referred as an intra-subject study [5–7]; and when the classification model is trained on data from a set of subjects, but tested on data from a subject which it has not been trained on, the study could be termed as inter-subject or cross-subject [8–10].

In the present research scenario, a lot of works have adopted an intra-subject approach. The preference for this methodology could be attributed to the high variability of EEG data, from one experiment to another, which makes generalization across subjects even more challenging task. Moreover, the cross-subject or cross-dataset studies prefer to use binary classification for HV/LV classification. Thus, there is a motivation to undertake a cross-subject study that considers richer dimensionality of emotions.

Prior studies in EEG analysis using machine learning methods have developed multiple handcrafted features to improve the performance. With the advent of deep learning adoption in various fields, many recent studies have experimented with these techniques in automated EEG analysis domain. However, the common deep learning frameworks, like Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN), could, at a time, capture only either the spatial content or the temporal content of the signals without intricate handcrafted features. Hence, we have explored the training of a relatively newer deep learning algorithm, Capsule Networks (CapsNet) [11], on a form of data representation that combines both spatial and temporal features.

The rest of the paper is arranged into a review of different studies done in this field, followed by a description of materials and methods used for experimentation. Then we present the results and our analysis and comparison with other works in the similar domain. Finally, we conclude discussing the limitations and scope for future work. The main contributions of this study are as follows.

1.1. Main contributions

We have listed the following contributions that this work has made in terms of novelty. In this work, we apply and test the efficacy of:

- A spatio-temporal frame group (STF Group) rearrangement of EEG recordings, as a time-domain stack of 2D EEG spatial samples. These samples are obtained by arranging the corresponding recordings snapshot from all the channels into a sparse matrix representative of actual electrode positions on the scalp corresponding to the prescribed positions in International 10–20 system.
- Dataset split into testing and cross-validation sets to mirror real-world application setting where the model is expected to perform on previously unseen data, unseen subjects and unseen stimulus (video) to subjects.
- Use of STF group in 2D convolution algorithms, by passing the 2D frames concatenated along the time domain as the channels for the input layer without any further processing, and achieve comparable performance on binary classification task with other representation methods.
- Applying and tuning CapsNet architecture for the binary classification task of emotion recognition under completely unseen data to obtain detection accuracy comparable to state-of-the-art models.

2. Related work

Research on EEG recordings based emotion recognition saw its inception with Davidson et al. [12], and has ever since expanded into an active research area. A notable development in the task of emotion recognition has been the enunciation of 2D and 3D model of emotions [13], which has found wide use in the field. Present day studies into the field explore both Machine Learning and Deep Learning approaches for emotion recognition and classification. We focus on the developments made in Computer Vision based Deep Learning approaches in general, and specifically the cross-subject studies.

2.1. Developments in deep learning methods applied to emotion recognition

Applying Deep Learning methods to Emotion Recognition task had a focus on CNN based methods, where many researchers have attempted various EEG signal preprocessing mechanisms [14,15,7]. C. Cheng et al. [14] proposed Emotion Recognition Algorithm based on CNN (ERACNN), where they fed a (channel × down-sampled signal × time of video) into CNN and tuned the network learning rate, weight initialization function, and momentum to optimize for accuracy. H.Mei and X. Xu [15] construct a Pearson Correlation Coefficient (PCC) Matrix for each sub-band, assuming the 32-channels as nodes of EEG-based brain functional networks. The individual PCC Matrices for sub-bands concatenated along the channel dimension becomes the input for CNN, which outperformed SVM and GELM on 2-class, 3-class and 4-class classification. While in [14] the spatial arrangement of electrodes is ignored, in [15], the authors do not include the temporal information in EEG recordings in any direct form in their features.

The caveats from the [14] and [15] have been addressed by Jungchan Cho et al. [7], where they have used a 3D EEG stream formed by stacking 2D EEG matrices to train 3D CNNs. Authors have mentioned that the 2D EEG matrices were obtained by rearranging the channel data according to actual channel position on scalp, and then interpolation has been done for missing channels. This work achieved classification accuracy of 99.74%, 99.11%, and 99.73% in the binary classification of arousal and valence, and, in four-class classification, respectively.

Few recent works have resorted to using Capsule Networks instead of Convolutional Neural Networks [5,6,8]. J. Guo et al. [5] have used a CapsNet model to classify emotions based on Granger Causality Matrices between channels computed after decomposing the EEG signals into alpha, beta and gamma bands by wavelet transform. This approach, however, completely excludes the information in the spatial arrangement of electrodes. Yu Liua et al. [6] have used a multi-level features primary capsule that combines features learned from multiple layers. The Capsule Network is trained subject-wise on matrices of (channel number × segment size) on both DEAP and DREAMER datasets to classify Valence, Arousal and Dominance. The scores reported for DEAP dataset with the proposed approach is lower than normal CapsNet for Valence and Dominance. The (channel number × segment size) arrangement does not contain any original electrode arrangement information, nor does it definitively ensure incorporation of temporal features.

Hao Chao et al. [8], composed a multiband feature matrix (MFM) by separating the parent EEG signal into different bands (α, β, γ, δ) and then, arranging the channel information into a 18x18 matrix (MFM), 4 matrices each for different band produced by mapping actual channel positions to a matrix. Duration of the sliding window was set to 3 s, while splicing the 60 s videos into segments. Individual samples formed from these sections, and inherited the labels of the original sample. In this experiment, CapsNet has been trained on these MFMs with 10-fold cross-validation on the combined data. The caveat in this approach is that the MFMs prepared do not contain any temporal information.

Bao G et al. [18] separated the EEG into 5 bands (Delta, Theta, Alpha, Beta, Gamma), and computed topology preserving differential entropy (TP-DE) features. The TPDE features were extracted using CNN and then passed through two-level domain adaptation neural network (TDANN) for classification. In [23], researches applied multivariate convolution network on a 3D feature matrix composed of time-domain features, and the channel recordings arranged in the 2D-plane according to the electrode positions. The performance of this approach was evaluated using 10-fold CV on DEAP dataset. While these works included both spatial and temporal information in the EEG signals, the cross-validation tests performed were not cross-subject.

Among machine learning algorithms applied to the task of emotion recognition, SVM and Naïve Bayes have been explored. Harsh Dabas et al. [17] have used the DEAP dataset to test the proposed 3

dimensional model of emotion without any specific feature rearrangement or preprocessing beyond what was provided by DEAP. In this experiment various machine learning algorithms like SVM, Naïve Bayes, have been applied to the preprocessed EEG DEAP data set.

2.2. Cross-subject studies for EEG emotion recognition

All the works discussed above do not employ any mechanism to separate their train and test sets in a way that would represent a real-world scenario, where a model trained from a couple of subjects needs to perform reliably on other subjects. To develop intuition on model performance in this scenario, the following works have carried out a cross-subject testing of methods and models.

There has been an interest in using feature selection from the high dimensional EEG signals for cross-subject emotion recognition. Fu Yang et al. [19] proposed a method for cross-subject emotion recognition based on multiple features that were extracted in order to form high-dimensional features by integrating the significance test/sequential backward selection and the support vector machine (ST-SBSSVM). The proposed ST-SBSSVM has been trained on DEAP [20] and Shanghai Jiao Tong University Emotion EEG Dataset (SEED) [21] datasets. Authors in this work have claimed that their model achieved an accuracy score of 0.72 for Valence classification and 0.89 respectively using leave-one-subject-out validation. A Decision Tree Classifier based on Sequential Backward Selection was used on PSD features by W. Jiang et al. [30] and tested on DEAP and self-made photograph EEG (PEEG). They achieved 65.8% on PEEG and 65.2% on DEAP for valence recognition. Researchers in [31] employed various ML models upon features selected via cross-subject Recursive Feature Elimination (RFE). The tests were run on DEAP and MAHNOB-HCI on Arousal and Valence classification adopting n-fold CV method, where n is the number of subjects. Highest accuracy achieved on DEAP was by ANN: 0.6461 for arousal and 0.6529 for valence classification.

Approaches for extracting features from the EEG signals to either reduce the dimensionality or obtain specific signal properties by using various transforms have also been adopted. Li X et al. [25] extracted nine time–frequency domain features and nine dynamical system features to train a SVM using "leave one out" verification strategy. The authors claimed that proposed model has achieved a highest average emotion recognition accuracy of 83.33% (AUC = 0.904) on the SEED dataset [21] and of 59.06% (AUC = 0.605) on the DEAP [20] dataset. V. Gupta et al. [28] extracted features by applying Information Potential (IP) to EEG sub bands obtained by Flexible Analytic Wavelet Transform on EEG signals. Random Forests and SVMs were used for prediction to obtain 90.48% on SEED, and 79.99% on Arousal classification, 79.95% on Valence classification, and 71.43% on HVHA/HVLA/LVLA/LVHA on DEAP for leave one out classification task.

Researchers in [24] employ Empirical Mode Decomposition (EMD) and Variational Mode Decomposition (VMD) for feature extraction to obtain intrinsic mode functions (IMF). A 3-layered Deep Neural Network was used for Classification based on IMFs. The evaluation was performed on DEAP dataset, by training on 30 subjects and testing on 2 subjects. The highest accuracy of 61.25% for Arousal and 62.50% for Valence was obtained for VMD based feature extraction. Samarth et al. [16] divided the entire signal train of each channel into 10 segments and computed several statistical features for the segments as well as the train overall. They use a matrix of (channel × statistical features) to feed into the CNN while the DNN takes the same matrix in flattened format. The CNN model outperformed the Deep Neural Network (DNN) model. Fdez, J. et al. [29] used stratified normalization, which employs a participant-based feature normalization to subtract inter-participant variability from EEG features extracted using Welch's method, multitaper and differential entropy, while retaining emotion information in their study for classification of emotions using neural networks. They obtained 91.6% accuracy on Positive-Negative and 79.6% on Positive-Negative-Neutral classification with leave-one-out validation on SEED dataset.

Table 1
A review of state-of-the-art methods in emotion recognition on DEAP.

Sl. No.	Work	Description
1	P. Pandey and K. R. Seeja [24]	- Use EMD and VMD for feature Extraction to obtain intrinsic mode functions (IMF) (EMD = Empirical Mode Decomposition, VMD = Variational Mode Decomposition) - **Pros**: Performed learning on selected channels. They had a light model which was only 3 layers deep, and performed better than more deeper and wider models - **Cons**: Use handcrafted features from EEG. Do not exploit the topology of electrode placements. Although testing is cross-subject, the leave one out validation is not strictly followed. Using selected channels make the method highly specific for emotion recognition task.
2	Y. Cimtay, E. Ekmekcioglu and S. Caglar-Ozhan [27]	- Use multimodal approach including facial expression, GSR, and EEG using different CNNs for each mode and training a decision tree to output the final result - **Pros**: Model has the ability to detect actual emotional state when it is dominant or hidden. Introduce a time-delay parameter between EEG and other signals – which is empirically determined. Used features extracted automatically via CNN. - **Cons**: Use multiple input instruments for different modalities. The time-delay introduced is empirically determined by comparing against accuracies obtained for different delays. Topology of electrodes was not considered in the EEG feature matrix
3	V. Gupta, M. D. Chopda and R. B. Pachori [28]	- Features from selected channels were extracted by applying Information Potential (IP) to EEG sub bands obtained by Flexible Analytic Wavelet Transform on EEG signals. Random Forests and SVMs were used for prediction - **-Pros:** Results verified on multiple datasets. Analysis of classification based on each channel - - **Cons:** Used handcrafted features for the emotion recognitions task. The information in electrode topology is not considered. Using selected channels make the method highly specific for emotion recognition task.
4	W. Jiang et al [30]	- Decision Tree Classifier based on Sequential Backward Selection was used on PSD features from DEAP and self-made photograph EEG(PEEG) - **Pros:** Used a self-made dataset (PEEG). - **Cons:** Used manually determined PSD features and relied on feature selection. Did not account for the topology information in electrode placements.
5	W. Zhang and Z. Yin [31]	- Various ML models are employed upon features selected via cross-subject Recursive Feature Elimination from numerous handcrafted features from both time and frequency domain. - **Pros:** Testing on different Machine Learning Models was performed. A rich handcrafted feature representation was obtained combining both frequency and time domain features upon which RFE was performed. Used two datasets for testing - **Cons:** Use of hand-crafted features for preparing the initial feature vectors. The features considered do not exploit the topological arrangement of electrodes.
6	Pandey, P., Seeja, K.R [33]	- A 12-layered CNN was used on scalogram images obtained by applying continuous

(continued on next page)

Table 1 (*continued*)

Sl. No.	Work	Description
		wavelet transform (CWT) to perform cross-subject and cross-dataset inference - **Pros:** Rely on the CNN to perform the entire feature engineering, hence the feature engineering part is also trainable. Cross-dataset testing also done. Comparison between using selected electrodes and all electrodes was also made. - **Cons:** The scalograms generated do not accommodate any information about the topology of electrodes.
7	J. Liu et al [47]	- Subject clustering based domain adaptation, where subjects are grouped together based on similarity of emotion-specific EEG response. A feedforward Neural Network is trained for inference - **Pros:** Clustering of subjects on the basis of emotion-specific EEG response gives a more logical discrimination method. - **Cons:** The features employed were manually extracted. The features did not account for the topology of electrode arrangement
8	Yingdong Wang et al. [48]	- Generate a source subject adapted list, and then train three models (MLP models) on each selected source subject and target subject. Target emotion label is obtained by distilling the classifiers - **Pros:** Performs cross-subject inference with limited target data using the adversarial domain adaptation. Use MLP models for feature extraction - **Cons:** The feature spaces did not consider the topological arrangement of electrodes in EEG montages.
9	Zhen Liang et al. [49]	- A hybrid model fusing CNN, RNN and GAN is proposed to extract EEG features and fuse them in an unsupervised learning scenario - **Pros:** Features extracted using unsupervised deep learning. Developed a mechanism to fuse features from CNN, RNN and GAN and verified the model with Leave One Out CV - **Cons:** The feature maps fed into the model, did not take into account the topology of electrode arrangements. Also, fusing three deep-learning models could make the overall model complex and computation heavy.
10	Arjun et al. [50]	- The authors utilize an unsupervised LSTM with channel-attention autoencoder to get a latent space feature representation of EEG signals on which a CNN with attention is trained for classification. - **Pros:** They have used unsupervised feature extraction methods, and tested the method on multiple datasets covering different application domains. - **Cons:** Although CNN is used, the preceding LSTM which is used to generate the latent space features ignores the topological arrangement of electrodes.

In [33], a 12-layered CNN was used on scalogram images obtained by applying continuous wavelet transform (CWT) to EEG signals, and study was conducted on both DEAP and SEED datasets. For DEAP highest Valence Accuracy obtained was 61.5, and Arousal Accuracy was 58.5 for 10 frontal electrodes; and 59.5 and 58 for all 32 electrodes, respectively.

All these works, however, did not take into account the spatial arrangement of electrodes while formatting the feature matrices in any form, and instead rely on manually extracted features. These features are either fed directly into a Machine Learning model, or are arranged in a 2D matrix form when the authors have chosen to use some variant of CNNs. Also, using a preprocessing step to produce features before feeding it into a deep learning model as used in [16,24,29,33] does not take advantage of feature extraction capability of deep learning methods to the fullest.

Treating EEG signals originating from different subjects to be having different domains, several domain adaptation techniques have also been adopted. He Li et al. [9] used Deep Adaptation Networks (DAN) to address the subject transfer problem in the domain of EEG-based emotion recognition on SEED [21] and SEED-IV [21]. The authors claimed that the proposed model achieved the maximum mean accuracy of 0.8381 on SEED dataset, and that of 0.5887 on SEED-IV. J. Li et al. [10] proposed a Transfer Learning approach to explore and exploit the models trained on existing subjects, by making the new subject data statistically similar to the sources previously seen with the use of style transfer mapping (STM). In this experiment authors have used SVM classifier to perform classification task on SEED [21] dataset. Maximum mean accuracy reported by the authors is 91.31% on MS-Semi-STM and 88.92% on MS-S-STM. Deep Domain Confusion (DDC) was used in [26] to minimize the difference between source and target domain feature distributions, which were obtained by passing Electrode Frequency Distribution Maps (EFDMs) constructed from EEG signals through Residual Blocks. Researchers reported 82.16%/4.43% for mean accuracy and standard deviation for cross-subject inference task on SEED dataset.

A number of works have also relied solely on deep learning based feature extraction and training for cross-subject emotion recognition. A multimodal approach based deep learning architecture was used in [27] wherein facial expression, GSR, and EEG were used as the different modes. On LUMED-2 for 3 emotion classes: sad, neutral and happy, the maximum one subject out accuracy reported was 81.2% and mean one subject out accuracy reported was 74.2%. On DEAP, maximum one subject out accuracy reported was 91.5% mean one subject out accuracy reported was 53.8%. Yucel Cimtay et al. [46] used raw EEG data from three different publicly available datasets, DEAP [20], SEED [21] and Loughborough University Multimodal Emotion Dataset (LUMED) [22] after applying windowing with N/6 overlap, pre-adjustments and normalization with employing a median filter to eliminate the false detections along a prediction interval of emotions. No manual features are extracted, and validation is done by leave-one-subject-out. Authors claimed that they obtain mean cross-subject accuracy of 86.56% (for two class) and 78.34 % (for three class) on the SEED [21] dataset, 72.81% on DEAP [20] dataset and 81.8% on the (LUMED) [22] for two emotion classes. Researchers have also used an adversarial neural network to train and test on SEED in one-subject-out fashion to obtain average classification accuracy of 75.31% with standard deviation of 7.33% [32].

Table 1 included below summarizes the recent state of the art models in emotion recognition performed on DEAP dataset and analyze the techniques employed and their shortcomings in the context of present proposed work.

The review of recent works makes it clear that the studies in cross-subject emotion recognition have not integrated the spatial arrangement of electrodes on scalp, along with the temporal information in EEG signal trains while creating their feature matrices. Using the spatial arrangement has yielded good results in some studies which do not explicitly perform a cross-subject inference. Many works that do adopt a spatial rearrangement (applied to intra-subject recognition), either discard the temporal features, or use manual feature extraction methods. Also, there are no studies on cross-subject emotion recognition that have produced results excluding both degrees of familiarity to the data that a model can have (refer 3.2.2) by performing classification on unseen-records from an unseen-subject.

This provides a motivation for this work to use a completely Deep Learning based model for cross-subject emotion recognition using feature maps that embody both spatial arrangement of electrodes, and the temporal features of EEG signal train. CapsNet is chosen as the deep

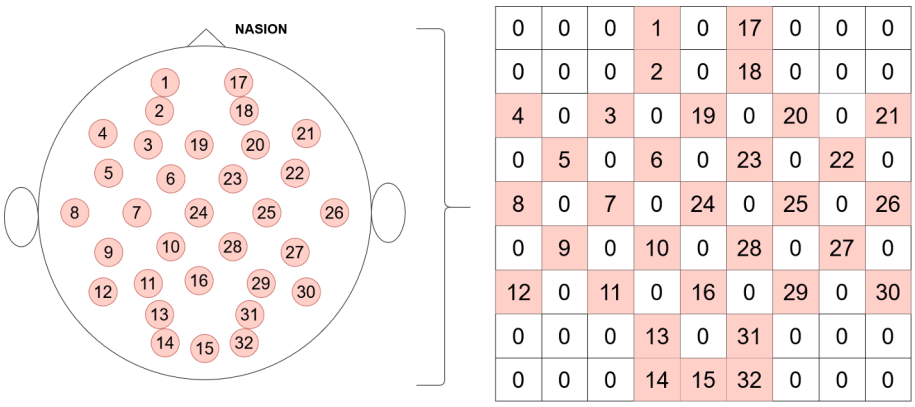

Fig. 1. Mapping Scalp Electrode Positions to a Matrix.

learning model, since, it has been claimed to perform better on sparse matrices and have better 3D pose inference. Using CapsNet, a baseline is provided where classification is performed on unseen-records from the left-out subject, which would serve as a better estimate of how the model will perform in a more realistic application setting.

3. Materials and methods

In this section we have described about the experimental data and methods using three subsections. Subsection-3.1 describes about the experimental dataset, Subsection-3.2 describes the methods that have been used for data preparation, and Subsection-3.3 describes the methods that have been used to perform classification.

3.1. Experimental dataset

In this study, we have used DEAP [20] dataset to estimate the performance of our proposed approach. DEAP [20] (Database for Emotion Analysis using Physiological Signals) dataset contain physiological signals that are interlinked with valence, arousal, dominance, and liking emotion states for the different trials and different subjects. DEAP [20] dataset contains EEG and other peripheral physiological signals. These signals were recorded using the standard 10–20 notation placement of electrodes on scalp, when the subjects were watching clips of music videos having lengths of 60 s each. Each of the 40 such different videos formed a single trial, and this set of trials was repeated over a total of 32 participants as subjects for the study. The subjects rated these videos on different levels of emotional dimensions namely, valence, arousal, dominance, and liking. The readings from 32 scalp EEG electrodes were recorded at a sampling rate of 512 Hz. Recorded signals were downsampled to 128 Hz; to sample the data correctly a filter was applied to extract the data signals between 4.0 Hz and 45.0 Hz.

After the application of above steps, the data finally formed into the matrix of [40 × 40 × 8064] (40 video/trails × 40 channels (described above) × 8064 signal points/data). The signal data of 8064 was obtained for 63 s of video and 128 Hz downsampled frequency (128 Hz × 63 s = 8064 samples/channel). All the involved channels would be generating the same data range. The labels are derived from the continuous ratings that range from 1 to 9, which were provided by the participants (subjects) after each trial for the different parameters, valence, arousal, dominance, and liking. The matrix of the labels is of the dimension 40 × 4 (40 trials × 4 labels (Valence, Arousal, Dominance, Liking)).

3.2. Methods for data preparation

We perform some data preparation to increase its compatibility with the class of algorithms studied, which has been described in this section with the help of two subsections. Subsection-3.2.1 described about the data preprocessing and augmentation process and Subsection-3.2.2 described the data splitting process which has been adopted for special arrangement of the data to perform cross-subject validation and testing.

3.2.1. Data preprocessing and augmentation

In this study we utilize both temporal and spatial information present in the recordings, hence we rearrange the EEG data into spatio-temporal frame groups (STF groups). The data downloaded from the dataset source had shape of [40 × 40 × 8064] for [trials × channels × samples], however, of the 40 channels, only 32 channels were from the EEG electrodes, while other 8 channels recorded different physiological signals. Also, the labels were continuous for each emotion, which needed to be converted to binary labels. So, before rearrangement, we remove these last 8 channels. The 8064 samples represent sampling at 128 Hz for 60 s of video and 3 s of baseline signal before the beginning of the trial $\left\{B_1, B_2, B_3, \hat{S_1}, \cdots, \hat{S_{60}}\right\}$. Hence, we take the sample-wise average of the baseline signals and subtract the 1 s duration signal so obtained $((B_1 + B_2 + B_3)/3)$ from the remaining duration of signals sample-wise to yield data, $\left\{S_1 \cdots, S_{60}\right\}$, which has a shape of [40 × 32 × 7680]. We then apply a z-score normalization to data so obtained, by subtracting the subject-wise mean and dividing the subject-wise standard deviation to obtain the final working set for signals, $\{S_1 \cdots, S_{60}\}$. The labels are converted to a binary representation, by applying a fixed threshold mid-way between the extremes (i.e., 4.5).

In literature deep learning techniques have been used over EEG signals, where CNN based approach [34,35] has been used over spatial features and LSTM based approach [36,37] has been used over temporal features of EEG signals. Since EEG recordings generally includes both spatial and temporal features, any of the aforementioned techniques used in an isolated way could not capture the entirety of the information. Hence, some works have also used some combination of both. For models derived from the computer vision domain, the data needs to be reshaped into a 2D matrix format. Traditionally, this has been achieved by stacking recording segments from different channels into a frame [38,14,5].

A defect in such rearrangement of EEG data is that the spatial information present in the original data is lost. The 10–20 electrode placement system, being an arrangement of electrodes on the surface, contains spatial information of the distributions of brain's potential. Since, computer vision techniques specialize in inference of spatial features, using a representation as above would undermine the efficacy of these algorithms. Hence, in recent literature [7,8,39], a 2D matrix mapping the positions of electrodes has been used, to create spatial frames against each sample as illustrated in Fig. 1. The mappings of the electrode numbers to standard 10–20 electrode names can be found in *Appendix-1*.

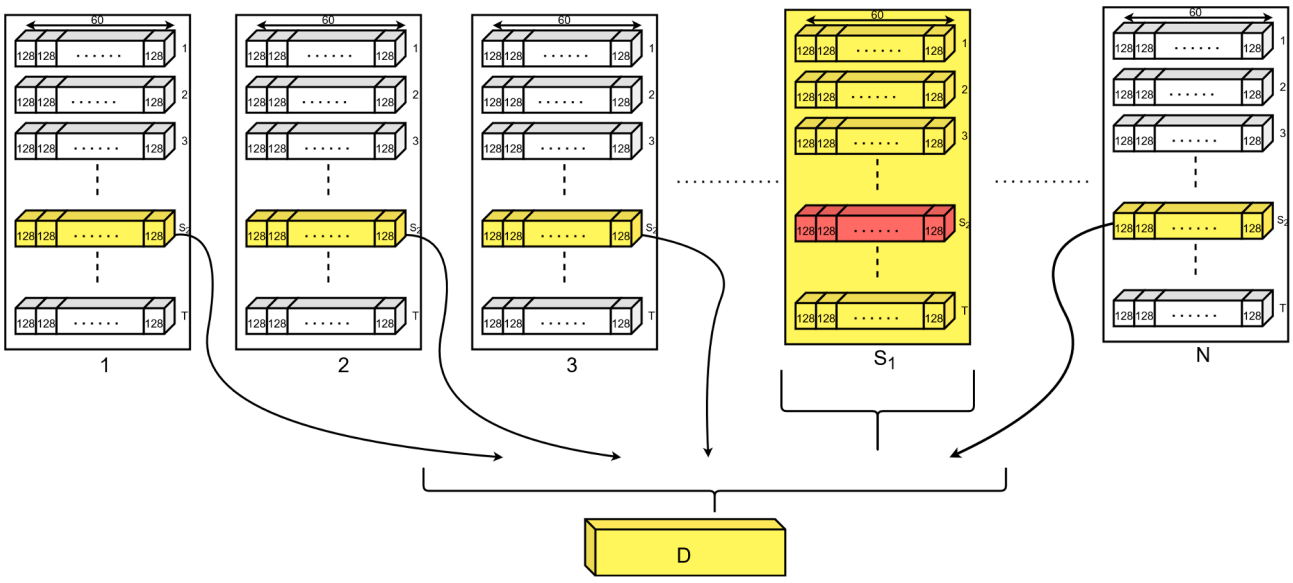

Fig. 2. Illustration of data split method.

Though the 2D matrix representation could incorporate the spatial relations of the signals, the temporal dimension is not available. To create a unified representation that would incorporate both spatial and temporal data into a single entity, we stack these 2D matrices or spatial frames along the third dimension to create spatio-temporal frame (STF).

Stacking the spatial frames in the third dimension encodes the temporal information into what would be the channel dimension of a normal RGB image, and hence it can be used by other computer vision algorithms without any further rearrangement being necessary. The number of spatial frames that were stacked in STFs were chosen so that each STF would cover a time duration of 1 s which has been established as the appropriate signal duration for EEG analysis in previous studies [40]. Since the sampling frequency is 128 Hz, we get 128 frames in each STF of shape [$9 \times 9 \times 128$].

3.2.2. *Data split*

When developing algorithms with the motivation for use in the field of medical diagnosis, it is desirable that the algorithms generalize well enough to an unseen data point. Ideally this unseen data should be in the form of a subject never encountered by the algorithm, which would be what the trained algorithms are supposed to do in a practical setting. Thus, we adopt a data split that would retain unknown data for both cross-validation and test sets.

Since the DEAP [20] data has 32 subjects with 40 videos each, there are two levels of familiarization the model could have when a conventional K-fold cross-validation or a randomized train-test split is used. If all the data samples are randomly mixed, there is a chance that a sample in the testing or cross-validation set belongs to:

(i) The same subject as in training set
(ii) The same video file or trial as in training set

In order to train a model which can generalize to completely unseen data, it would be required to have a set where the algorithm has neither been trained on any data from the unknown subject, nor on any data from any known subject recorded using the unknown video. While this kind of generalization is ideal, practically it is expected to be sufficient with the algorithm generalizing to an unknown subject.

It must be noted that in studies with subject-wise training and testing, with an appropriate split (cross-trial split), familiarity (ii) could be eradicated, but in order to eliminate familiarity (i), cross-subject split is necessary. However, in this work, we adopt a mixture of the splits to make the testing and cross-validation sets, so that the final performance evaluation metric is representative of both.

To achieve a mixed data split representative of a real-world scenario where model has to perform on unseen subjects, we choose a random subject (S1) from all the subjects (N) and a random trial (S2) from all the trials (T), separate all the samples from the selected subject, and the samples from the selected trial across all subjects into test set. We repeat the same process for the remaining subjects (now, N-1) and trials (now,

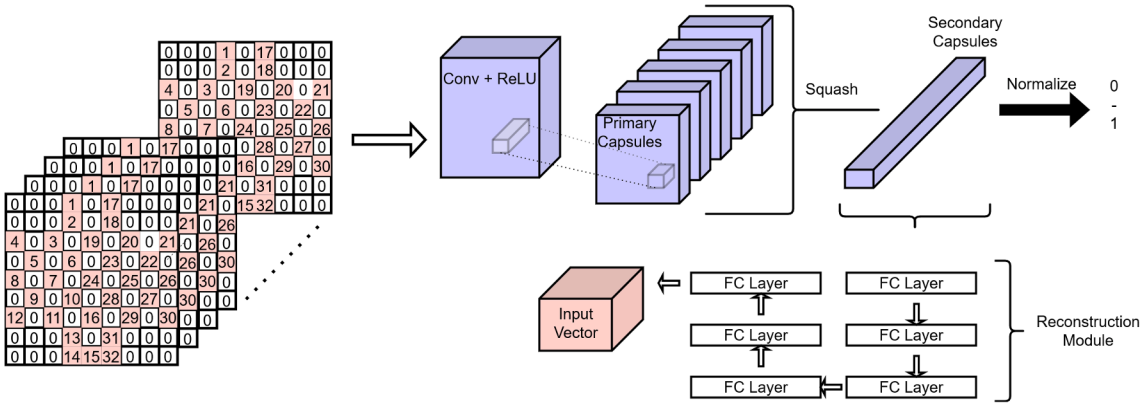

Fig. 3. Capsule Network Architecture.

T-1) to obtain the cross-validation set. The algorithm (Fig. 2) for splitting is as stated below:

A1 S1 = random subject from N (remaining) subjects;
S2 = random trial from T (remaining) trials
A2 Add all samples of S1 into a different set D;
A3 Iterate through N-1 remaining subjects and accumulate all samples corresponding to S2 trial;
A4 Add all such accumulated samples to the set D;

3.3. Methods and Models for Classification

In this section we have discussed about the classification models which have been used in the experiment. Our proposed approach based on Capsule Network (described in subsection 3.3.1) and the performance of the proposed approach has been compared with the performance of the Convolution Neural Network and Residual Neural Network.

Convolutional Neural Networks and Residual Networks are two popular network architectures in Deep Learning for processing spatial patterns in images. However, both these methods rely on the technique of pooling. The CapsNet architecture is different from these widely used models in that it does not utilize pooling. Hence, we describe the CapsNet architecture in detail in the subsection below.

3.3.1. Capsule network

Capsule Network consists of a vector of neurons rather than consisting of single neurons nodes as compared to traditional neural networks [11]. Hence, they contain more information about the input as compared to the information present in neurons in a traditional neural network [11]. Capsule Network (Fig. 3) consists of Convolutional Layer, Primary Capsule Layer, Secondary Capsule Layer (Emotion Capsules in our case) and Fully Connected Layers. The first three layers are called Encoder and the Fully Connected Layers are called Decoder. Capsule network trains the network model by dynamic routing method [11].

Capsules are group of neurons [11]. Various parameters are represented by the activity of these neurons and the probability of the existence of a particular entity are represented by the length of these vectors. The main drawback of CNN models are their pooling layers [11]. These pooling layers reduce the dimensionality of the feature matrix thereby losing most of the information about the input.

In Capsule networks tackle this loss of information in pooling layers by replacing this methodology with Routing by Agreement [11]. Based on these criteria all the parent capsule layers in the next layer receive the output, however their coupling coefficient are not the same [11]. The output of the parent capsule is predicted by each capsule and if the prediction is same as the actual output then the coupling coefficient between the two capsules increases. If u_i is considered the output of capsule, then the prediction for parent capsule j is calculated as,

$\widehat{u}_{j\vee i} = W_{ij}u_i$ (2)where, $\widehat{u}_{j\vee i}$ is the prediction vector of output of j th capsule in higher layer computed by i th capsule in lower level, W_{ij} is the weight matrix.

Coupling coefficients c_{ij} are calculated based on the degree of conformation between the lower level capsules and the parent capsules, using the SoftMax activation function,

$c_{ij} = \frac{e^{b_{ij}}}{\sum_k e^{b_{ik}}}$ (3)where, b_{ij} represents the probability whether the capsule i will be coupled to capsule j, k being all such capsules in the subsequent layer. This is initially set to 0, at the beginning of routing by agreement. Hence the parent capsule input vector can be calculated as:

$s_j = \sum_i c_{ij}\widehat{u}_{j\vee i}$ (4)

At last the nonlinear squashing function is applied to prevent the output vector of capsules from exceeding one as it represents the probability and probability is never greater than 1. The final output of each Capsule is given by:

$v_j = \frac{\|s_j\|^2}{1+\|s_j\|^2}\frac{s_j}{\|s_j\|}$ (5)where, s_j is the Capsule j input vector and v_j is its output.

To enable the capsules to learn a better representation, a decoder network is also coupled to the Capsule layer, which takes the outputs of the capsules, and tries to reconstruct the input. The training loss involves a component both from the reconstruction loss and the classification loss.

3.4. Experimentation

In this section, we describe our approach to experimentation with the aforementioned spatiotemporal rearrangement of EEG data with the different classification algorithms. In subsection 3.4.1, we discuss the reason for our choice of experiment algorithms, and in subsection 3.4.2 we describe the methodology adopted to reach at the results discussed in the subsequent section.

From Table 1 it was evident that many works in cross-subject inference have resorted to using hand-crafted features. This excludes the feature extraction method to be trainable and adapt according to the problem. Those that do use deep-learning features, ignore the topology of electrode arrangement. To fill this gap in research we incorporate the spatial information from topological arrangement of electrodes using the spatio-temporal rearrangement of EEG recordings as described in section 3.2.1. However, such arrangement leads to a sparse feature matrix. In order to exploit deep learning based feature extraction on this sparse feature matrix, we employ a Capsule Network architecture.

3.4.1. Why CapsNet based proposed approach has been compared with CNN and ResNet?

With the novel representation of EEG data formulated in this paper, we seek to investigate the efficacy of different deep learning algorithms on the data. We choose Convolutional Neural Network as the baseline model for the performance, and compare and contrast the performance of CNN with ResNets, which have a deeper architecture; and CapsNet (Refer Table 6), which has an architecture that forgoes the pooling layers used in the previous two.

The main advantage of using Capsule networks over Convolutional Neural Networks and Residual Networks are mentioned as follows as per [11,41]:

- CapsNets are better adapted for sparse image pattern recognition
- CapsNets are better adapted for 3D pose inference

3.4.2. Experimental details

In this study, experiments were carried out with the focus of testing the inference of capsule networks on a cross-subject dataset that would mimic the real-world scenario of model performing on unseen subjects, with a rich spatiotemporal feature representation. Hence, the spatiotemporal feature frames generated were fed into the selected classification algorithms. All the algorithms were trained from scratch.

The CapsNet was trained on the input 3D frames using 2D convolutions along the spatial dimensions with temporal features being combined in an early fusion [42] on the custom train set (Refer Fig. 3). The training was done using a single GPU on 4-compute node having two NVIDIA TESLA V-100(16 GB) GPGPU of HPC Cluster in Central Computing Facility, IIITA. We choose to use 2D convolutions in place of 3D convolutions used in [7], since numerous previous works [6,9,10,19,40] have carried out the task of EEG inference treating each sample point in the 1 s frame as a separate feature. Moreover, 2D convolutions considerably reduce the number of trainable parameters, which was a major limitation of the said work [7].

In order to tune the network a number of experiments were performed with different number of primary and secondary capsules. Each configuration was trained for 20 epochs in batches of 10 STFs and the performance was tuned by minimizing the loss obtained on a separate

Table 2
Best and Average Case Accuracies for CapsNet.

Labels	Arousal		Dominance		Valence		Liking	
Cases Accuracies	LOO Subject	LOO Record	LOO Subject	LOO Record	LOO Subject	LOO Record	LOO Subject	LOO Record
BestCase	84.249	84.167	100.00	90.625	63.042	78.906	94.292	84.375
Average Case	58.525 ± 15.181%	54.002 ± 19.001%	60.966 ± 16.645%	62.358 ± 19.059%	48.219 ± 7.832%	39.937 ± 23.097%	60.951 ± 18.305%	56.940 ± 23.125%

cross-validation set. The models were tuned for optimal combination of length of primary capsules, depth-wise count of primary capsules, number of filters in the convolution layer, and length of emotion capsules.

We employ a Bayesian Optimization methodology for finding the best set of hyperparameters, since each iteration for parameter optimization involves training on the entire set. Bayesian Optimization is best suited for problems where each iteration with a certain set of parameters is costly in terms of resources. We model the Bayesian Optimization problem using Gaussian Processes. We relate the hyperparameters of the model and the maximum cross-validation loss of the model with a function, (X) , in the hyperparameters plane, with $X = (x_1, x_2, \cdots .., x_N)$, $N = 4$. We assume that the known values of $f(X)$ at t previous observations, F_t for their corresponding values of X_t , $0 \leq t < T$, follows the distribution of a Gaussian Process with a prior distribution:

$$p(F_t \vee X_t) = \frac{e^{\left(\frac{-1}{2}F_t K_t^{-1} F_t^{-1}\right)}}{\sqrt{(2\pi)^t det(K_t)}} \quad (6)$$

Here, K_t is a $t \times t$ covariance matrix whose coefficients are derived from a kernel as $K_{mn} = K(X_m, X_n, \theta)$, θ being the kernel hyperparameters. This K_t represents prior assumption of the function in our Bayesian Optimization. In order to the determine the next set X_{t+1}, that would minimize the error, a posterior distribution $p_{t+1}\left(f_{t+1}|F_t, X_{t+1}\right.$ is estimated using the prior, and the surrogate model f_{t+1} is obtained using conditional probability operations on this posterior. Using Expected Improvement acquisition function on this surrogate model, a X_{t+1} is selected so that the difference between the current maximum $f_{max}(X)$ and $f_{t+1}(X_{t+1})$ is maximized [43,44].

$$EI(X) = E\Big(max\Big(\Big|f_{t+1}(X_{t+1}) - f_{max}(X)\Big|, 0\Big)\Big) \quad (7)$$

The prior is updated every time we evaluate the function and obtain the actual $f_{t+1}(X_{t+1})$ under the Bayesian paradigm. This step is repeated for a maximum of $T = 100$ steps. We start with an assumption of the hyperparameters $X_0 = (8, 32, 5, 16)$ and we define the search space for each hyperparameter as:

(i) length of primary capsules $(x|1)$, $x_1 \in [8, 32]$
(ii) depth-wise count of primary capsules $(x|2)$, $x_2 \in [30, 40]$
(iii) number of filters in the convolution layer $(x|3)$, $x_3 \in [3, 7]$
(iv) length of emotion capsules $(x_4), x_4 \in [16, 32]$

We repeat this process for each emotion class and find the most suitable hyperparameters for each binary classification task. After the best hyperparameter combination is obtained, we train a model with those hyperparameters and test it on the test set to report the accuracy.

We also test the performance of a CNN model and a Resnet model on the dataset, trained using 10-fold cross-validation; and hyperparameters tuned using Grid Search [45]. The hyperparameters chosen for the exhaustive search primarily related to the model architecture, and the range was limited. The comparison of performances of the deep learning models is included in Table 6.

4. Result and discussion

The testing has been performed in epochs. Each epoch considers a single trial duration for prediction. We report both the best-case and average case accuracies obtained during test epochs, since we are performing prediction on a cross-subject test case, and the models do not have any knowledge of the domain or distribution of the test signals. The accuracies reported are tabulated in Table 2. Since, for each test we have combined the EEG records from left-out-subject and left-out-records from the other subjects, each test-epochs can belong either to left-out-subject or the left-out-record. We report the best and average accuracies obtained for these different sets separately. Also, from among the test epochs, there would be one epoch that corresponds to the left-out-record of the left-out-subject. This is the epoch in the unseen subject, unseen record (USUR) domain, where the model has *zero familiarity*. We report the best accuracy and accuracies averaged over all the leave-one-out runs for the unseen records of the unseen subject in Table 3. The best performing model hyperparameters used to achieve the above scores are listed in Table 5.

Table 3
Unseen Subject, Unseen Record Accuracies for CapsNet.

Labels Accuracies	Arousal	Dominance	Valence	Liking
Best Case	100.000	100.000	100.000	100.000
Average Case	44.381	67.598	46.571	45.735

The best case accuracies listed belong to a run that performed best with regard to the specific criteria stated, e.g., best performance on the unseen records. In other words, the best case accuracy for subject and record originate from different test epochs, as can be seen from the plots of LOO subject and record accuracies (Figs. 4–7).

The mean of best-case performances for all leave one subject out across all prediction classes, thus comes out to be, 85.396%, and that of leave one record out comes out to be 84.518%. Similarly, the mean of averaged accuracies for leave one subject out across all prediction classes comes out to be 57.165%, and that of leave one record out comes out to be 53.309%. On the same line, the accuracies for unseen-subject-unseen record averaged over all the classes for the best predictions turn out to be 100.00% and for the averaged predictions, 51.071%. We argue that this accuracy of 51.071%, which is computed from the performance of the model on completely unseen data with no familiarity, is a proper gauge of what the model would perform in the real world.

There is a considerable degree of variation in the performance of the algorithm on different subjects. Table 4 shows a measure of variance in accuracies obtained by the model on dataset. Along with the standard deviations reported in Table 2 for the average cases, the quantities demonstrate the spread of the performance on the subjects.

It is however worthy to notice that the subjects on which the model does perform better have high precision and recall scores as well, which implies the model learns overall better feature representations on these subjects. We plot charts of accuracy, precision and recall (Fig. 4 – Fig. 7) for various subject + record combination we used during our training and validation. The two cases for LOO subject and LOO trials are shown in the graph for comparison of performance of proposed model when applied to LOO subject or LOO trials separately. The charts also provide an insight into how well the model would perform in a LOO subject or LOO record, when a particular subject/record is left out. We only consider precision and recall scores for predicted label of 1, which is the label predicted when the particular emotion class (arousal, dominance, valence, liking) is determined to be present in the sample processed.

Another interesting observation is that although for the labels where

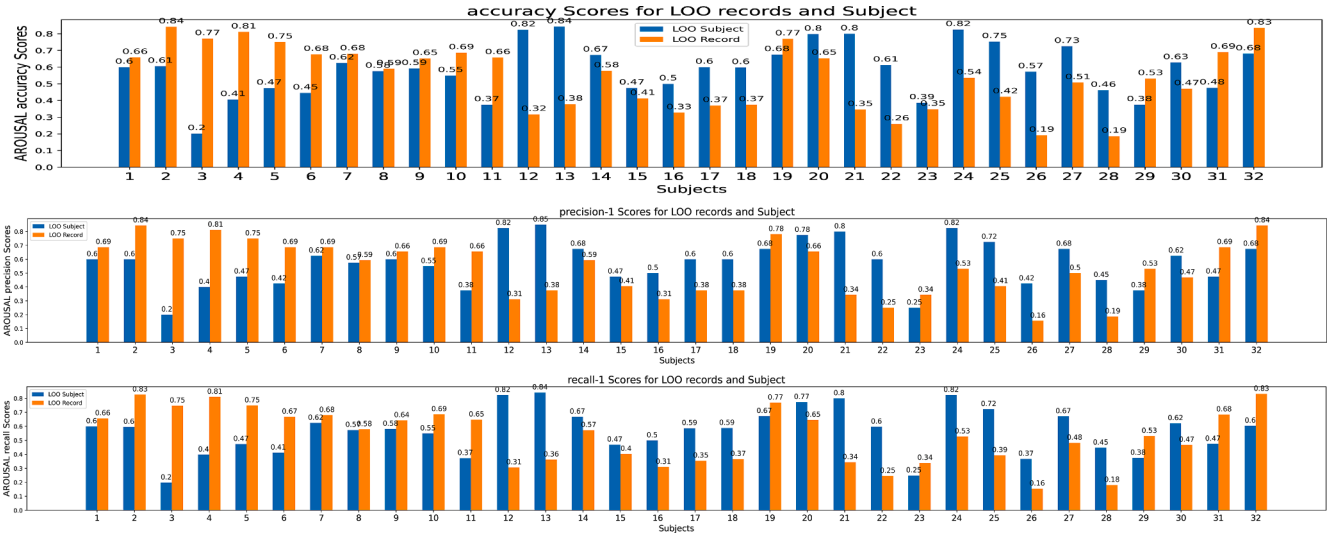

Fig. 4. Performance chart for arousal label.

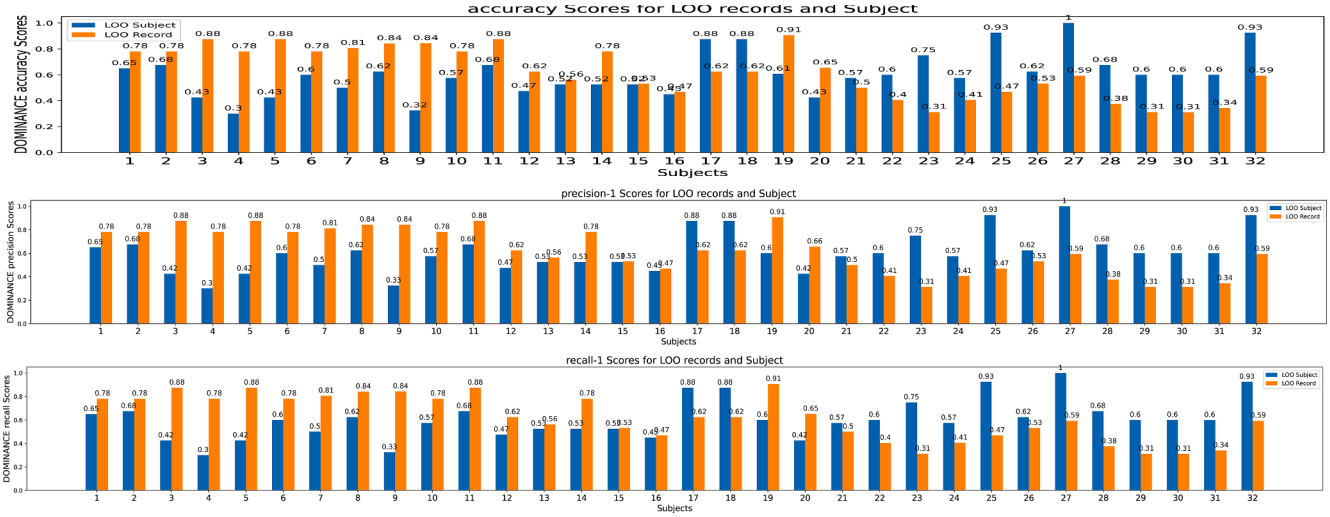

Fig. 5. Performance chart for dominance label.

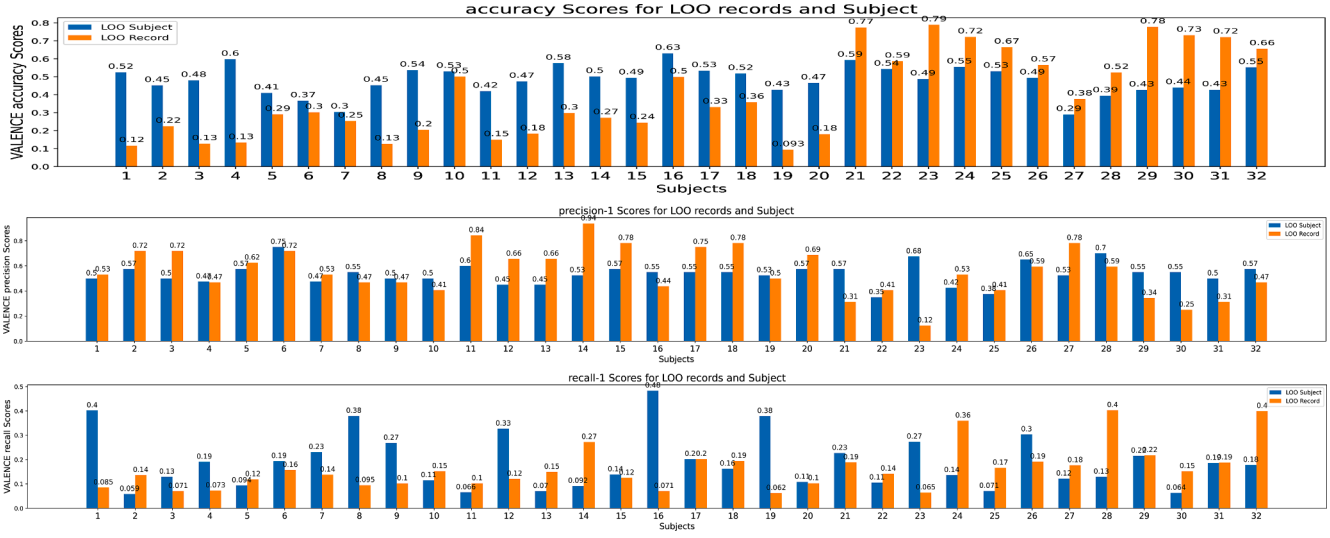

Fig. 6. Performance chart for valence label.

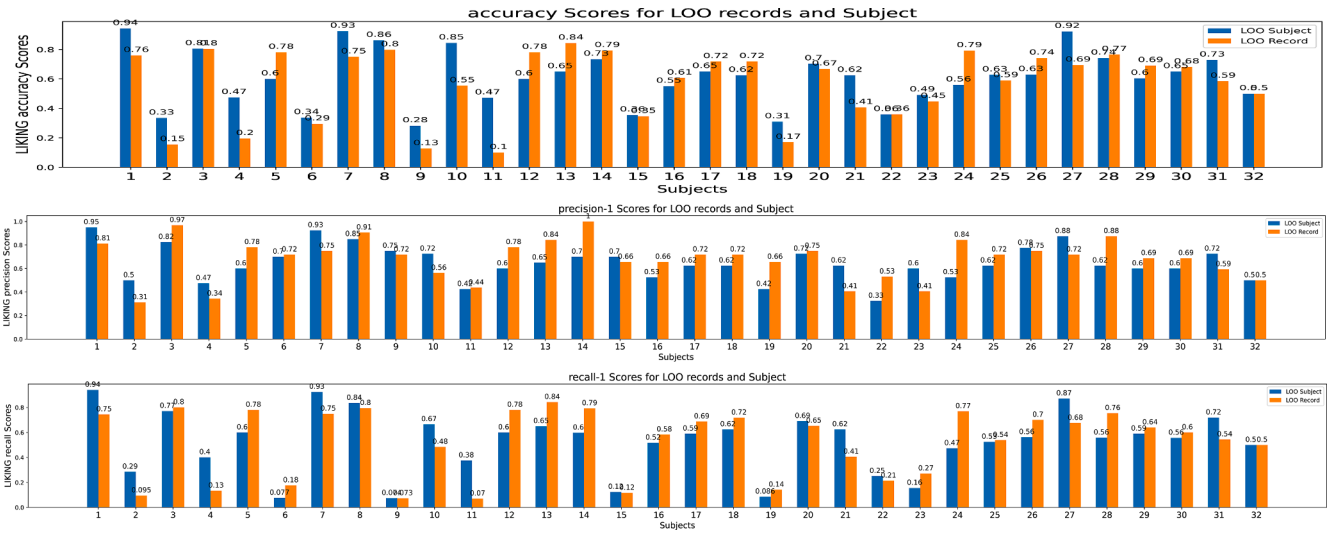

Fig.7. Performance chart for liking label.

Table 4
Variance in accuracies for CapsNet (when accuracies ϵ [0,1])

Labels	Arousal		Dominance		Valence		Liking	
Cases Accuracies	LOO Subject	LOO Record	LOO Subject	LOO Record	LOO Subject	LOO Record	LOO Subject	LOO Record
Average Case	0.02305	0.03610	0.02770	0.03633	0.00613	0.05335	0.03351	0.05348

Table 5
Best Fit Hyperparameters for CapsNet.

Hyperparameters	**Arousal**	**Dominance**	**Valence**	**Liking**
Primary Capsule Length	32	23	23	10
Count of primary capsules	31	38	40	31
No. of filters (Conv. Layer)	3	4	5	7
Emotion Capsules Length	22	16	31	32

accuracies are relatively higher, the precision and recall scores seem to closely mirror one another. However, as for the case of *valence* (Fig. 6), the recall scores are much lower. This could explain the poorer performance of the model on this particular emotional label. This is also evident in the performance chart for *liking* label (Fig. 7). Although the precision scores maybe decent for particular subjects, the poor recall of the model in these cases, renders a lower performance. Thus, this is a caveat where a scope for improvement has been identified.

We test our spatiotemporal frames on other popular deep learning based computer vision algorithms, viz., CNNs and ResNet, trained using 10-fold cross validation, and then tested on the hold-out data. The model hyperparameters were tuned using grid-search [45] cross validation, and the accuracy scores obtained for the best classifiers are comparable to that obtained in previous works with a similar setting. The comparison with the CapsNet model is done on the basis of CapsNet's performance on left-out subject. The accuracies obtained on the hold-out data was less than CapsNet models for arousal and liking labels, and a significant outperformance of CapsNet was observed in case of dominance label, indicating that the CapsNet model could learn better mapping from other deep learning based computer vision models used in literature. We have detailed the results of our experimentation in Table 6.

Hyperparameter tuning was performed on both CNN and ResNet models using grid-search [45]. The hyperparameters chosen to optimize in CNN was the nodes in the fully connected layer and the number of filters in convolution layers. The best hyperparameter set obtained for CNN was 128 nodes in the fully connected layer and 64 filters in the convolutional layer. For Resnet, the number of filters in 1st convolution layer,

Table 6
Comparison of performance achieved by proposed CapsNet approach with CNN and ResNet.

Model	**Arousal**	**Dominance**	**Valence**	**Liking**
CNN (128/64)	52.71	41.86	60.47	60.47
ResNet50	46.51	41.86	60.47	60.47
CapsNet (Best - Subject)	**84.249**	**100.00**	**63.042**	**94.292**
CapsNet (Average)	**58.525**	**60.966**	48.219	**60.951**

Table 7
Comparison of our results with other studies that used CapsNet on DEAP.

Study	Method Used	**Best Accuracy Reported**			
		Arousal	Dominance	Valence	Liking
J. Guo et al [5]	CapsNet with wavelet Transform	0.8737	–	0.8809	–
Yu Liu et al [6]	Multi-level Features	0.9831	0.9832	0.9797	–
Hao Chao. et al [8]	Multi-band Feature Matrix	0.6828	0.6725	0.6673	–
Proposed	Spatio-Temporal Frame Group	0.8417	0.9063	0.7891	0.8438

subsequent convolution blocks, and identical blocks were chosen. The best hyperparameters set obtained for ResNet was 64 filters for filters of 1st convolution layer, and 16 for each subsequent convolutional block; while maintaining the ideal number of filters, 16, in the identity blocks.

We compare our results to other prior works utilizing Capsule Networks for Emotion Recognition from EEG signals in Table 7. It could be observed in Table 6 that the proposed approach based on CapsNet algorithm gives a comparable performance to the best reported accuracy of other works, when considering the performance on left-out-records. Also, it must be noted that in this study classification on all four classes found in the dataset, i.e., arousal, dominance, valence and liking has been performed and reported.

Table 8
Performance Comparison with Different Studies that performed cross subject inference on DEAP dataset.

Study	Method	Accuracy Reported			
		Arousal	Dominance	Valence	Liking
Fu Yang et al [19]	ST-SBSSVM	–	–	72%	–
Pallavi Pandey et al [24]	VMD + DNN	61.25%	–	62.50 %	–
Li X et al [25]	SVM and specialized feature extraction	–	–	59.06%	–
V.Gupta et al [28]	Random Forests and SVMs	79.99%	–	79.95%	–
W. Jiang et al [30]	Decision Tree with SBS	–	–	65.2%	–
W. Zhang et al [31]	ANN + RFE	64.61%	–	65.29%	–
Pandey, P. et al [33]	CNN on CWT scalograms	58.5%	–	61.5%	–
Y. Cimtay. et al [46]	CNN with median filtering	–	–	72.81%	–
J. Liu et al [47]	Domain Adaptation with subject clustering	68.8%	–	73.9%	–
Zhen Liang et al [49]	CNN-RNN-GAN (EEGFuseNet)	58.78%	61.69%	56.27%	66.30%
Arjun et al [50]	Attention Driven Neural Networks	69.5%	–	65.9%	–
This Study (Average)	STFG + CapsNet	58.525%	60.966%	48.219%	60.951%

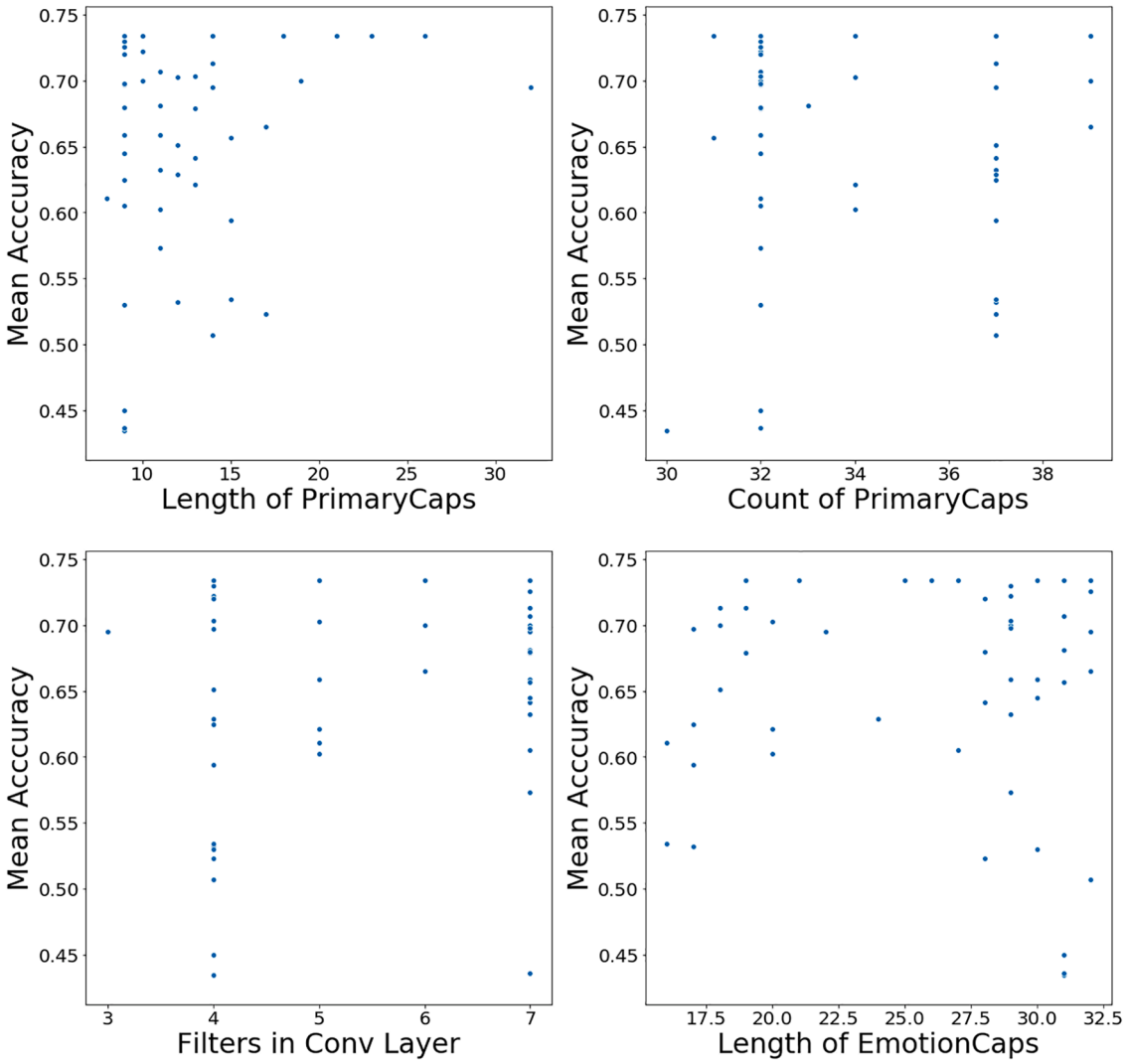

Fig. 8. Plots of Accuracy vs Hyperparameters for Liking Label; with the hyperparameters being, top-left: Length of Primary Capsules, top-right: Count of Primary Capsules, bottom-left: Number of Filters in Convolution Layer, bottom-right: Length of EmotionCaps.

We also make a comparison with prior works which adopted a cross-subject testing in Table 8. We consider multiple works that carried out cross-subject analysis on DEAP. It can be seen that most of the works employing leave-one-subject-out validation have considered only valence class to report the accuracy on DEAP dataset.

We have focused on a single dataset and the performance of the algorithm on the various classes of that dataset [20]. Comparing the performance on valence class of DEAP dataset, the proposed method has given a comparative performance to prior works on emotion recognition using EEG signals. Same could be said about the relative performance of the proposed method on arousal class; with the best case accuracy outperforming prior methods. This establishes that using spatiotemporal frames along with CapsNet, could achieve competent performance without any additional feature processing.

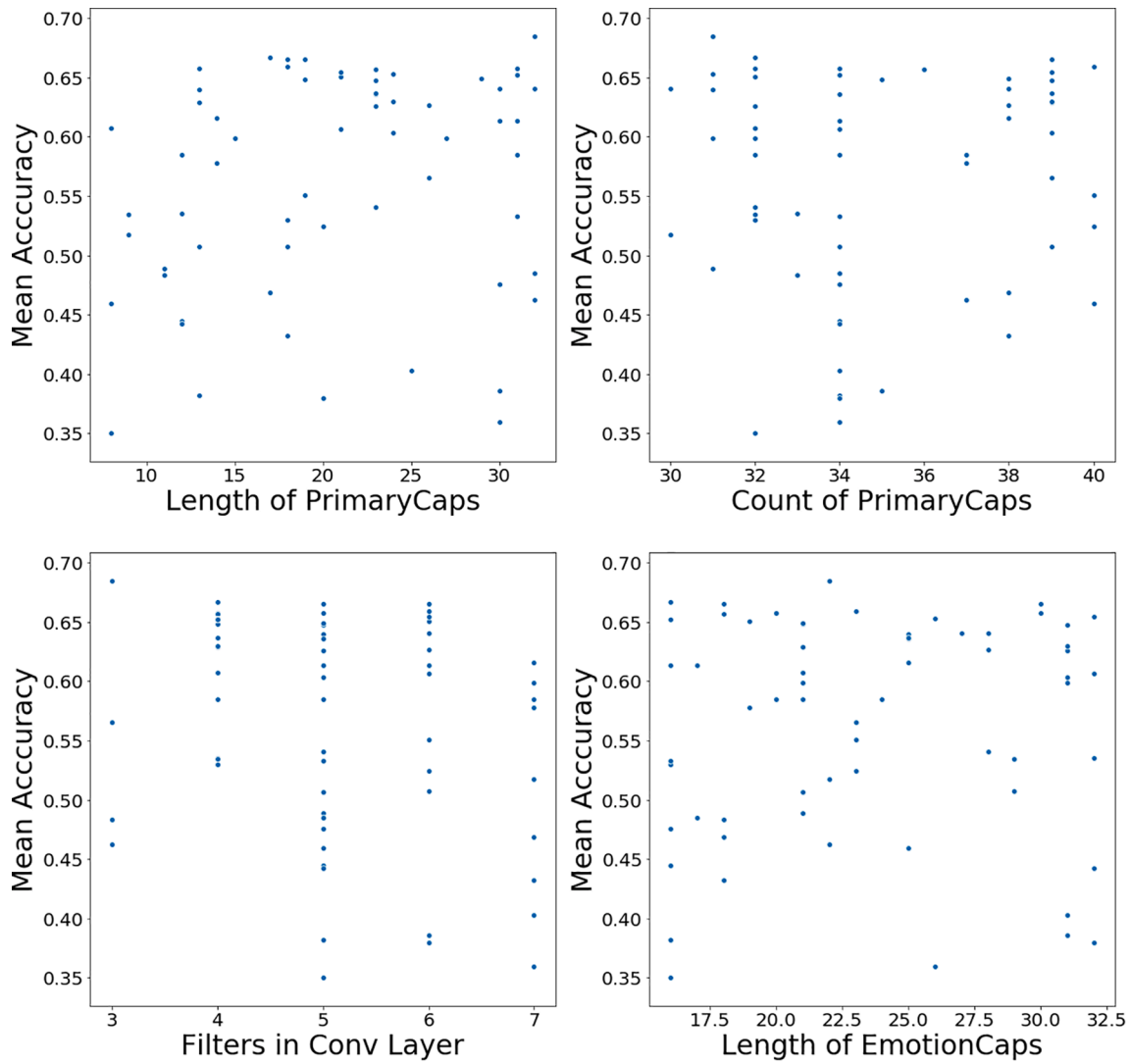

Fig. 9. Plots of Accuracy vs Hyperparameters for Arousal Label; with the hyperparameters being, top-left: Length of Primary Capsules, top-right: Count of Primary Capsules, bottom-left: Number of Filters in Convolution Layer, bottom-right: Length of EmotionCaps.

This could be attributed to the two distinct novelty elements proposed in this work. The first being the arrangement of signals in spatiotemporal frames which encode both spatial information of electrode position and the temporal recordings. The rearrangement is a trivial task as it does not involve any additional processing of signals pre or post rearrangement. This ensures minimal contribution of any other factors to the performance of the model once the data is provided in an appropriate format expected by the network.

Also, we rely on the power of convolutional layer to extract the spatiotemporal features like in [10], and expect the dynamic routing network of capsules to learn the mappings from these condensed rich features to the corresponding class labels. Since Capsule Networks have the innate ability to infer from complex non-linearities in the data using a rich vector representation of the different extracted features, and propagating the learned information without any loss to pooling.

4.1. Analysis of Adaptability of Capsnet Hyperparameters

We further analyze the relation between primary capsule length and emotion capsule length when applying dynamic routing to the problem of EEG emotion recognition. The trends are illustrated in Figs. 8–11. Considering the scatterplots for mean accuracy in Fig. 8, it could be observed that for arousal class, a lot of experiments were focused on a higher number for length of primary capsule, and on both extremes for the count of primary capsule, where the most suitable parameter set was obtained; however for filters in convolution layer the experiments focused on higher numbers, though it must be noted that best-fit set for filters came from the point with which least number of experiments were run. And, for Length of Emotion Caps, experiments were evenly distributed.

Interpreting Fig. 9, Fig. 10, and Fig. 11 along with Fig. 8, the trend for count of primary capsule is persistent through all the different class labels, with experiments being focused on either extreme where the best-fit lies. The experiments on filter numbers also have a skewed distribution throughout all the classes, with least number of experiments for 3 filters. However, for other classes the best fit value for filter number was found to be a different number than 3.

For classes other than arousal, the experiments for length of emotion caps are focused on the extremes of the experimental range, where the best-fit values are found. It must be noted that the classes with higher number of Emotion Capsules in their best-fit have achieved higher accuracy scores, which hints at the necessity of additional dimensionality in the EmotionCaps layer.

For the length of primary capsules, experiments were almost uniformly distributed throughout, especially when we consider experiments that yielded an accuracy score higher than 0.5. The only exception to this being the liking class, with more focus on the lower end of the domain,

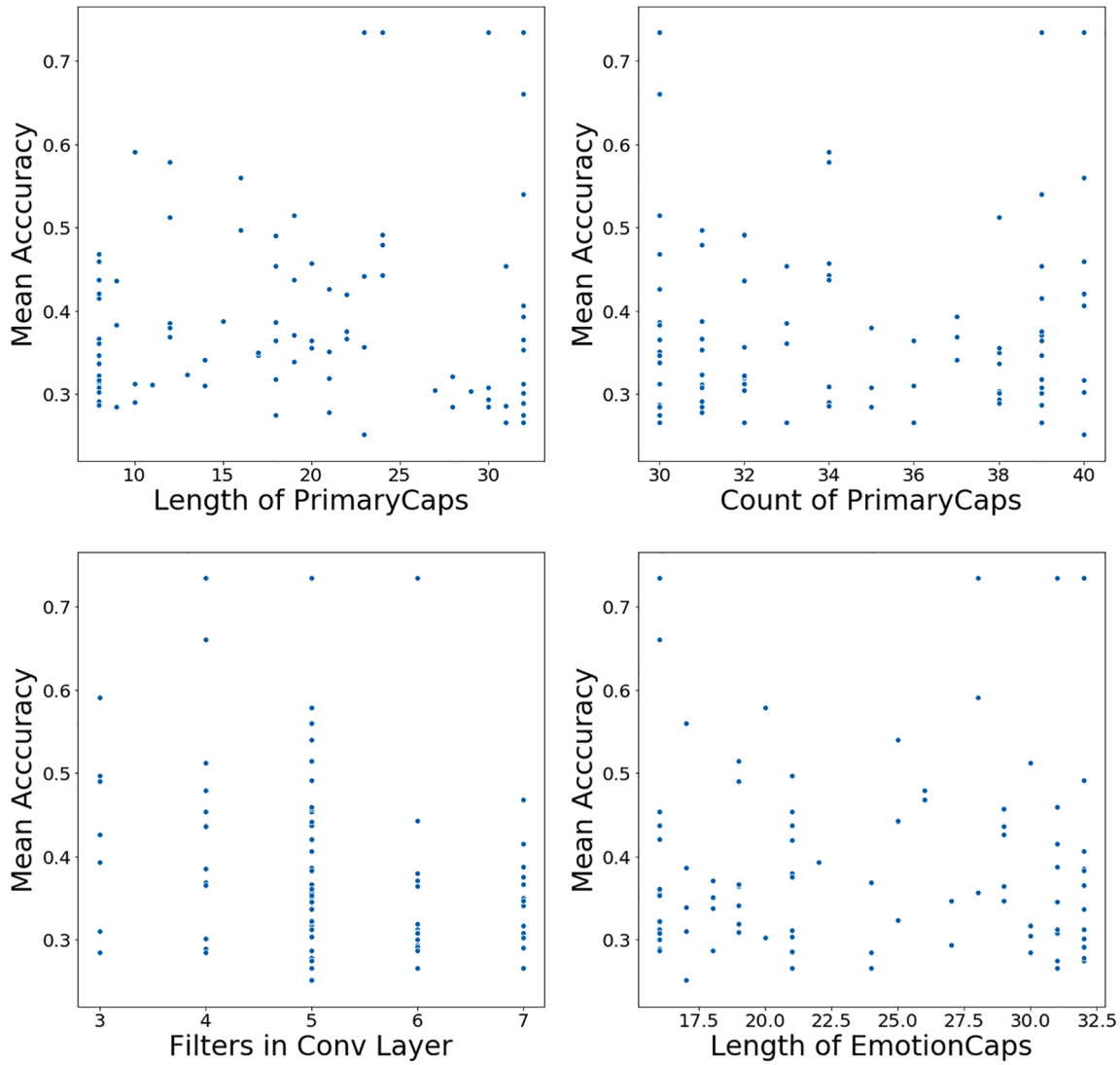

Fig. 10. Plots of Accuracy vs Hyperparameters for Valence Label; with the hyperparameters being, top-left: Length of Primary Capsules, top-right: Count of Primary Capsules, bottom-left: Number of Filters in Convolution Layer, bottom-right: Length of EmotionCaps.

where the best-fit parameter setting is found. On overall, however, the liking class has a clear-cut trend for distribution of experimental focus for each of the hyperparameters, and is the one best-classified.

In a previous work [7] on the same problem, authors utilized a different approach that had similarities to our method. They created spatiotemporal frames using interpolation to fill the sparse matrix, and then expanded the images so created to be 64×64, and employed 3D convolution network models for inference. This work however, does not shed any light on the justification for the method of choice used to interpolate, or how the artefacts introduced in temporal domain due to such interpolation are accounted for in the training. Hence, we chose to stick with simpler frame structures. The total model trained has a considerable complexity, with about 129.7 M trainable parameters when trained from scratch. This includes both the inference module and the reconstruction module, with the majority of parameters being in the reconstruction module, which is deeper than the inference module (6 FC Layers) in order to transform the $\sim$ 32dimensional vector of EmotionCaps to 128×81 dimensional vector of the corresponding input.

4.2. Limitations and Future Scope

In spite of the better performance of our approach, we identify a few limitations. Though the final result is representative of the splitting mechanism in order to minimize both degrees of familiarity, the representation however is not uniform, since the majority of samples are from the unknown subject - known video rather than unknown video - known subject. Also, the unknown subject-unknown video set was not separated at the time of experimentation. The splits, although made by selecting random subjects and videos, are static sets. Producing these sets in a K-fold manner is left for future work.

It is assumed that the uniform thresholding of labels is an accurate enough thresholding for the more granular subject ratings, with any subject ratings greater than this threshold being considered as indicative of the considered emotion. The actual threshold might change depending on the persona, biases and other unknown psychological factors of the subjects, for example, the ratings from 27th subject are all greater than 5. A variable threshold methodology for labelling is left for a future work.

The problem addressed in this study was of binary classification for each class. However, it is desirable to have a model trained to perform four-class classification on the dataset. Similarly, study of the use of 3D convolutions based CapsNet on the specific problem, and comparison of its performance to the 2D convolution based CapsNet is also desirable. These are saved for a future study.

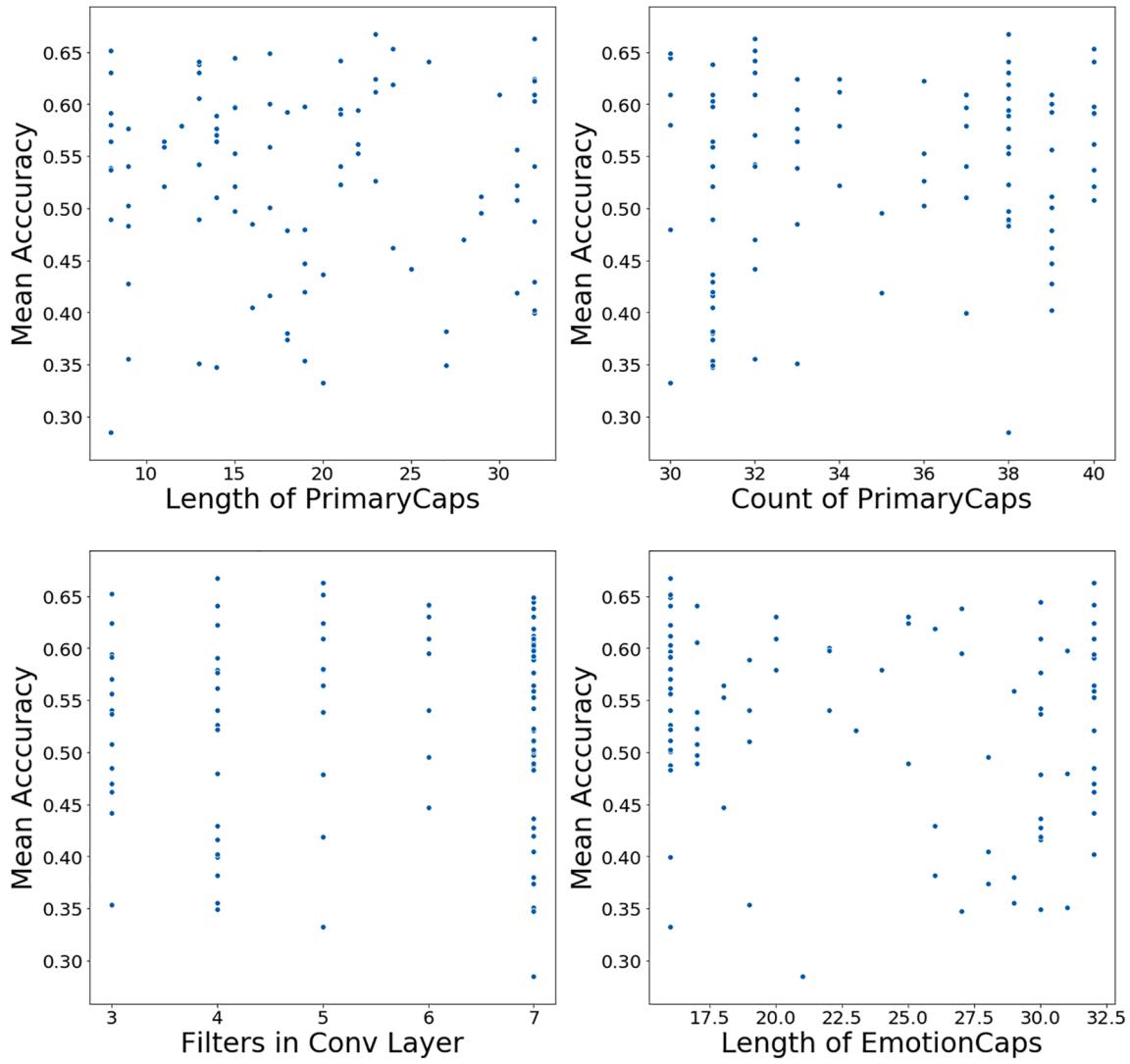

Fig. 11. Plots of Accuracy vs Hyperparameters for Dominance Label; with the hyperparameters being, top-left: Length of Primary Capsules, top-right: Count of Primary Capsules, bottom-left: Number of Filters in Convolution Layer, bottom-right: Length of EmotionCaps.

5. Conclusion

In this study we present an application of capsule networks to the cross subject inference of emotions from EEG signals, using a special rearrangement of EEG signals to incorporate both spatial and temporal information. We also demonstrate a form of data splitting such that the model has to perform on a set with which it has the minimum degree of familiarity, forcing it to generalize better. Hence, the result obtained is representative of the inference of the model on completely unseen data. The model reports a best case accuracy of 85.396% and average case accuracy of 57.165%, when averaged across all the classes. Also, the proposed methodology beat the best accuracy reported in other works employing capsule networks for intra-subject inference. We also study the application of Bayesian Optimization for the specific problem and analyze the relation between the different hyperparameters and the accuracy score, for all the classes in the dataset. It could be concluded from the study that using a spatio-temporal representation for EEG signals proposed in this work, satisfactory classification scores could be obtained on the task of cross-subject emotion recognition, when such signal representation is used in training of deep learning algorithms, like, CapsNet, ResNet and CNN. The key highlights of the work are:

- A method of incorporating the topology of electrode arrangement along with the temporal information by forming sparse spatiotemporal frame based features is proposed and implemented
- We use a data-split methodology that would help us test the model with multiple degrees of familiarity
- We propose using CapsNet to process and infer from the sparse spatiotemporal frame based features and obtain a best case accuracy of 0.85396 when averaged across all the classes.
- We also analyze the variation of accuracy for different settings of Capsule Network parameters.
- We provide a baseline for emotion recognition for unseen subject unseen record classification (Table 3)

Experimental Data and Source Code Availability

Experimental dataset is publicly available is at http://www.eecs.qmul.ac.uk/mmv/datasets/deap/ and source code of this experiment will be available at https://sites.google.com/site/gcjanahomepage/publications/Publications-Source-Codes (https://doi.org/10.5281/zenodo.5688674).

Ethical Approval

This study is performed on a publicly available dataset. None of the authors have directly involved in way of experimentation, survey, data acquisition, or otherwise, with any human or animal participants for the purpose of this study. Since the data used in this study has been acquired

from data available publicly on request, the authors cannot be held accountable for any concerns arising about the nature of acquisition of data.

CRediT authorship contribution statement

Gopal Chandra Jana: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Anshuman Sabath:** Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Anupam Agrawal:** Supervision, Resources, Validation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was carried out at Interactive Technologies & Multimedia Research Lab (ITMR Lab) supported by the Department of Information Technology, Indian Institute of Information Technology Allahabad (https://www.iiita.ac.in/), UP, India. The computational results reported in this work were performed on the Central Computing Facility of IIIT-Allahabad. The authors are grateful for this support.

Appendix

Table 9
EEG channel names (according to the 10/20 system) as per DEAP Dataset [20].

Channel no.	Channel name
1	Fp_1
2	AF_3
3	F_3
4	F_7
5	FC5
6	FC1
7	C3
8	T7
9	CP5
10	CP1
11	P3
12	P7
13	PO3
14	O1
15	Oz
16	Pz
17	Fp2
18	AF4
19	Fz
20	F4
21	F8
22	FC6
23	FC2
24	Cz
25	C4
26	T8
27	CP6
28	CP2
29	P4
30	P8
31	PO4
32	O2

References:

[1] Dzedzickis, Andrius, et al. "Human Emotion Recognition: Review of Sensors and Methods" Sensors, vol. 20, no. 3, Jan. 2020, pp. 592, https://doi.org/10.3390/s20030592.
[2] Klonowski, Wlodzimierz. "Everything You Wanted to Ask about EEG but Were Afraid to Get the Right Answer" Nonlinear Biomedical Physics, vol. 3, no. 1, Dec. 2009, pp. 2, https://dx.doi.org/10.1186%2F1753-4631-3-2.
[3] M. Feidakis, T. Daradoumis, S. Caballe, Endowing e-Learning Systems with Emotion Awareness, Third International Conference on Intelligent Networking and Collaborative Systems (2011) 68–75, https://doi.org/10.1109/INCoS.2011.83.
[4] Verma, Gyanendra K., and Uma Shanker Tiwary. "Affect Representation and Recognition in 3D Continuous Valence–Arousal–Dominance Space" Multimedia Tools and Applications, vol. 76, no. 2, Jan. 2017, pp. 2159–83, https://doi.org/10.1007/s11042-015-3119-y.
[5] J. Guo, et al., EEG Emotion Recognition Based on Granger Causality and CapsNet Neural Network, in: 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), 2018, pp. 47–52, https://doi.org/10.1109/CCIS.2018.8691230.
[6] Liu, Yu, et al. "Multi-Channel EEG-Based Emotion Recognition via a Multi-Level Features Guided Capsule Network" Computers in Biology and Medicine, vol. 123, Aug. 2020, pp. 103927, https://doi.org/10.1016/j.compbiomed.2020.103927.
[7] Cho, Jungchan, and Hyoseok Hwang. "Spatio-Temporal Representation of an Electoencephalogram for Emotion Recognition Using a Three-Dimensional Convolutional Neural Network" Sensors, vol. 20, no. 12, June 2020, pp. 3491, https://doi.org/10.3390/s20123491.
[8] Chao, Hao, et al. "Emotion Recognition from Multiband EEG Signals Using CapsNet", Sensors, vol. 19, no. 9, May 2019, pp. 2212, https://doi.org/10.3390/s19092212.
[9] Li, He, et al. "Cross-Subject Emotion Recognition Using Deep Adaptation Networks" Neural Information Processing, edited by Long Cheng et al., vol. 11305, Springer International Publishing, 2018, pp. 403–413, https://doi.org/10.1007/978-3-030-04221-9_36.
[10] Li, Jinpeng, et al. "Multisource Transfer Learning for Cross-Subject EEG Emotion Recognition" IEEE Transactions on Cybernetics, 2019, pp. 1–13, https://doi.org/10.1109/TCYB.2019.2904052.
[11] Sabour, Sara, et al. "Dynamic Routing Between Capsules" ArXiv:1710.09829 [Cs], Nov. 2017, http://arxiv.org/abs/1710.09829.
[12] R. Davidson, N. Fox, Asymmetrical brain activity discriminates between positive and negative affective stimuli in human infants, Science 218 (4578) (1982) 1235–1237, https://doi.org/10.1126/science.7146906.
[13] P. Ekman, R. Davidson, The Nature of Emotion: Fundamental Questions, Oxford University Press, 1994.
[14] C. Cheng, et al., Emotion Recognition Algorithm Based on Convolution Neural Network, in: 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), 2017, pp. 1–5, https://doi.org/10.1109/ISKE.2017.8258786.
[15] Mei, Han, and Xiangmin Xu. "EEG-Based Emotion Classification Using Convolutional Neural Network" International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), 2017, pp. 130–35, https://doi.org/10.1109/SPAC.2017.8304263.
[16] Tripathi, Samarth, et al. "Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset", Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI Press, 2017, pp. 4746–52.
[17] Dabas, Harsh, et al. "Emotion Classification Using EEG Signals" Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence, Association for Computing Machinery, 2018, pp. 380–84, https://doi.org/10.1145/3297156.3297177.
[18] Bao, Guangcheng, et al. "Two-Level Domain Adaptation Neural Network for EEG-Based Emotion Recognition" Frontiers in Human Neuroscience, vol. 14, Jan. 2021, pp. 605246, https://doi.org/10.3389/fnhum.2020.605246.
[19] F.u. Yang X. Zhao W. Jiang P. Gao G. Liu 13 10.3389/fncom.2019.00053.
[20] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, DEAP: A Database for Emotion Analysis Using Physiological Signals, IEEE Transactions on Affective Computing 3 (1) (2012) 18–31, https://doi.org/10.1109/T-AFFC.2011.15.
[21] SEED Dataset. https://bcmi.sjtu.edu.cn/home/seed/. Accessed: 17th August 2021.
[22] Loughborough University EEG based Emotion Recognition Dataset. https://www.dropbox.com/s/xlh2orv6mgweehq/LUMED_EEG.zip?dl=0. Accessed: 17th August 2021.
[23] H. Chao, L. Dong, Emotion Recognition Using Three-Dimensional Feature and Convolutional Neural Network from Multichannel EEG Signals, IEEE Sensors Journal 21 (2) (2021) 2024–2034, https://doi.org/10.1109/JSEN.736110.1109/JSEN.2020.3020828.
[24] Pandey, Pallavi, and K. R. Seeja. "Subject Independent Emotion Recognition from EEG Using VMD and Deep Learning" Journal of King Saud University - Computer and Information Sciences, Nov. 2019, pp. S1319157819309991, https://doi.org/10.1016/j.jksuci.2019.11.003.
[25] Li, Xiang, et al. "Exploring EEG Features in Cross-Subject Emotion Recognition" Frontiers in Neuroscience, vol. 12, Mar. 2018, pp. 162, https://doi.org/10.3389/fnins.2018.00162.
[26] Zhang, Weiwei, et al. "Cross-Subject EEG-Based Emotion Recognition with Deep Domain Confusion" Intelligent Robotics and Applications, edited by Haibin Yu et al., vol. 11740, 2019, pp. 558–70, https://doi.org/10.1007/978-3-030-27526-6_49.

[27] Cimtay, Yucel, et al. "Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion" IEEE Access, vol. 8, 2020, pp. 168865– 168878, https://doi.org/10.1109/ACCESS.2020.3023871.
[28] Gupta, Vipin, et al. "Cross-Subject Emotion Recognition Using Flexible Analytic Wavelet Transform From EEG Signals" IEEE Sensors Journal, vol. 19, no. 6, Mar. 2019, pp. 2266–2274, https://doi.org/10.1109/JSEN.2018.2883497.
[29] Fdez, Javier, et al. "Cross-Subject EEG-Based Emotion Recognition Through Neural Networks With Stratified Normalization" Frontiers in Neuroscience, vol. 15, Feb. 2021, pp. 626277, https://doi.org/10.3389/fnins.2021.626277.
[30] W. Jiang, et al., Cross-Subject Emotion Recognition with a Decision Tree Classifier Based on Sequential Backward Selection, in: 11th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2019, pp. 309–313, https://doi.org/10.1109/IHMSC.2019.00078.
[31] W. Zhang, Z. Yin, EEG Feature Selection for Emotion Recognition Based on Cross-Subject Recursive Feature Elimination, 39th Chinese Control Conference (CCC) (2020) 6256–6261. https://doi.org/10.23919/CCC50068.2020.9188573.
[32] S. Hwang, et al., Subject-Independent EEG-Based Emotion Recognition Using Adversarial Learning, in: 8th International Winter Conference on Brain-Computer Interface (BCI), 2020, pp. 1–4, https://doi.org/10.1109/BCI48061.2020.9061624.
[33] Pandey, Pallavi, and K. R. Seeja. "Subject Independent Emotion Recognition System for People with Facial Deformity: An EEG Based Approach" Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 2, Feb. 2021, pp. 2311–2320, https://doi.org/10.1007/s12652-020-02338-8.
[34] Wen, Zhiyuan, et al. "A Novel Convolutional Neural Networks for Emotion Recognition Based on EEG Signal" International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), 2017, pp. 672–677, https://doi.org/10.1109/SPAC.2017.8304360.
[35] J.X. Chen, P.W. Zhang, Z.J. Mao, Y.F. Huang, D.M. Jiang, Y.N. Zhang, Accurate EEG-Based Emotion Recognition on Combined Features Using Deep Convolutional Neural Networks, IEEE Access 7 (2019) 44317–44328, https://doi.org/10.1109/Access.628763910.1109/ACCESS.2019.2908285.
[36] Du, Xiaobing, et al. "An Efficient LSTM Network for Emotion Recognition from Multichannel EEG Signals" IEEE Transactions on Affective Computing, 2020, pp. 1–1, https://doi.org/10.1109/TAFFC.2020.3013711.
[37] Zhang, Tong, et al. "Spatial–Temporal Recurrent Neural Network for Emotion Recognition" IEEE Transactions on Cybernetics, vol. 49, no. 3, Mar. 2019, pp. 839–847, https://doi.org/10.1109/TCYB.2017.2788081.
[38] M. Yanagimoto, C. Sugimoto, Recognition of Persisting Emotional Valence from EEG Using Convolutional Neural Networks, in: 9th International Workshop on Computational Intelligence and Applications (IWCIA), 2016, pp. 27–32, https://doi.org/10.1109/IWCIA.2016.7805744.
[39] Y. Yang, et al., Emotion Recognition from Multi-Channel EEG through Parallel Convolutional Recurrent Neural Network, International Joint Conference on Neural Networks (IJCNN) (2018) 1–7, https://doi.org/10.1109/IJCNN.2018.8489331.
[40] Wang, Xiao-Wei, et al. "Emotional State Classification from EEG Data Using Machine Learning Approach" Neurocomputing, vol. 129, Apr. 2014, pp. 94–106, https://doi.org/10.1016/j.neucom.2013.06.046.
[41] Hinton, Geoffrey E., et al. Matrix Capsules with EM Routing. 2018. openreview.net, https://openreview.net/forum?id=HJWLfGWRb.
[42] A. Karpathy, et al., Large-Scale Video Classification with Convolutional Neural Networks, Conference on Computer Vision and Pattern Recognition (2014) 1725–1732, https://doi.org/10.1109/CVPR.2014.223.
[43] Bayesian optimization. http://krasserm.github.io/2018/03/21/bayesian-optimization/. Accessed 17th August 2021.
[44] C. Brecque, The Intuitions behind Bayesian Optimization with Gaussian Processes, Medium, 2 Apr. (2021). Accessed 17th August 2021, https://towardsdatascience.com/the-intuitions-behind-bayesian-optimization-with-gaussian-processes-7e00fcc898a0.
[45] Simon Chan, Philip Treleaven, "Chapter 5 - Continuous Model Selection for Large-Scale Recommender Systems", Editor(s): Venu Govindaraju, Vijay V. Raghavan, C. R. Rao, Handbook of Statistics, Elsevier, Volume 33, 2015, pp. 107-124, ISSN 0169-7161, ISBN 9780444634924.
[46] Cimtay, Yucel, and Erhan Ekmekcioglu. "Investigating the Use of Pretrained Convolutional Neural Network on Cross-Subject and Cross-Dataset EEG Emotion Recognition" Sensors, vol. 20, no. 7, Apr. 2020, pp. 2034, https://doi.org/10.3390/s20072034.
[47] J. Liu, X. Shen, S. Song and D. Zhang, "Domain Adaptation for Cross-Subject Emotion Recognition by Subject Clustering" 10th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 904-908, https://doi.org/10.1109/NER49283.2021.9441368.
[48] Yingdong Wang, Jiatong Liu, Qunsheng Ruan, Shuocheng Wang, Chen Wang, "Cross-subject EEG emotion classification based on few-label adversarial domain adaption", Expert Systems with Applications, Volume 185, 2021, 115581, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2021.115581.
[49] Zhen Liang, Rushuang Zhou, Li Zhang, Linling Li, Gan Huang, Zhiguo Zhang, and Shin Ishii, "EEGFuseNet: Hybrid Unsupervised Deep Feature Characterization and Fusion for High-Dimensional EEG with An Application to Emotion Recognition", arXiv:2102.03777 [cs.HC], https://arxiv.org/abs/2102.03777v.
[50] Arjun, Aniket Singh Rajpoot, Mahesh Raveendranatha Panicker, "Subject Independent Emotion Recognition using EEG Signals Employing Attention Driven Neural Networks", arXiv:2106.03461 [cs.NE], https://arxiv.org/abs/2106.03461v1.