Received 20 November 2023, accepted 12 December 2023, date of publication 18 December 2023, date of current version 3 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3344476

RESEARCH ARTICLE

A Graph Neural Network for EEG-Based Emotion Recognition With Contrastive Learning and Generative Adversarial Neural Network Data Augmentation

SAREH SOLEIMANI GILAKJANI, (Student Member, IEEE), AND HUSSEIN AL OSMAN, (Member, IEEE)
Department of Electrical and Computer Engineering, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Corresponding author: Sareh Soleimani Gilakjani (ssole037@uottawa.ca)

ABSTRACT The limited size of existing datasets and signal variability have hindered EEG-based emotion recognition. In this paper, we present a solution that simultaneously addresses both problems. Generative Adversarial Networks (GANs) have recently shown notable data augmentation (DA) success. Therefore, we leverage a GAN-based DA technique to enhance the robustness of our proposed emotion recognition model by synthetically increasing the size of our datasets. Moreover, we employ contrastive learning to improve the quality of the learned representations from EEG signals and mitigate the adverse impact of inter-subject and intra-subject variability in signals corresponding to the same stimuli or emotions. We do so by maximizing the similarity in the representation of such EEG signals. We perform EEG-based emotion classification using a Graph Neural Network (GNN), which learns the relationship between the extracted EEG features. We compare the proposed model with several recent state-of-the-art emotion recognition models on the DEAP and MAHNOB datasets. The experimental results demonstrate that the proposed model outperforms previous models with a 64.84% and 66.40% emotion classification accuracy on the test set of the DEAP dataset and a 66.98% and 71.69% emotion classification accuracy on the test set of the MAHNOB-HCI dataset for the valence and arousal emotional dimensions, respectively. We perform an ablation study to demonstrate how contrastive learning, GAN, and GNN contribute to improving the proposed solution's performance.

INDEX TERMS Contrastive learning, data augmentation, emotion in human-computer interaction, graph neural network, machine learning.

I. INTRODUCTION

Automated emotion recognition technologies can be incorporated into diverse medical, educational, and entertainment applications [1]. Humans express emotions through different physical and physiological modalities. Compared with conventional modalities used for emotion recognition, such as facial expression [2] and voice signal [3], internal physiological signals can provide a more effective human emotional state recognition [4]. Electroencephalography (EEG) is the most used Brain-Computer Interface (BCI) [5] that provides a direct measurement on the cerebral cortex of the brain. EEG is advantageous for its relatively low-cost and high temporal resolution compared to other neuroimaging technologies, such as functional Magnetic Resonance Imaging (fMRI) and Magnetoencephalography (MEG) [6]. In recent years, due to dry electrode technology's rapid development, which decreases this BCI's invasiveness, EEG has become even more suitable for affective computing applications [7]. Consequently, EEG-based emotion recognition

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

models have lately received substantial attention [8], [9], [10], [11].

A. EEG FEATURE EXTRACTION

A typical EEG emotion recognition network comprises two major parts, i.e., discriminative EEG feature extraction and emotion classification. EEG signals are stored as time-domain series, but EEG features used for emotion recognition can be captured through their time-dependent or frequency-dependent components or a mix of both. One of the most used methods for frequency-dependent feature extraction in EEG is to divide the signal frequency range into several bands, namely δ (0 to 4Hz), θ (4 to 7Hz), α (8 to 12Hz), β (13 to 30Hz) and γ (31 to 50Hz) [12], and extract features from each band. There are extensive studies that investigate EEG representations for emotion recognition [13], [14]. Differential Entropy (DE) [15] features have been widely used for state-of-the-art emotion recognition models [16], [17]. These features have outperformed other EEG feature sets such as Differential Asymmetry (DASM), Rational Asymmetry (RASM), Differential Caudality (DCAU), and Power Spectral Density (PSD) [18], [19], [20]. DE is generally equivalent to the logarithmic spectral energy for a fixed-length EEG signal in a specific frequency band [15]. Cheng et al. constructed a manual 2D frame for each sample according to the EEG channels' distribution and used these frames as input features to their proposed classification model [21]. They used a Cascade Forest, using Deep Forest (DF) algorithms, to classify emotions. They benchmarked the performance of their model against various existing emotion recognition models on two standard datasets and reported higher accuracy rates for their model. Li et al. processed raw EEG signals for their proposed classification model [22]. They implemented a dense layer paired with a softmax activation function for the purpose of emotion classification. They refined their model using Neural Architecture Search (NAS), which was guided by reinforcement learning strategies. They compared their model with several existing emotion recognition models such as SVM, Decision Tree (DT), Multilayer Perceptron (MLP) Convolution Recurrent Attention, Dynamic Graph Convolutional Network (DGCNN), and Continuous CNN. The outcomes demonstrated that the performance of their model surpassed that of the others in performance.

B. REPRESENTATION LEARNING USING DEEP NETWORKS

In addition to manual features, researchers have also leveraged deep networks to learn EEG representations and model the relationship between different EEG channels. For instance, since researchers observed that EEG data for emotion recognition exhibits long-term dependencies [23], a Recurrent Neural Network (RNN) [24] was used to capture temporal dependencies in sequential data. There are two commonly used types of RNNs: Long-Term Short Memory (LSTM) and Gated Recurrent Unit (GRU). LSTM was first introduced by Hochreiter and Schmidhuber in 1997 [25] to address the issue of long-term dependencies in data, as standard RNNs are often limited by the gradient vanishing/exploding problem, which hampers learning for long data sequences [26]. Yang et al. deployed a Bidirectional LSTM (BiLSTM) network [27] using DE features where the network models sequential information in the backward and forward directions.

Convolutional Neural Networks (CNNs) and attention mechanisms have been deployed to extract emotion-related EEG representations [28], [29], [30], [31], [32]. Li et al. deployed a model consisting of a deep CNN called CapsNet and an attention mechanism to learn shared representations from multi-tasks [30]. They compared their model with several benchmark emotion recognition models and achieved a better performance with their proposed model [32]. Graph Neural Networks (GNNs) have also been employed to capture the spatial relation between EEG features across different EEG electrodes [33], [34], [35].

Some researchers combined various deep networks to extract deep EEG representations [36], [37], [38]. Yin et al. combined Graph Convolutional Neural Network (GCNN) with LSTM to capture both spatial and temporal relationships among EEG channels [36]. Li et al. combined an attention mechanism with a bidirectional LSTM [37]. Du et al. proposed a model constructed using an attention-based auto-encoder, an LSTM-based feature extractor, and a domain discriminator [38]. They demonstrated that their proposed model outperforms Support Vector Machine (SVM), Deep Belief Network (DBN), Graph Convolutional Neural Network (GCNN), and DGCNN models. Yang et al. utilized the combination of CNN and LSTM to learn deep representations of EEG signals [39].

Researchers have also deployed deep auto-encoders and attention-based auto-encoders to extract high-quality EEG features by encoding and decoding input data [40], [41], [42]. Zhang et al. deployed a deep recurrent autoencoder (AE) to recognize emotions from EEG signals. The AE is trained separately to extract EEG representations, which are then passed to two fully connected dense layers to perform the emotion classification task [40]. In [41], the authors proposed a combination of a CNN and a deep sparse autoencoder to extract EEG latent representations. These representations were then fed into a deep neural network (DNN) consisting of three fully connected dense layers to perform emotion classification. Rajpoot et al. proposed a model consisting of an LSTM with a channel attention autoencoder to extract high-level EEG representations [42]. They also deployed CNN with an attention mechanism to perform emotion classification.

Transformer-based emotion recognition models using EEG signals have also been deployed lately in the context of automated emotion recognition [43], [44]. Wang et al. proposed a transformer-based model to perform emotion recognition using EEG signals [43]. Their approach involved categorizing EEG signals based on different brain regions and then utilizing transformers to synthesize the information from these regions. They tested their proposed model on two benchmark datasets and achieved better performance compared to the state-of-the-art emotion recognition models. Liu et al. proposed an emotion transformer model consisting of variants of self-attention blocks exclusively [44]. They compared their model with several benchmark emotion recognition models, and their results demonstrated that their model achieves a superior performance.

C. DATA AUGMENTATION

EEG datasets for emotion recognition are often limited in size, making it challenging to achieve satisfactory accuracy using machine learning techniques, especially deep learning approaches. This is because deep learning models have a large number of model parameters, requiring a significant amount of data for effective training [45]. To overcome this problem, Data Augmentation (DA) techniques were introduced to enlarge the size of datasets synthetically [46]. DA refers to the process of generating new data samples by transforming existing samples in a dataset. This technique can increase the classification's accuracy and stability, including EEG-based emotion classification [47]. Generative Adversarial Networks (GANs), which were first introduced by Goodfellow et al. [48], are widely used for the augmentation of EEG data [49], [50], [51]. For instance, Lue et al. proposed a Conditional Wasserstein GAN (CWGAN) to augment EEG data for emotion classification [52]. They adopted a set of indicators to judge the quality of the generated data. Their results indicated that their proposed CWGAN significantly improved the emotion classification accuracy when appending generated EEG data to the actual dataset. Luo et al. adopted a Conditional Boundary Equilibrium GAN (CBEGAN) [53] to generate synthetic training data for multiple modalities [51]. They tested their model on two different datasets, and their results indicated that using CBEGAN improved emotion recognition accuracy.

D. CONTRASTIVE LEARNING

A novel recent framework named SimCLR was proposed in [54] for Contrastive Learning (CL) of visual representations. Their proposed CL approach learns data representations by maximizing the agreement between various augmented transforms of the same data example via a contrastive loss function. The idea of maximizing the agreement between representations of a sample was first introduced by Becker and Hinton [55].

CL has been widely used for general data-representation learning [56], [57]. CL is a self-supervised learning method to project the data into a space where different views of the same input sample have highly similar representations. While initially developed for image classification [54], [58] and computer vision applications [59], CL has also been applied to emotion recognition based on physiological signals [60], [61], [62]. Mohsenvand et al. [60] used a similar method to SimCLR, which they called SeqCLR, to learn similarities between differently augmented transforms of the same EEG data sample, disregarding the emotional state of the data sample. Augmented transforms were generated using temporal masking, linear scaling, time shifting, DC shifting, band-stop filtering, and Gaussian noise adding. They also compared their model with several state-of-the-art models on EEG-based emotion recognition and concluded that it achieves a higher classification accuracy.

Pinitas et al. [61] proposed a model to learn general affect-infused multi-modal representations from audio, video, Electrocardiography (ECG), and Electrodermal Activity (EDA) modalities. The model was built upon the contrastive learning framework introduced in [58]. Their results show that using CL improves multi-modal affect modeling tasks. However, their proposed CL-based solution only considered a single emotional dimension for their pairing mechanism.

Shen et al. employed CL to address the problem of inter-subject variability, which they define as the disparity in the EEG signals of any two subjects exposed to the same stimulus [62]. Hence, their proposed model maximizes the similarity between the representations of the EEG signals collected in response to identical stimuli. However, the effect of intra-subject variability was not considered in their work.

Inter-subject variability generally refers to the difference in brain functionality across different subjects, whereas intra-subject variability refers to the difference in brain functionality within one subject [63]. In this work, we propose a model that addresses the problem of both inter-subject variability, where different subjects are exposed to the same stimulus, and intra-subject variability, where one subject is exposed to different stimuli that evoke a similar emotional state. To the best of our knowledge, this is the first instance of using CL for EEG-based affect modeling to address both inter-subject and intra-subject variability.

E. CONTRIBUTIONS

We identify several challenges not addressed in existing EEG-based emotion estimation models. Firstly, the reported results for existing models typically pertain to the average of n-fold cross-validation without verification on testing datasets [21], [22], [32], [33], [34], [36], [38], [42], [43], [44], [52], [60], [61], [62]. Assessing the performance of the testing datasets is necessary to ensure that the models are not overfitting and to verify their generalizability. Secondly, most current models do not address inter-subject and, most importantly, intra-subject variability, which has posed significant challenges for emotion recognition [63], [64], [65]. EEG signals exhibit significant inter-subject variability in response to the same stimulus, leading to reduced robustness of trained classifiers for emotion recognition across different individuals. Furthermore, the problem of intra-subject variability further exacerbates the lack of robustness and generalizability many EEG-based emotion classifiers exhibit. Thirdly, in most existing work, the topological structure of EEG channels is not considered effectively, which may limit the model's ability to learn discriminative EEG representations. Lastly, most existing models are trained on limited datasets, given the difficulty and cost associated with data collection.

This paper proposes an integrated solution that addresses all the challenges mentioned above. Existing models have addressed aspects of these challenges; however, to the best of our knowledge, our proposed model is the first to provide a comprehensive solution to the identified challenges.

The main contributions of this paper are as follows:

- We propose a novel solution for recognizing emotions from EEG signals called Contrastive Learning GAN-based Graph Neural Network. It leverages self-supervised and supervised learning to capture high-quality EEG representations and address inter-subject and intra-subject emotion variability.
- We systematically investigate the effect of employing GAN data augmentation, CL, and GNN on emotion recognition performance by isolating each component and analyzing its impact through a comparison with several benchmark models.
- We test the proposed model on two popular benchmark emotion recognition datasets: DEAP [66] and MAHNOB-HCI [67]. Our experimental results indicate a testing emotion recognition accuracy of 64.84% and 66.98% for valence and 66.40% and 71.69% for the arousal classification task on the DEAP and MAHNOB databases, respectively. The results also show that our proposed model performs better than recent state-of-the-art EEG-based emotion recognition models.

The rest of this paper is organized as follows: in Section II, we introduce the main components of our proposed model, namely CL, GAN, and GNN, and present the datasets we employ to train, validate, and test our solution. In Section III, we describe the proposed solution. In Section IV, we perform experimental analysis and evaluation of the proposed model. Finally, in Section V, we provide concluding remarks and ideas for future work.

II. PRELIMINARIES

This section presents preliminary knowledge about the different components of our proposed model, including CL, GNN, and GAN, which form the basis for our proposed approach. Moreover, we describe the datasets we adopt for our work.

A. CONTRASTIVE LEARNING

CL is a self-supervised learning algorithm that captures a deep representation of data by maximizing the similarity between two signals (called a positive pair) using contrastive loss. CL has achieved state-of-the-art performance in various fields, such as bioinformatics [68], natural language processing (NLP) [69], and computer vision [54].

The overall structure of the CL component we deploy in our solution is depicted in Fig. 1. The component contains four sub-components: pair loader, channel encoder, channel projector, and contrastive loss function. The pair loader creates a batch of several positive pairs of EEG signals. The channel encoder extracts representations from EEG signals. The channel projector maps the extracted representations to another latent space to maximize their similarities using a contrastive loss function. The parameters of the channel encoder and channel projector are optimized such that the contrastive loss is minimized.

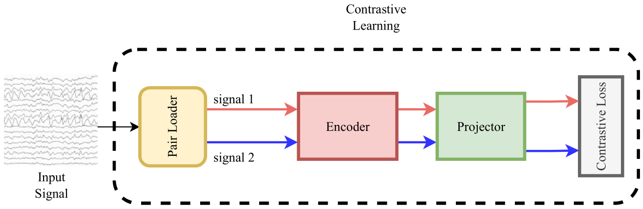


FIGURE 1. The CL architecture used for our proposed model.

In general, the most crucial step in CL is the choice of positive pairs, which highly impacts the quality of CL's output signal [61]. We propose a novel pairing method, which we describe in Section III-A.

B. GENERATIVE ADVERSARIAL NETWORK (GAN)

The original GAN is a network comprised of two competing parts: generator and discriminator, which are both parameterized as deep neural networks. The generator learns how to generate synthetic data that resembles the real data. The discriminator evaluates the probability that a sample originates from the real data distribution. In the training process of a GAN, the generator (G) attempts to deceive the discriminator (D) by generating synthetic data. In contrast, the discriminator tries to improve its discrimination to avoid being deceived by the synthetically generated data. The two parts are optimized simultaneously to reach Nash equilibrium eventually. The adversarial training procedure is formulated as a minimax problem, expressed as:

$$\underset{\theta g \theta_d}{minmax}\, L(X, Z) = E_{x_i} \sim X\,[log\,(D\,(x_i))] + E_{z_i} \sim Z[log\,(1 - D\,(G\,(z_i)))] \quad (1)$$

where θ_g and θ_d denote the parameters of the generator and discriminator, respectively. X represents the real data distribution, and Z represents a noise distribution, which can be uniform or Gaussian. The training of GAN involves two steps: maximizing the discriminator's loss and minimizing the generator's loss. In the first step, the optimal D is obtained by maximizing the above function with fixed G and Z. In the second step, the function is minimized to find the optimal G using the previously computed optimal D. Section III will elaborate on this minimax problem and describe the two steps in greater detail.

C. GRAPH REPRESENTATION

A graph is defined as G={V, E, W}, where V denotes a set of nodes with number |V| =N, E represents a set of edges connecting the nodes and W∈$R^{N\times N}$ represents an adjacency matrix defining the connection between any two nodes. The element w_{ij} of the adjacency matrix with i and j representing the row and column numbers, shows the weight which corresponds to the importance of the connection between nodes i and j. Data on $v$can be represented by $X \in R^{N \times d}$ where d represents the dimension of input features. In the proposed model, each EEG channel corresponds to a node, the relationship between each two channels corresponds to the edges of the graph, and the elements of the adjacency matrix describe the importance of the channels' relationship. A greater value of an element on the adjacency matrix indicates a closer relationship between the two channels.

To determine the elements of the adjacency matrix (w_{ij}), we can employ a Gaussian kernel function [65] expressed as follows:

$$w_{ij} = \begin{cases} exp\left(-\frac{[dist\ (i,j)]^2}{2\theta^2}\right) & if\ dist\ (i,j) \leq \lambda \\ 0 & otherwise \end{cases} \quad (2)$$

where λ and θ are two constants and dist(i,j) represents the Euclidean distance between channel i and j, which can be computed from the 3D channel coordinates found on the data sheet of the EEG recording device.

However, in this paper, we use a Graph Neural Network where we allow the entries of W to be learned dynamically [30] within the network instead of being prespecified. The above formula is only used for the initialization of the adjacency matrix.

D. DATASETS

To evaluate the performance of the proposed model, we conducted experiments on the publicly available DEAP [66] and MAHNOB-HCI [67] emotion recognition databases. The MAHNOB-HCI database was created under similar experimental conditions as the DEAP database. Although these datasets contain multiple modalities, we only utilized the EEG modality from the DEAP and MAHNOB-HCI databases to evaluate our proposed model.

The DEAP dataset includes EEG and peripheral physiological signals of 32 subjects aged between 19 and 37 years. This dataset has been widely used in research on emotion recognition [17], [28], [29], [35], [36]. During the dataset collection, subjects were asked to watch 40 segments of music videos that may evoke a variety of emotions. Each music video segment is 1-minute long. The subjects were also asked to rate the videos they watched based on the level of arousal, valence, dominance, and liking on a scale of 1 to 9 with 1 being the least intense. All un-processed EEG data was stored in BioSemi.bdf format at a 512Hz sampling rate. The recordings consist of a 3-second pre-trial baseline followed by a 60-second period during which the participant watches the music video, totaling a 63-second signal recording for each trial. However, for our work, we only used the pre-processed down-sampled EEG signals at 128Hz where the 3s pre-trial baseline recording and Electrooculography (EOG) artifacts which are caused by eyeball movement are removed. Moreover, in the pre-processed data, the EEG signals are band-pass filtered to preserve frequency components in our region of interest (from 4 to 45Hz). The EEG signals are averaged to the common reference signal to construct a spatial voltage distribution with zero-mean. In this dataset, EEG data were collected by a Biosemi ActiveTwo device with 32 active AgCl electrodes according to the international 10-20 system.

The second dataset is MAHNOB-HCI which presents data collected from 27 healthy adults aged between 19 and 40 years old. During the dataset collection, each subject was fitted with EEG and peripheral physiological sensors. each subject watched 20 music videos which resulted in 20 trials per subject. Although the length of the videos ranged from 94 to 176 seconds, only the recordings captured during the final 60 seconds of each stimulus were used for subsequent processing and analysis [71]. At the end of each trial, the subjects were asked to self-rate their arousal, valence, dominance, and sense of predictability on a scale of 1 to 9. Furthermore, they were told to self-report their emotions using emotional keywords. To ensure more efficient emotion recognition, we applied the same pre-processing steps that were used on the EEG signals in the DEAP dataset. The EEG data in this dataset was also recorded using a 32-channel Biosemi ActiveTwo device.

A summary of the DEAP and MAHNOB-HCI datasets information is presented in Table 1.

TABLE 1. Information about the DEAP and MAHNOB-HCI databases.

Feature	DEAP / MAHNOB-HCI Description
Number of subjects	32 / 27
Recorded Signals	EEG, respiration signal, PSG, EOG, EMG, GSR, skin temperature / EEG, respiration signal, ECG, GSR, skin temperature
Recorded video	face / face and body (6 cameras)
Number of experiments	40 /20
Number of EEG channels	32 / 32
Experiment length	60s (128Hz) / 94s-176s (256Hz)
Rating scales	Valence, Arousal, dominance and liking / Emotional keywords, Valence, Arousal, dominance, and predictability
Rating values	1-9 / 1-9

E. EMOTION MODELS

Emotions can be represented in various ways [72]. The most common approach is the categorical model, which assigns discrete labels to emotions such as fear and happiness. However, the categorical model has some limitations. For instance, emotions do not always have exact translations in different languages. For example, the word "disgust" does not have an exact equivalent in the Polish language [73]. Alternatively, emotions can be represented using a dimensional model [74]. One of the most widely used dimensional models specifies two dimensions, valence and arousal [75]. For our work, we focus on these two emotional dimensions and exclude other labels such as dominance, liking, and predictability, as they are not commonly used in the literature. Valence refers to the pleasantness of the emotional experience, ranging from negative (very unpleasant) to positive (very pleasant), while arousal refers to the intensity of the emotion, ranging from passive (very calm) to active (very excited). Discrete emotions can be located in the four quadrants of the valence-arousal (VA) dimensional model, as illustrated in Fig. 2. Therefore, discrete emotions can be estimated using this model. For example, if both valence and arousal are greater than 5, the emotional state falls into the first quadrant, which can correspond to an excited or happy emotion.

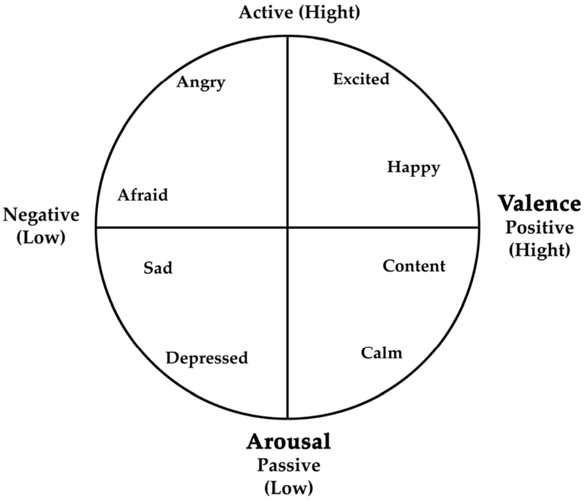


FIGURE 2. Valence-arousal dimensional emotion model.

In our work, we classify the valence and arousal emotional dimensions into two classes: high and low with a threshold set to 5 to distinguish between the high or low intensity of each emotional dimension, i.e., ratings from 6 to 9 refer to the high class and ratings from 1 to 5 refer to the low class.

III. PROPOSED MODEL

To ensure a fair performance assessment, we trained and evaluated our model using two strategies:

1. We shuffled all the trials and partitioned the dataset, allocating 80% for training and 20% for testing. This approach ensures that the model is never exposed to the test set during the training process.
2. We implemented a leave-one-subject-out cross-validation (LOSOCV) strategy. For each cross-validation fold, data from one subject is reserved for validation, while data from the remaining subjects is used for training. This method has its limitations as it results in a small testing set that corresponds to the data of a single subject. However, importantly, it provides a subject-independent evaluation strategy.

In both strategies, we augment the training data with synthetic data using a GAN component (Section III-B). The GAN component exclusively uses the training data to generate synthetic data.

Our model is comprised of three main components: CL, GAN, and GNN as depicted in Fig. 3. We will further describe these components in the subsequent sections.

A. CL COMPONENT

The overall architecture of the proposed model is presented in Fig. 3. The input signal is fed to the encoder of the CL component (Fig. 1) which in turn produces a latent representation. As described in Section II-A, the encoder is trained using a CL approach.

However, before deploying the encoder, the CL component (Fig. 1) needs to be trained and the CL loss must be minimized. CL has recently been adopted in the context of physiology-based emotion recognition [60], [61], [62]. However, in [60], the focus was mainly on creating augmented instances of each EEG signal to construct positive pairs, without addressing inter-subject and intra-subject emotion variabilities. In contrast, the CL method presented in [61] and [62] was applied to limited data, as they did not leverage augmented transforms of the EEG signals, which may cause overfitting issues. Additionally, in [62], the pairing was based on the EEG signals of the same trials over different subjects, neglecting intra-subject emotion variabilities for trials pertaining to a subject and evoking the same emotional state.

The CL component has a dual role. It acts as an encoder to extract high-quality features from the raw EEG signals and minimizes inter-subject and intra-subject emotion variabilities.

Fig. 1 shows that the first step for training the CL component is to load signal pairs. We propose a pairing mechanism that leverages trial categorization based on the valence-arousal emotional model to train the CL component. By doing so, we aim to maximize the representation similarity across EEG signals corresponding to the same emotional state, regardless of the subject or trial number. Therefore, in our CL strategy, the model learns to recognize whether two EEG signals correspond to the same emotional state.

To achieve our goal, we categorized the EEG data into four emotional categories: High Valence/High Arousal (HVHA), High Valence/Low Arousal (HVLA), Low Valence/High Arousal (LVHA), and Low Valence/Low Arousal (LVLA). Then, we employ a pair loader to create the mini-batches we use for training. We describe the pairing mechanism for the DEAP dataset below. The same approach was used for the MAHNOB-HCI dataset.

We flattened all EEG signals as:

$$S = \left\{s_{1,1,1}, s_{1,1,2}, \ldots, s_{i,j,k}\right\} \tag{3}$$

where s∈R^{7680} represents each DEAP EEG signal described in Section II-D) with 7680 samples. i represents the number

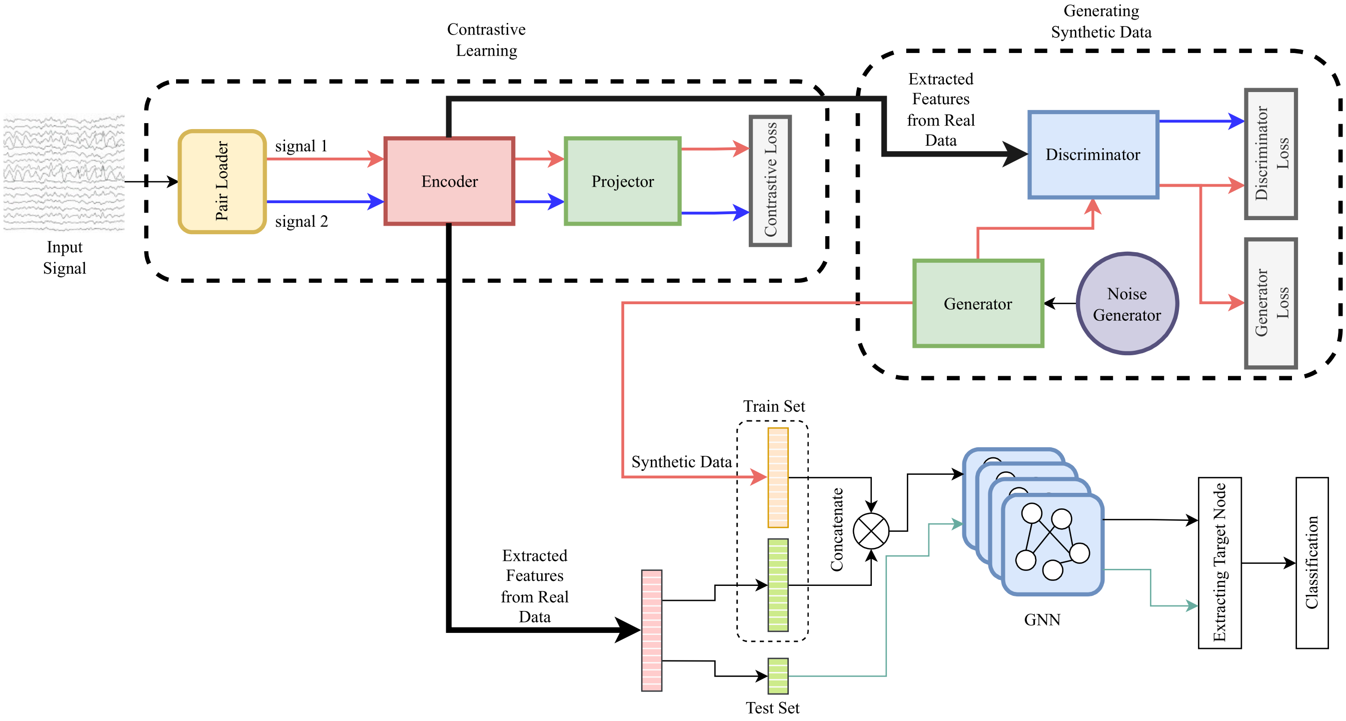


FIGURE 3. **Architecture of the proposed model.**

of subjects ($1 \leq i \leq 32$), j represents the number of trials ($1 \leq j \leq 40$) and k represents the number of EEG channels ($1 \leq k \leq 32$). We segmented the pre-processed EEG signals with a length of 20s as S becomes S' where $s' \in R^{2560\times3}$ to prepare the data to be fed to the channel encoder.

Then, we evenly divided the EEG signals across two arrays, A and B (Fig. 4) as follows:

$$\begin{aligned} A &= \{a_1, a_2, \ldots, a_n\} \\ B &= \{b_1, b_2, \ldots, b_n\} \\ Count(A) &= Count(B) \\ Count\left(A_{xy}\right) &= Count(B_{xy}) \end{aligned} \tag{4}$$

where a_n and b_n represent signals in arrays A and B, respectively ($1 \leq n \leq 16380$). *Count* is a function that returns the array length. Also, x and y represent valence and arousal labels, respectively. Where $x,y \in \{L,H\}$ and L and H represent low and high class for each label.

We used a pair loader to create mini-batches for training. The mini batches contained pair samples, with each signal in the pair originating from a different array. The signals in the mini-batches were pulled randomly from the arrays as follows:

$$pair\_loader\ (l) \rightarrow A_{minibatch}, B_{minibatch} \tag{5}$$

where $\boldsymbol{A}_{\text{minibatch}}$ and $\boldsymbol{B}_{\text{minibatch}}$ are mini-batches of size l that are loaded from the $\boldsymbol{A}$ and $\boldsymbol{B}$ arrays, respectively. Each signal from $\boldsymbol{A}_{\text{minibatch}}$ is paired with all the signals in $\boldsymbol{B}_{\text{minibatch}}$. Therefore, the total number of pairs formed using both mini-batches is 1 × 1. We considered a pair as positive if both of its signals had the same label, and we considered it as negative otherwise. Positive and negative pairs are labeled as 1 and 0, respectively.

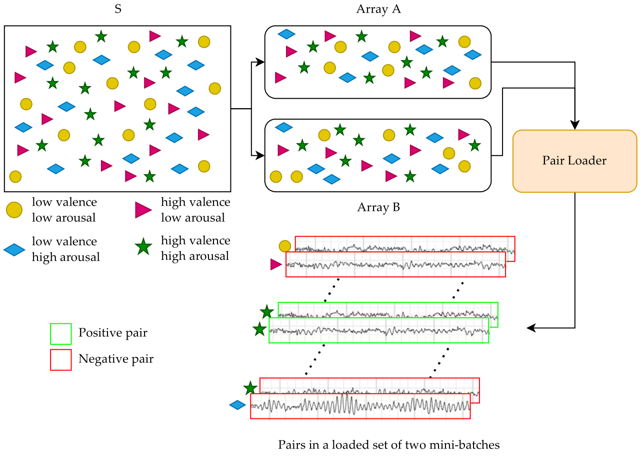


FIGURE 4. **Demonstration of the proposed pairing model based on emotional categories.**

As described in Section II-A, the CL component we use consists of four sub-components. Therefore, when pairs are loaded, the EEG signals of any typical positive pair of a_i and b_i are passed to the channel encoder of the CL component to generate high-quality inter-subject/intra-subject aligned representations for the EEG signals over trials with the same emotional state.

The output of the encoder is forwarded to a simple multilayer perceptron channel projector. As it was found useful in [54], we apply the contrastive loss on the output of the channel projector.

As in the contrastive loss function used in [57] and [60], we deploy the normalized temperature-scaled cross-entropy loss which has been modified according to our proposed pairing mechanism. The loss attempts to increase the similarity between the two EEG signals of a positive pair. The contrastive loss function for our proposed pairing mechanism is defined as follows:

$$loss\left(A_{minibatch}, i\right) = -\sum\nolimits_{j} \log \frac{exp\left(sim\left(a'_i, b'_j\right)/\tau\right)}{1+\sum\nolimits_{k=1}^{l} \mathbb{I}\left(a_i, b_k\right) exp\left(sim\left(a'_i, b'_k\right)/\tau\right)} \tag{6}$$

where l is the total number of pairs that can be constructed with an EEG signal a_i from $\boldsymbol{A}_{\text{minibatch}}$, j represents indices of signals in $\boldsymbol{B}_{\text{minibatch}}$ which construct a positive pair with a_i, τ is the temperature parameter to adjust the scaling of the similarity scores, and a'_i and b'_i are the output of the channel projector in response to signals a_i and b_i for a positive pair derived from:

$$a'_i = projector(encoder(a_i)) \tag{7}$$

sim(a,b)is the cosine similarity of a and b which is calculated as follows:

$$sim(a, b) = \frac{a.b}{\parallel a \parallel \parallel b \parallel} \tag{8}$$

where, $||a||$ and ||b|| are the Euclidean norms of a and b, respectively. $\mathrm{I(a_i,b_k)}$ $\in\{0,1\}$ which is set as 1 if a_i and $\mathrm{b_k}$ makes a negative pair otherwise, it is 0.

The final loss of the set of two mini-batches is computed as follows:

$$L = \sum\nolimits_{i=1}^{l} \frac{Loss\left(A_{minibatch}, i\right) + Loss\left(B_{minibatch}, i\right)}{2} \tag{9}$$

Our channel encoder and channel projector network architectures are inspired by [60] and presented in Table 2 and Table 3, respectively. The hyper-parameters of these networks were further manipulated and the networks with the presented parameters achieved the top performance. k, f, and s refer to kernel size, filter size, and stride number, respectively.

For the CL training, we used a mini-batch size of 30 and trained the model for 8 epochs on the training set. Algorithm 1 presents the pseudo-code for the proposed CL component.

After completing the CL training, we integrated the trained encoder of the CL component into the overall solution presented in Fig. 3.

TABLE 2. The architecture of the channel encoder.

Layer	Parameter	Activation
Conv1D	k=20, f=100, s=2	ReLU
Batch Norm	—	—
Conv1D	k=10, f=90, s=2	ReLU
Batch Norm	—	—
Conv1D	k=5, f=50, s=2	ReLU
Batch Norm	—	—
Conv1D	k=3, f=10, s=2	ReLU
Batch Norm	—	—
Flatten	—	—
Dense	70	Sigmoid
Batch Norm	—	—

TABLE 3. Architecture of the channel projector.

Layers	Parameter	Activation
Dense	100	ReLU
Dense	10	—

B. GAN COMPONENT

Due to the small size of the dataset which may cause overfitting issues in the training process, we increased the number of data samples using GAN to generate synthetic realistic-like data.

Therefore, as depicted in Fig. 3, the features extracted from the CL's channel encoder are the real data that are fed to a GAN to generate the synthetic data.

Equations (10) and (11) show the discriminator and generator's loss functions, respectively.

$$loss_D = \nabla_{\theta_d} \frac{1}{m}\left[logD\left(x^i\right) + log\left(1 - D\left(G\left(z^i\right)\right)\right)\right] \tag{10}$$

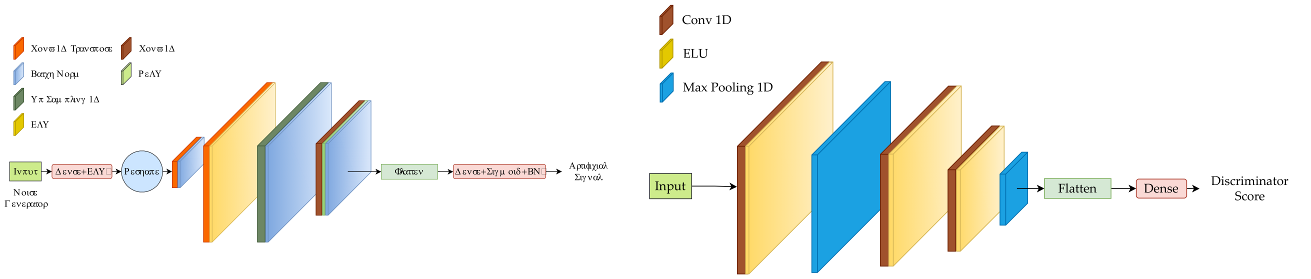
$$loss_G = \nabla_{\theta_g} \frac{1}{m} log\left(1 - D\left(G\left(z^i\right)\right)\right) \tag{11}$$

where G is a generator, D is a discriminator, $\mathrm{z^i}$ $\in\mathrm{R}^{1\times256}$ is ith sample noise vector with i representing the trial number, $\mathrm{x^i}$ $\in\mathrm{R}^{1\times32\times70}$ is our ith input data which is aimed to be reconstructed. θ_d is the parameter set for the discriminator, and θ_g is the parameter set for the generator.

We utilized the GAN component to produce four distinct sets of labeled synthetic data representing high valence, low valence, high arousal, and low arousal.

To train the GAN component, we used 100, 100, and 256 as batch size, number of epochs, and code size, respectively. Algorithm 2 illustrates the pseudo-code of the proposed GAN network.

After generating synthetic data using GAN, the synthetic data is appended to the real data obtained from the channel encoder of CL. Then, we conducted two rounds of training and classification of the GNN, the first time for the valence



(a)

(b)

FIGURE 5. Architecture of the GAN network. (a) Architecture of the generator which generates synthetic data (b) Architecture of the discriminator which distinguishes the real and synthetic data.

Algorithm 1 Contrastive Learning Algorithm

```
Inputs: Training data $\{A, B\}$, the learning rate $\alpha$, the mini-batch size $l$, the training epochs T
1: initialize parameters of the base encoder $\theta_e$ and the projector $\theta_p$
2: for epoch = 1 to T do
3: repeat
4:     sample $l$ signals from $A$ and $B$
5:     obtain $\{\mathrm{sim}_{i,j} = 1, 2, \ldots 1\}$ by (8)
6:     calculate loss by (9)
7:     update $\theta_p$ and $\theta_p$ by loss with $\alpha$
8: until all possible pairs enumerated
Outputs: Features of data using parameters $\theta_e$
```

Algorithm 2 GAN Algorithm

```
Inputs: Features of training trial's data $\{X\}$, the learning rate $\alpha_g$ and $\alpha_d$, the Batch size N, the training epochs T
1: initialize parameters of the base generator $\theta_g$ and the discriminator $\theta_d$
2: for epoch = 1 to T do
3: for iterate = 1 to 5 do
4:     sample from trials $\{x_i | i = 1, \ldots, N\}$
5:     sample from noise generator $\{z_i | i = 1, \ldots, N\}$
6:     calculate loss by (10)
7:     update $\theta_d$ by loss with $\alpha_d$ learning rate
6:     sample from noise generator $\{z_i | i = 1, \ldots, N\}$
7:     calculate loss by (11)
8: update $\theta_g$ by loss with $\alpha_g$ learning rate
9: generate synthetic data using $\theta_g$
Outputs: synthetic realistic-like data
```

and the second time for the arousal dimension. We used the DEAP and MAHNOB-HCI databases for this study with 1280 and 530 total number of trials, respectively.

As in [53], the discriminator parameters were updated 5 times per epoch, while the generator parameters were updated once per epoch.

The generator and discriminator networks used in this study were optimized through a trial-and-error process. Table 4 and Table 5 show the deployed architectures for the generator and discriminator networks, respectively. The architecture of the GAN component is illustrated in Fig. 5.

TABLE 4. Architecture of the generator network.

Layer	Parameters	Activation Function
Dense	768	ELU
Reshape	(3, 256)	—
Conv1D Transpose	k=5, f=40, s=1	—
Batch Norm	—	—
Conv1D Transpose	k=5, f=40, s=1	ELU
UpSampling1D	k=2	—
Conv1D Transpose	K=5, f=35, s=1	—
Batch Norm	—	—
Conv1D	k=5, f=32, s=1	ReLU
Batch Norm	—	—
Flatten	—	—
Dense	2240	Sigmoid
Reshape	(32, 70)	—
Batch Norm	—	—

TABLE 5. Architecture of the discriminator network.

Layer	Parameters	Activation Function
Conv1D	k=5, f=50, s=1	ELU
MaxPool1D	k=2	—
Conv1D	k=5, f=50, s=1	ELU
Conv1D	k=5, f=40, s=1	ELU
MaxPool1D	k=3	—
Flatten	—	—
Dense	1	—

TABLE 6. Architecture of the proposed GNN.

Layer	Number of hidden units	Activation Function	Dropout rate
Dense	45	Sigmoid	
Dropout	—	—	22%
Dense	26	Sigmoid	
Dense	7	Sigmoid	
Dropout	—	—	5%
Dense	2	Softmax	

C. GNN COMPONENT

The appended data is fed to the GNN component for classification. We propose a GNN with multiple linear layers. Each layer is composed of a dense layer, sigmoid activation function. We also deploy the dropout layer on the first and before the last dense layer. We also applied regularization over the network.

First, we setup a dynamic square adjacency matrix with a dimension corresponding to the number of EEG channels (32 for both DEAP and MAHNOB-HCI databases). The following equation describes each layer's forward propagation:

$$\begin{aligned} output &= WX_{i,j} + b,\ \ W \in R^{u\times l} \\ X_{i,j} &= S_{i,j}A,\ \ S_{i,j} \in R^{l\times c},\ \ A \in R^{c\times c} \end{aligned} \tag{12}$$

where $S_{i,j}$ is the extracted features corresponding to the ith subject and jth experiment. W and b are the parameters of the layer. A is the adjacency matrix where the *ith* row and jth column define the relation between the *ith* node and jth one. c is the number of channels, l is the length of extracted features, and u is the number of hidden units of the corresponding linear layer.

The architecture of our proposed GNN is described in Table 6. The hyper-parameters of the proposed GNN model are optimized using Gaussian search [76]. We adopted the learning rate, dropout, layer's output size (hidden size), and the number of dense layers for hyperparameter optimization. The hyperparameter optimization process rendered the following values: a learning rate of 0.00005, 4 dense layers with hidden sizes of [50, 28, 7, 2], respectively, and a dropout rate of 22% and 5% as in Table 6. Algorithm 3 presents the pseudo-code of the proposed GNN.

We introduced a new parameter to our network, which we call the "target node", to update the weights during backpropagation. The model is fed input features of a target node and its neighboring nodes and is tasked with predicting the target output value. Target node optimization in a graph neural network involves training a model to predict the properties or behaviors of specific target nodes in a graph. Node-level backpropagation attempts to categorize nodes into several classes, which can improve performance.

Algorithm 3 GNN Algorithm

```
Inputs: Features of training trial's data {X}, Features of synthetic trial's data {F}, the learning rate $\alpha_c$, the batch size N, the training epochs T, the number of channels C
initialize parameters of the GNN $\theta_{gnn}$ and the adjacency matrix $\{A_{i,j}|i, j = 1, \ldots, C\}$ by (2)
$\{S\} \leftarrow$ merge {X} and {F}
for epoch = 1 to T do
repeat
    sample from trials $\{s_i|i = 1, \ldots, N\}$
    calculate the output of all layers by (12)
    get p-value by (13)
calculate loss with p-value
    update $\theta_{gnn}$ parameters and $A_{i,j}$ by loss with $\alpha_c$
until all possible batches enumerated
Outputs: $\theta_{gnn}$ and $A_{i,j}$ parameters
```

In our paper, we considered the number of nodes to be equal to the number of EEG channels in each trial, which is 32. We used grid search to optimize this parameter before performing hyperparameter optimization. Hence, from (12) we have:

$$out \in R^{b\times 32\times 2} p = out_t 1 \leq t \leq 32, out_t \in R^{b\times 2} \tag{13}$$

where *out* is the last output from our GNN. b is the batch size of the processed data. t refers to the target node and p denotes the output processed data of the corresponding node which is used further for the calculation of the loss.

For the loss function and optimizer, we used Categorical Cross Entropy loss and Adam optimizer. The batch size and number of epochs are 100 and 400, respectively. However, we used early stopping to avoid overfitting.

IV. RESULTS AND EXPERIMENTAL ANALYSIS

This section describes the experimental setup and evaluation of the proposed model using two testing scenarios. Firstly, we performed an ablation study to analyze the impact of each component (CL, GAN, and GNN) of the proposed model on improving the emotion classification accuracy. To do so, each component is replaced with a competing similar component from the literature or omitted entirely. Secondly, we compared the performance of the proposed model with that of several recent competing emotion recognition models. We implemented the existing models and trained and evaluated them on the same datasets for a fair comparison. All the models were implemented using the Keras framework libraries with a Tensorflow backend in Python and trained on an NVIDIA GeForce RTX 2080 Ti GPU.

A. EXPERIMENTAL SETUP

We performed two sets of evaluations. The first evaluation involved the partitioning of the dataset into training and testing subsets. The training set is used to train both the proposed model and existing models (for comparison purposes), computing cross-entropy loss, and updating model parameters using the Adam optimizer [77]. The testing set is used to evaluate the trained model's ability to identify the level of arousal and valence of the testing samples.

To split a dataset into training and testing subsets, we shuffled all trials for different subjects and separated 20% of the data as testing samples while using the remaining 80% for training. We used 1024 trials as training datasets and 256 trials as the testing dataset for the DEAP database. Similarly, for the MAHNOB-HCI database, we used 424 and 106 trials as training and testing datasets, respectively. Moreover, we performed four-fold cross-validation on the training set and reported the average accuracy of the four folds as the training accuracy.

For the second evaluation, we assess the performance of the proposed model in a subject-independent manner where the data of one subject was excluded as a testing dataset and the data of the remaining subjects (31 and 26 for DEAP and MAHNOB-HCI, respectively) were used for training and validating of the proposed model using the LOSOCV evaluation strategy. We reported the average accuracy of the folds (31 folds and 26 folds for DEAP and MAHHNOB-HCI datasets, respectively) as the training accuracy. This test provides an unbiased estimate of the model performance for individual subjects since each subject serves as a test set in LOSOCV.

For both evaluations, when comparing to existing state-of-the-art methods, we detail the accuracy of the training and testing subsets in the case of valence and arousal. Additionally, we provide average accuracy across both emotional dimensions for the training and testing datasets.

B. FIRST EVALUATION STRATEGY: SPLITTING DATASET INTO TRAINING AND TESTING SETS

In this section, we evaluate the proposed model by splitting the dataset into training and testing subsets. We start with an ablation study (Section IV-B.1) and then proceed to compare the proposed model to state-of-the-art models on the same datasets (Section IV-B.2).

1) COMPONENT-BASED ANALYSIS (ABLATION STUDY)

a: PERFORMANCE OF THE GAN COMPONENT

In this section, we evaluate the performance of the GAN component of the proposed model based on the amount of generated data added to the training set for both the DEAP and MAHNOB-HCI databases. Table 7 and Table 8 present the performance of the proposed model for different amounts of synthetically generated data appended to the training sets for the two databases. The amount of appended synthetic data that results in the best performance is used for further classification.

The training and testing accuracies of the valence and arousal emotional dimensions are used to determine the amount of synthetic data to be appended to the training sets. The amount of appended data is expressed as a multiple of the real training set. For instance, 0 denotes that we will use only the original real training dataset without any synthetic data added. On the other hand, 0.5 indicates that we will add synthetic data that corresponds to half the size of the original real set. Specifically, we would add 512 synthetic trials for the DEAP database and 212 synthetic trials for the MAHNOB database.

TABLE 7. Performance evaluation on the volume of appended synthetically generated EEG data in DEAP dataset.

Amount of appended Synthetic data	Valence		Arousal	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.
× 0 dataset (0)	66.01	59.37	69.33	62.10
× 0.5 dataset (512)	72.39	62.10	70.89	63.67
× 1 dataset (1024)	**74.65**	**64.84**	**74.21**	**66.40**
× 1.5 dataset (1536)	76.71	64.06	73.63	64.84
× 2 dataset (2048)	77.53	64.84	74.27	64.45

TABLE 8. Performance evaluation on the volume of appended synthetically generated EEG data in MAHNOB-HCI dataset.

Amount of appended Synthetic data	Valence		Arousal	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.
× 0 dataset (0)	72.68	66.03	73.15	61.32
× 0.5 dataset (512)	**74.32**	**66.98**	**78.12**	**71.69**
× 1 dataset (1024)	72.92	66.03	71.25	62.26
× 1.5 dataset (1536)	74.32	63.20	76.42	67.92
× 2 dataset (2048)	77.88	64.15	74.50	64.15

From the tables, we can see that adding a number of synthetic samples equal to that of the real ones (i.e., the amount of appended synthetic data is equal to 1) achieves the best results for the DEAP database while adding synthetic samples that correspond to half the number of real samples (i.e., amount of appended synthetic data is equal to 0.5) performs best for the MAHNOB database based on the accuracies obtained for the valence and arousal classification tasks. However, both tables show that employing synthetic data still improves the performance of the model. For further analysis, we selected 1 and 0.5 times as the amount of appended synthetic data for the DEAP and MAHNOB-HCI databases, respectively, as they achieve the best accuracies on both emotional dimensions and the least discrepancy between the training and testing datasets.

b: PERFORMANCE OF THE CL COMPONENT

In the next step, we compared the performance of the encoder from the proposed CL component to other feature extractors, namely SeqCLR Encoder [60] and CLISA Encoder [62], which are encoders trained using CL with their proposed pairing mechanism for EEG-based classification, Label-Based

TABLE 9. Performance evaluation on different CL pairing methods and feature extractors on DEAP dataset.

Method	Valence		Arousal	
	Train	Test	Train	Test
SeqCLR Encoder [60]	87.84	64.06	69.33	64.06
Label-Based CL Encoder [61]	80.95	60.54	74.41	62.89
CLISA Encoder [62]	65.11	61.66	75.96	65.41
VGG16 [31]	90.86	60.15	92.52	62.10
Attention Encoder [42]	74.65	59.76	76.26	61.71
Proposed CL component Encoder	**74.65**	**64.84**	**74.21**	**66.40**

CL Encoder [61], which is an encoder trained using CL with their proposed pairing mechanism for multi-modal classification, Attention Encoder [42], which is an encoder with an attention mechanism for EEG-based emotion classification, and VGG16 [31], which is a widely used CNN for feature extraction employed in various applications. To evaluate the performance of a feature extractor, we replace the proposed encoder in our model (shown in Fig. 3) with the feature extractor. Then, we train and test the resulting model on the DEAP dataset. Our comparison results are presented in Table 9. As shown in Table 9, the proposed CL component encoder outperforms the other components.

We evaluated the use of VGG16 [31] for feature extraction from EEG signals. Although VGG16 is a deep convolutional neural network typically used for image feature extraction, it has been employed for other feature extraction applications as well. Based on our results, we observed that using the VGG16 feature extractor caused the model to overfit, leading to high performance on the training dataset but relatively low performance on the testing dataset. The Attention Encoder proposed in [42] performs relatively well, but our proposed CL component encoder shows approximately a 5% improvement in testing accuracies for the valence and arousal emotion classification. The CLISA Encoder [62] achieves more reliable results than other existing models, but our proposed CL component encoder achieves higher testing accuracies for both valence and arousal classification. The SeqCLR Encoder [60] and Label-Based CL Encoder [61] show a relatively high discrepancy between training and testing accuracies, which might be due to overfitting.

c: PERFORMANCE OF THE GNN

In the last step, we investigated the effect of using GNN as a classifier by comparing it with some benchmark classification models presented in Table 10. All evaluated classifiers have been proposed for EEG-based emotion classification in recent work [25], [28], [38], [47]. To evaluate the performance of a classifier, we replace the proposed GNN classifier in our model (shown in Fig. 3) with the classifier. Then, we train and test the resulting model on the DEAP dataset.

TABLE 10. Performance evaluation of various classification models on the DEAP dataset.

Model	Valence		Arousal	
	Train	Test	Train	Test
DNN [41]	99.99	52.73	79.79	48.44
CNN [30]	98.73	52.34	74.22	53.52
BiLSTM [27]	95.07	56.25	84.64	60.16
SVM [52]	84.76	50.78	66.16	59.76
Proposed GNN	**74.65**	**64.84**	**74.21**	**66.40**

As shown in Table 10, the choice of classifier has a significant impact on the emotion classification accuracy. Our proposed classifier outperformed the existing popular classification models. The training and testing accuracies for the valence and arousal dimensions have a high discrepancy for the DNN [41], CNN [30], and BiLSTM [27] models which indicates that the model overfits the training set. SVM [52] performs more reliably on the arousal dimension but performs poorly on the testing dataset for both the arousal and valence dimensions. BiLSTM [27] achieves better testing accuracies compared to DNN [41] and CNN [30]. Nonetheless, our proposed classifier still showed a significant improvement over the other methods.

2) PERFORMANCE COMPARISON

In this section, we compare our proposed model with several recent emotion classification models on both the DEAP and MAHNOB-HCI datasets. The DEAP dataset is particularly valuable as it has the highest number of participants compared to other publicly available EEG datasets for emotion recognition.

We compared our proposed model with existing benchmark EEG-based deep models, namely:

- Attention-based LSTM combined with domain discriminator denoted as ATDD-LSTM, where DE features from different frequency bands, are used as input [38]
- Attention-based convolutional recurrent neural network (ACRNN) using EEG features of time and frequency domains as input [29]
- Graph convolutional neural network combined with LSTM (ECLGCNN) with DE features as input [36]
- Conditional Wasserstein GAN (CWGAN) using DE features as input [52]
- CapsNet with attention mechanism (ACapsNet) using segmentations of raw EEG signals as input [32]
- NAS-optimized emotion recognition model with raw EEG signals as input [22]
- Multi-task learning using DF with a manual 2D frame for each sample according to EEG channels' distribution as input [21]

- Attention-based LSTM autoencoder (ALSTM autoencoder) + attention-based CNN (ACNN) using raw EEG signals with additive white Gaussian noise as input [42]
- Hierarchical Spatial Learning Transformer (HSLT) model with DSP features as input [43]
- EEG emotion Transformer model (EeT) using segmentations of raw EEG signals as input [44]
- Contrastive learning followed by three dense layers for emotion recognition (CLISA) using EEG signals as input [62].

Table 11 shows the emotion recognition accuracies on the training and testing sets for the proposed and existing competing models on the valence and arousal for the DEAP database. Our model outperforms competing EEG-based emotion recognition models with significant improvement. However, the models presented in [62], [29], [32], [21], [22], [42], [43], and [44] achieve a relatively high training accuracy but present a significant discrepancy between the training and testing accuracies which may be due to overfitting on the training set. The models presented in [36], [38], and [52] achieve the least discrepancy between the training and testing accuracies, but overall, these three models perform poorly for emotion classification compared to the proposed model. Specifically, the GAN-based model presented in [52] achieves the highest training and testing emotion classification among the eleven state-of-the-art methods, while maintaining a moderate discrepancy between the training and testing set results. The model presented in [52] expands the dataset synthetically using a GAN-based method. Hence, the robust performance of this model demonstrates the positive effect of dataset expansion on deep learning model performance. It suggests that by incorporating a larger and more diverse dataset, there is a potential enhancement in the model's performance. The proposed model achieves a training accuracy of 74.65% and 74.21% for valence and arousal, respectively. The same model achieves a testing accuracy of 64.84% and 66.40% for valence and arousal, respectively.

This enhancement of the proposed model indicates that the EEG representation learning, achieved by the proposed CL component, outperforms simple DE features in emotion recognition classification.

We have also calculated the average accuracy over the valence and arousal emotional dimensions across all models to assess their performance relative to the proposed model. As can be seen from Table 11, the proposed model achieves a higher average emotion recognition accuracy on the test and train datasets with less discrepancy between the average accuracies on the training and testing datasets. However, the model presented in [52] demonstrates a 0.39% increase in the average test accuracy compared to the proposed model. However, the proposed model achieves 4.76% higher training accuracies on the valence and arousal emotion dimensions than the model presented in [52]. The model presented in [36] achieves a moderate average accuracy for the test dataset compared to the other nine benchmark models, which is 7.17% lower than the average test accuracy of the proposed model. However, this model results in a low average training accuracy of 57.57%.

TABLE 11. Comparison of emotion recognition models on DEAP database.

Model	Valence Accuracy (%)		Arousal Accuracy (%)		Average Accuracy for Valence and Arousal (%)	
	Train	Test	Train	Test	Train	Test
CLISA [62]	90.72	50.59	85.11	58.79	87.91	54.69
ACRNN [29]	68.26	45.31	79.00	55.47	73.63	50.39
ECLGCNN [36]	56.64	56.25	58.50	60.55	57.57	58.45
ATDD LSTM [38]	54.93	51.12	59.37	55.76	57.15	53.44
CWGAN [52]	69.92	61.71	69.43	70.31	69.67	66.01
ACapsNet [32]	99.90	51.56	96.19	50.39	98.04	50.97
NAS [22]	99.31	52.73	98.24	48.04	98.77	50.38
DF [21]	96.87	53.51	95.01	55.46	95.94	54.48
ALSTM autoencoder + ACNN [42]	99.80	54.68	99.12	57.42	99.46	56.05
HSLT Transformer [43]	73.14	42.18	71.48	54.29	72.31	48.23
EeT Transformer [44]	88.96	48.43	87.10	53.51	88.03	50.97
Proposed model	**74.65**	**64.84**	**74.21**	**66.40**	**74.43**	**65.62**

We evaluated the effectiveness of our proposed model using the MAHNOB-HCI database. Table 12 presents the training and testing accuracies of the compared models for the MAHNOB-HCI database on the valence and arousal emotion dimensions. The results indicate that the proposed model outperforms the competing recent models with 74.32% and 78.12% as training accuracies and, 66.98% and 71.69% as testing accuracies for valence and arousal, respectively. Despite achieving a moderate level of discrepancy between the training and testing accuracies, the model presented in [62] performs relatively poorly on emotion classification compared to our proposed model. As can be seen from Table 12, the average training and testing accuracies for the proposed model are relatively higher compared to state-of-the-art models, coupled with a moderate discrepancy between these accuracies. However, the GAN-based model presented in [52] shows the lowest discrepancy between the average training and testing accuracies compared to the proposed model. Despite this, the proposed model surpasses the performance of the GAN-based model in [52] for the training and testing sets for both emotional dimensions assessed.

TABLE 12. Comparison of emotion recognition models on MAHNOB-HCI database.

Model	Valence Accuracy (%)		Arousal Accuracy (%)		Average Accuracy for Valence and Arousal (%)	
	Train	Test	Train	Test	Train	Test
CLISA [62]	54.15	50.00	61.26	60.37	57.70	55.18
ACRNN [29]	84.60	45.28	83.41	61.32	74.00	53.30
ECLGCNN [36]	99.76	52.83	99.29	41.50	99.52	47.16
ATDD LSTM [38]	99.53	37.73	99.82	50.00	99.67	43.86
CWGAN [52]	66.98	66.03	66.50	67.92	66.74	66.97
ACapsNet [32]	99.52	37.73	97.64	34.90	98.58	36.31
NAS [22]	99.76	45.28	99.29	45.28	99.52	45.28
DF [21]	91.74	56.60	90.80	53.77	91.27	55.18
ALSTM autoencoder + ACNN [42]	91.98	45.28	95.28	42.45	93.63	43.86
HSLT Transformer [43]	68.39	53.77	70.28	50.00	69.33	51.88
EeT Transformer [44]	94.10	47.16	92.45	49.05	93.27	48.10
Proposed model	**74.32**	**66.98**	**78.12**	**71.69**	**76.22**	**69.33**

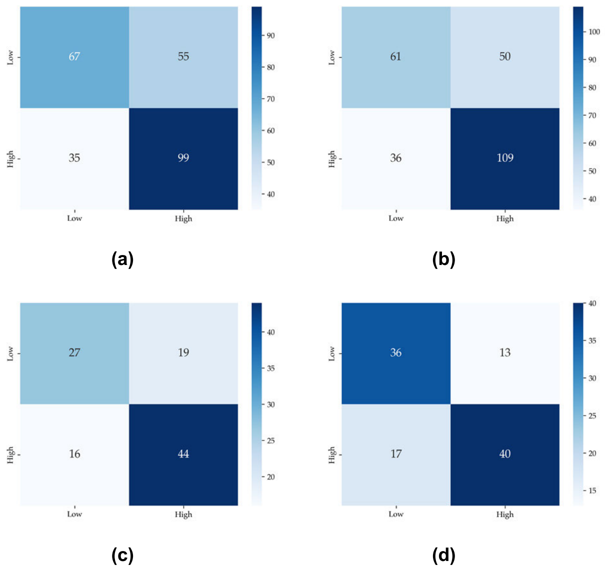


FIGURE 6. Confusion matrices of the proposed model. (a) Valence- DEAP dataset (b) Arousal-DEAP dataset (c) Valence-MAHNOB dataset. (d) Arousal-MAHNOB dataset.

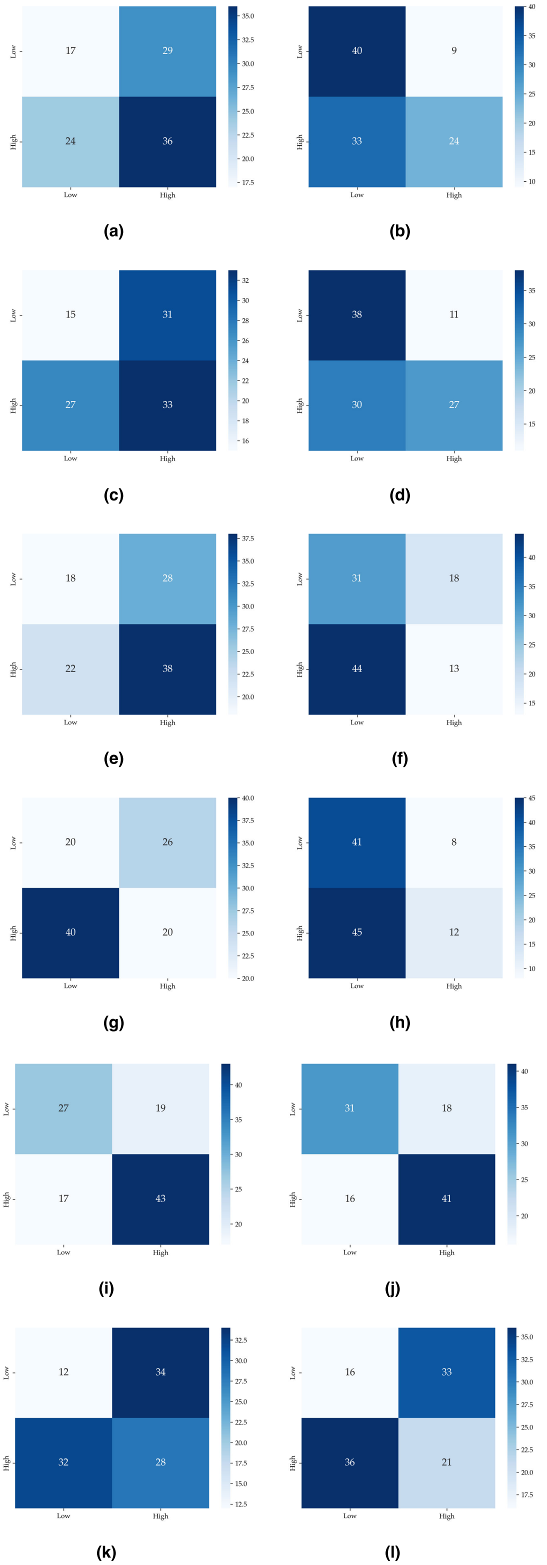
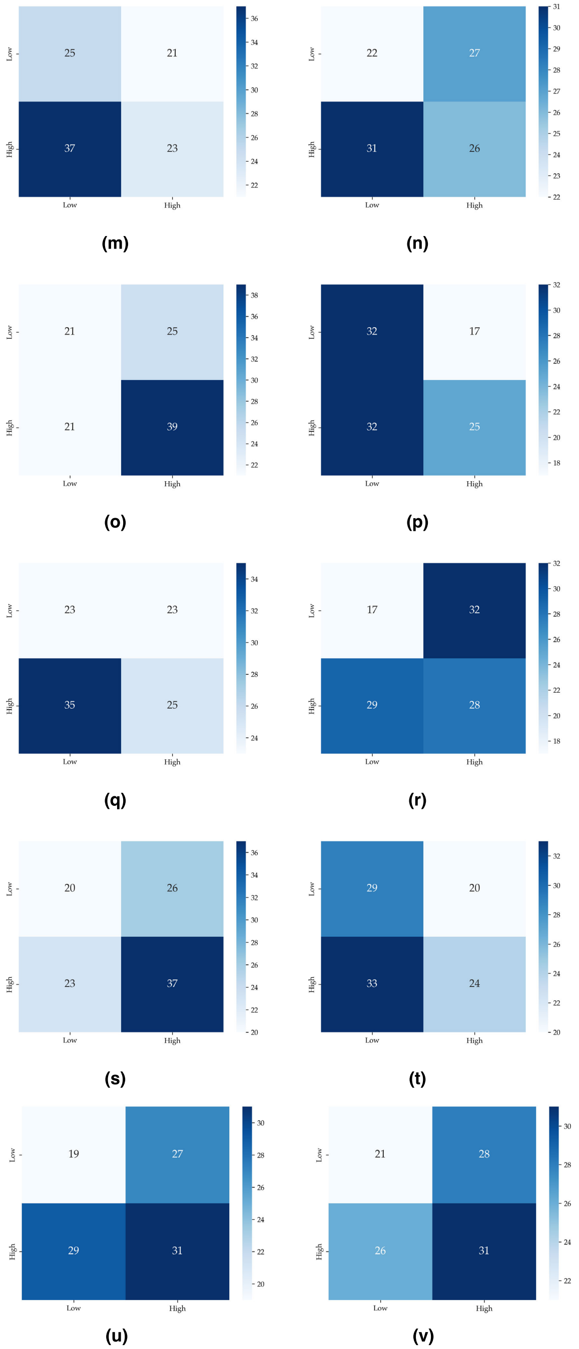


FIGURE 7. Confusion matrices of (a)Valence-CLISA [62] (b) Arousal-CLISA [62] (c) Valence-ACRNN [29] (d) Arousal-ACRNN [29] (e) Valence-ECLGCNN [36] (f) Arousal-ECLGCNN [36] (g) Valence-ATDD LSTM [38] (h) Arousal-ATDD LSTM [38] (i) Valence-CWGAN [52] (j) Arousal-CWGAN [52] (k) Valence-ACapsNet [32] (l) Arousal-ACapsNet [32].

The average testing and training accuracies over valence and arousal for the proposed model are 69.33% and 76.22 % with a 6.89% drop which is significantly lower than the difference between the average training and testing accuracies over



(m) (n) (o) (p) (q) (r) (s) (t) (u) (v)

FIGURE 7. *(Continued.)* **Confusion matrices of (m) Valence-NAS [22] (n) Arousal-NAS [22] (o) Valence-DF [21] (p) Arousal-DF [21] (q) Valence-ALSTM autoencoder + ACNN [42] (r) Arousal-ALSTM autoencoder + ACNN [42] (s) Valence-HSLT Transformer [43] (t) Arousal-HSLT Transformer [43] (u) Valence-EeT Transformer [44] (v) Arousal-EeT Transformer [44].**

TABLE 13. **LOSOCV of the proposed model on DEAP database.**

Model	Valence Accuracy (%)		Arousal Accuracy (%)		Average Accuracy for Valence and Arousal (%)	
	Train	Test	Train	Test	Train	Test
CLISA [62]	50.36	50.00	50.00	50.00	50.18	50.00
ACRNN [29]	80.24	45.00	93.87	40.00	87.05	42.50
ECLGCNN [36]	56.77	50.00	58.55	70.00	57.66	60.00
ATDD LSTM [38]	56.21	67.50	59.19	50.00	57.70	58.75
CWGAN [52]	71.29	52.50	70.72	62.50	71.00	57.50
ACapsNet [32]	99.83	45.00	98.62	47.50	99.22	46.25
NAS [22]	99.35	37.50	98.14	55.00	98.74	46.25
DF [21]	96.12	52.50	95.96	52.50	96.04	52.50
ALSTM autoencoder + ACNN [42]	98.87	55.00	99.03	55.00	98.95	55.00
HSLT Transformer [43]	75.88	47.50	73.30	52.50	74.59	50.00
EeT Transformer [44]	87.82	52.50	85.48	50.00	86.65	51.25
Proposed model (LOSOCV)	**68.77**	**57.50**	**69.85**	**70.00**	**69.31**	**63.75**

the benchmark emotion recognition models. This indicates that the model has learned the relevant patterns from the training data without memorizing and consequently, it can generalize well to new unseen data.

We present the confusion matrices for the test set for the proposed model for both databases in Fig. 6. The proposed model achieved a better performance in the classification of high arousal and high valence for the DEAP dataset (73.88% and 80.74%, respectively) compared to the MAHNOB-HCI dataset (73.33% and 70.17%, respectively), which could be due to the testing set in the DEAP dataset being slightly unbalanced in favor of high valence and high arousal.

We present the confusion matrices for the existing models for the MAHNOB-HCI database in Fig. 7. We observe that the values on the diagonal of the proposed model's confusion matrices (Fig. 6(c) and (d)) are relatively greater than those on the diagonal of the other confusion matrices. This observation suggests that our proposed model has lower misclassification rates compared to the competing models. However, for the arousal classification, CLISA [62], ACRNN [29], and ATDD LSTM [38] outperformed our proposed model in classifying low arousal with an 8%, 4%, and 10% higher accuracy, respectively and CWGAN [52] outperformed our proposed model in classifying high arousal with a 1% higher accuracy. Nevertheless, our proposed model still outperforms the other models on both the arousal and valence emotion classification.

TABLE 14. **LOSOCV of the proposed model on MAHNOB-HCI database.**

Model	Valence Accuracy (%)		Arousal Accuracy (%)		Average Accuracy for Valence and Arousal (%)	
	Train	Test	Train	Test	Train	Test
CLISA [62]	49.59	47.50	100.0	52.50	74.79	50.00
ACRNN [29]	88.71	55.00	75.56	45.00	82.13	50.00
ECLGCNN [36]	100.0	30.00	100.0	60.00	100.0	45.00
ATDD LSTM [38]	99.79	45.00	96.10	50.00	97.94	47.50
CWGAN [52]	67.25	60.00	67.05	45.00	67.15	52.50
ACapsNet [32]	99.79	50.00	97.84	40.00	98.81	45.00
NAS [22]	99.60	30.00	97.05	45.00	98.32	37.50
DF [21]	92.15	60.00	92.35	55.00	92.25	57.50
ALSTM autoencoder + ACNN [42]	89.80	40.00	93.33	45.00	91.56	42.50
HSLT Transformer [43]	69.41	60.00	72.15	55.00	70.78	57.50
EeT Transformer [44]	90.39	50.00	86.47	55.00	88.43	52.50
Proposed model (LOSOCV)	**71.52**	**60.00**	**67.95**	**55.00**	**69.73**	**57.50**

C. SECOND EVALUATION: LEAVE-ONE-SUBJECT-OUT CROSS-VALIDATION

The evaluation results of the proposed model under the LOSOCV evaluation strategy are presented in Table 13 and Table 14 for the DEAP and MAHNOB-HCI datasets, respectively.

As can be seen from Table 13 and Table 14, the proposed model presents a relatively small discrepancy between the training and testing accuracies on the DEAP and MAHNOB-HCI datasets for both emotional dimensions. This small gap suggests that the proposed model can generalize well to unseen data, thereby avoiding overfitting or underfitting on the training set. However, for the proposed model, test accuracies for valence on the DEAP dataset is as low as 57.50%, and 55% for arousal on the MAHNOB-HCI dataset. Nonetheless, the existing state-of-the-art models present similar performance limitations on the testing set. This could be attributed to limitations in the size of the test data, as we only included data from one subject as test data, which constitutes less than 4% of the entire dataset size for each dataset (40 and 20 experiments for the DEAP and MAHNOB-HCI datasets, respectively).

We calculated the average accuracies for the testing and training sets across the valence and arousal emotional dimensions to compare the overall performance of the proposed model with the existing models. As demonstrated by Table 13 and Table 14, the proposed model achieved the highest average test accuracy of 63.75% and 57.50% for the DEAP and MAHNOB-HCI datasets, respectively. In comparison to existing models, the proposed model demonstrates the smallest discrepancy between the average testing and training accuracies on both datasets. The existing models seem to overfit the MAHNOB-HCI dataset, possibly due to its limited size. Nonetheless, the GAN-based model introduced in [52] exhibits the most consistent training and testing accuracies across both emotional dimensions when benchmarked against the current models on both datasets. Even so, the proposed model maintains a superior performance overall. The models outlined in references [29], [32], [21], [22], [42], [43], and [44] overfit on the DEAP dataset. Although the models discussed in references [62], [36], [38] do not overfit on the DEAP dataset, our proposed model nonetheless achieves superior accuracy.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel CL GAN-based Graph Neural Network for the emotion recognition task, which addresses several issues in affective computing. The CL component is a self-supervised model that learns high-quality EEG representations and addresses inter-subject and intra-subject emotion variabilities. The GAN component adds synthetic realistic-like data to the real data to address the limitation of dataset size, while the GNN component considers the topological structure of EEG channels. Our proposed model was implemented and compared with several state-of-the-art models in a subject-independent experimental setting for EEG-based emotion recognition. The results demonstrated the superior performance of our model in achieving higher recognition accuracy on both arousal and valence dimensions for both DEAP and MAHNOB-HCI databases.

In future work, we can explore the potential benefits of combining multiple modalities to further improve classification accuracy. Additionally, we plan to investigate the effectiveness of our proposed model on a multi-class emotion classification task and evaluate its performance in a multi-modal emotion recognition setting.

REFERENCES

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.

[2] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017.

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Nov. 2009.

[4] W. Hu, G. Huang, L. Li, L. Zhang, Z. Zhang, and Z. Liang, "Video-triggered EEG-emotion public databases and current methods: A survey," *Brain Sci. Adv.*, vol. 6, no. 3, pp. 255–287, Sep. 2020.

[5] S.-E. Moon and J.-S. Lee, "Implicit analysis of perceptual multimedia experience based on physiological response: A review," *IEEE Trans. Multimedia*, vol. 19, no. 2, pp. 340–353, Feb. 2017.

[6] X. Hu, J. Chen, F. Wang, and D. Zhang, "Ten challenges for EEG-based affective computing," *Brain Sci. Adv.*, vol. 5, no. 1, pp. 1–20, Mar. 2019.

[7] M. Sreeshakthy and J. Preethi, "Classification of human emotion from DEAP EEG signal using hybrid improved neural networks with cuckoo search," *Broad Res. Artif. Intell. Neurosci.*, vol. 6, pp. 60–73, Jan. 2016.

[8] Y. Li, L. Wang, W. Zheng, Y. Zong, L. Qi, Z. Cui, T. Zhang, and T. Song, "A novel bi-hemispheric discrepancy model for EEG emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 2, pp. 354–367, Jun. 2021.

[9] Y.-J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-time movie-induced discrete emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 550–562, Oct. 2018.

[10] J. Li, S. Qiu, Y.-Y. Shen, C.-L. Liu, and H. He, "Multisource transfer learning for cross-subject EEG emotion recognition," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3281–3293, Jul. 2020.

[11] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, "Domain adaptation techniques for EEG-based emotion recognition: A comparative study on two public datasets," *IEEE Trans. Cognit. Develop. Syst.*, vol. 11, no. 1, pp. 85–94, Mar. 2019.

[12] *The McGill Physiology Virtual Lab*. Accessed: Feb. 9, 2023. [Online]. Available: https://www.medicine.mcgill.ca/physio/vlab/biomed_signals/eeg_n.htm

[13] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Jul. 2014.

[14] R. Nawaz, K. H. Cheah, H. Nisar, and V. V. Yap, "Comparison of different feature extraction methods for EEG-based emotion recognition," *Biocybernetics Biomed. Eng.*, vol. 40, no. 3, pp. 910–926, Jul. 2020.

[15] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu, "Differential entropy feature for EEG-based vigilance estimation," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 6627–6630.

[16] F. Cui, R. Wang, W. Ding, Y. Chen, and L. Huang, "A novel DE-CNN-BiLSTM multi-fusion model for EEG emotion recognition," *Mathematics*, vol. 10, no. 4, p. 582, Feb. 2022.

[17] Y. Yang, Q. Wu, Y. Fu, and X. Chen, "Continuous convolutional neural network with 3D input for EEG-based emotion recognition," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2018, pp. 433–443.

[18] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auto. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[19] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Nov. 2013, pp. 81–84.

[20] D.-W. Chen, R. Miao, W.-Q. Yang, Y. Liang, H.-H. Chen, L. Huang, C.-J. Deng, and N. Han, "A feature extraction method based on differential entropy and linear discriminant analysis for emotion recognition," *Sensors*, vol. 19, no. 7, p. 1631, Apr. 2019.

[21] J. Cheng, M. Chen, C. Li, Y. Liu, R. Song, A. Liu, and X. Chen, "Emotion recognition from multi-channel EEG via deep forest," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 2, pp. 453–464, Feb. 2021.

[22] C. Li, Z. Zhang, R. Song, J. Cheng, Y. Liu, and X. Chen, "EEG-based emotion recognition via neural architecture search," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 957–968, Apr. 2023.

[23] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 839–847, Mar. 2019.

[24] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, May 2013, pp. 6645–6649.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[26] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[27] J. Yang, X. Huang, H. Wu, and X. Yang, "EEG-based emotion classification based on bidirectional long short-term memory network," *Proc. Comput. Sci.*, vol. 174, pp. 491–504, Jan. 2020.

[28] H. Cui, A. Liu, X. Zhang, X. Chen, K. Wang, and X. Chen, "EEG-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network," *Knowl.-Based Syst.*, vol. 205, Oct. 2020, Art. no. 106243.

[29] W. Tao, C. Li, R. Song, J. Cheng, Y. Liu, F. Wan, and X. Chen, "EEG-based emotion recognition via channel-wise attention and self attention," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 382–393, Jan. 2023.

[30] D. Maheshwari, S. K. Ghosh, R. K. Tripathy, M. Sharma, and U. R. Acharya, "Automated accurate emotion recognition system using rhythm-specific deep convolutional neural network technique with multi-channel EEG signals," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104428.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[32] C. Li, B. Wang, S. Zhang, Y. Liu, R. Song, J. Cheng, and X. Chen, "Emotion recognition from EEG based on multi-task learning with capsule network and attention mechanism," *Comput. Biol. Med.*, vol. 143, Apr. 2022, Art. no. 105303.

[33] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1290–1301, Jul. 2022.

[34] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Jul./Sep. 2020.

[35] G. Zhang, M. Yu, Y.-J. Liu, G. Zhao, D. Zhang, and W. Zheng, "SparseDGCNN: Recognizing emotion from multichannel EEG signals," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 537–548, Jan. 2023.

[36] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Appl. Soft Comput.*, vol. 100, Mar. 2021, Art. no. 106954.

[37] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition," *Inf. Process. Manage.*, vol. 57, no. 3, May 2020, Art. no. 102185.

[38] X. Du, C. Ma, G. Zhang, J. Li, Y.-K. Lai, G. Zhao, X. Deng, Y.-J. Liu, and H. Wang, "An efficient LSTM network for emotion recognition from multichannel EEG signals," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1528–1540, Jul. 2022.

[39] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen, "Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–7.

[40] G. Zhang and A. Etemad, "Deep recurrent semi-supervised EEG representation learning for emotion recognition," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2021, pp. 1–8.

[41] J. Liu, G. Wu, Y. Luo, S. Qiu, S. Yang, W. Li, and Y. Bi, "EEG-based emotion classification using a deep neural network and sparse autoencoder," *Frontiers Syst. Neurosci.*, vol. 14, p. 43, Sep. 2020, doi: 10.3389/fnsys.2020.00043.

[42] Arjun, A. S. Rajpoot, and M. R. Panicker, "Subject independent emotion recognition using EEG signals employing attention driven neural networks," *Biomed. Signal Process. Control*, vol. 75, May 2022, Art. no. 103547.

[43] Z. Wang, Y. Wang, C. Hu, Z. Yin, and Y. Song, "Transformers for EEG-based emotion recognition: A hierarchical spatial information learning model," *IEEE Sensors J.*, vol. 22, no. 5, pp. 4359–4368, Mar. 2022.

[44] J. Liu, H. Wu, L. Zhang, and Y. Zhao, "Spatial–temporal transformers for EEG emotion recognition," in *Proc. 6th Int. Conf. Adv. Artif. Intell.*, Oct. 2022, pp. 116–120.

[45] F. Wang, S.-H. Zhong, J. Peng, J. Jiang, and Y. Liu, "Data augmentation for EEG-based emotion recognition with deep convolutional neural networks," in *Proc. 24th Int. Conf. MultiMedia Model. (MMM)*, Bangkok, Thailand, vol. 7, Feb. 2018, pp. 82–93.

[46] E. Lashgari, D. Liang, and U. Maoz, "Data augmentation for deep-learning-based electroencephalography," *J. Neurosci. Methods*, vol. 346, Dec. 2020, Art. no. 108885.

[47] J. Fan, C. Sun, C. Chen, X. Jiang, X. Liu, X. Zhao, L. Meng, C. Dai, and W. Chen, "EEG data augmentation: Towards class imbalance problem in sleep staging tasks," *J. Neural Eng.*, vol. 17, no. 5, Oct. 2020, Art. no. 056017.

[48] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[49] K. G. Hartmann, R. T. Schirrmeister, and T. Ball, "EEG-GAN: Generative adversarial networks for electroencephalographic (EEG) brain signals," 2018, *arXiv:1806.0187*.

[50] Y. Luo, L.-Z. Zhu, Z.-Y. Wan, and B.-L. Lu, "Data augmentation for enhancing EEG-based emotion recognition with deep generative models," *J. Neural Eng.*, vol. 17, no. 5, Oct. 2020, Art. no. 056021.

[51] Y. Luo, L. Z. Zhu, Z. Y. Wan, and B. L. Lu, "A GAN-based data augmentation method for multi-modal emotion recognition," in *Proc. Int. Symp. Neural Netw.* Cham, Switzerland: Springer, 2019, pp. 141–150.

[52] Y. Luo and B.-L. Lu, "EEG data augmentation for emotion recognition using a conditional Wasserstein GAN," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 2535–2538.

[53] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," 2017, *arXiv:1703.10717*.

[54] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[55] S. Becker and G. E. Hinton, "Self-organizing neural network that discovers surfaces in random-dot stereograms," *Nature*, vol. 355, no. 6356, pp. 161–163, Jan. 1992.

[56] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193907–193934, 2020.

[57] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 3875–3879.

[58] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.

[59] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.

[60] M. N. Mohsenvand, M. R. Izadi, and P. Maes, "Contrastive representation learning for electroencephalogram classification," in *Proc. Mach. Learn. Health*, 2020, pp. 238–253.

[61] K. Pinitas, K. Makantasis, A. Liapis, and G. N. Yannakakis, "Supervised contrastive learning for affect modelling," in *Proc. Int. Conf. Multimodal Interact.*, Nov. 2022, pp. 531–539.

[62] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, "Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2496–2511, Jul./Sep. 2023.

[63] N. S. Suhaimi, J. Mountstephens, and J. Teo, "EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–19, Sep. 2020.

[64] J. Liu, X. Shen, S. Song, and D. Zhang, "Domain adaptation for cross-subject emotion recognition by subject clustering," in *Proc. 10th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, May 2021, pp. 904–908.

[65] J. Teo, C. L. Hou, and J. Mountstephens, "Deep learning for EEG-based preference classification," *AIP Conf. Proc.*, vol. 1891, no. 1, 2017, Art. no. 020141.

[66] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[67] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[68] P. Li, J. Wang, Y. Qiao, H. Chen, Y. Yu, X. Yao, P. Gao, G. Xie, and S. Song, "An effective self-supervised framework for learning expressive molecular global representations to drug discovery," *Briefings Bioinf.*, vol. 22, no. 6, Nov. 2021, Art. no. bbab109.

[69] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL)*, 2019, pp. 4171–4186.

[70] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[71] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 98–107, Jan. 2018.

[72] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cognition Emotion*, vol. 23, no. 2, pp. 209–237, Feb. 2009.

[73] J. A. Russell, "Culture and the categorization of emotions," *Psychol. Bull.*, vol. 110, no. 3, pp. 426–450, 1991.

[74] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Personality*, vol. 11, no. 3, pp. 273–294, Sep. 1977.

[75] D. Garg and G. K. Verma, "Emotion recognition in valence-arousal space from multi-channel EEG data and wavelet based deep learning framework," *Proc. Comput. Sci.*, vol. 171, pp. 857–867, Jan. 2020.

[76] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced Lectures on Machine Learning*. Springer, 2004, pp. 63–71.

[77] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, Dec. 2014, pp. 1–15.

SAREH SOLEIMANI GILAKJANI (Student Member, IEEE) received the B.Sc. degree from Shahid Beheshti University, Tehran, Iran, and the M.Sc. degree in electrical and computer engineering from the University of Ottawa, Canada, where she is currently pursuing the Ph.D. degree in electrical and computer engineering. Her research interests include affective computing, deep learning, and its applications.

HUSSEIN AL OSMAN (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the University of Ottawa, Canada. He is currently an Associate Professor with the University of Ottawa and leads the Multimedia Processing and Interaction Group. His research interests include affective computing, specifically multi-modal affect estimation, human–computer interaction, serious gaming, and multimedia systems.