IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 May 2024, accepted 24 July 2024, date of publication 1 August 2024, date of current version 12 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3436556

RESEARCH ARTICLE

Deep Representation Learning for Multimodal Emotion Recognition Using Physiological Signals

MUHAMMAD ZUBAIR, SUNGPIL WOO, SUNHWAN LIM, AND CHANGWOO YOON
Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, South Korea

Corresponding author: Sungpil Woo (woosungpil@etri.re.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) under Grant RS-2024-00423362 and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) under Grant 2022-0-01032.

ABSTRACT Physiological signal analysis has gained a lot of interest in recent years and has been used in a variety of fields including emotion recognition, activity recognition, and health monitoring. However, emotion recognition based on physiological signals is not yet explored entirely using deep learning, and there are still some exciting challenges to be handled. For example, deep representation learning for spatio-temporal feature extraction, the discrimination between adjacent emotions with entangled features, and the imbalanced distribution of data are the most prominent issues in emotion recognition. This work focuses on deep multimodal representation learning of physiological signals to alleviate the aforementioned challenges. We introduce a novel deep learning architecture for emotion classification that effectively extracts spatio-temporal information from physiological signals. We proposed a mutual attention mechanism to extract emotion-specific features for improved classification. To handle the issue of adjacent emotions and imbalanced data, we introduce a dense max-margin loss function based on Gaussian similarity measure. Our experiments on different datasets reveal that the proposed emotion classification methodology effectively learns a balanced deep representation of physiological signals, significantly maximizes the inter-class margin, and reduces intra-class variance to discriminate between different classes of emotions.

INDEX TERMS Emotion recognition, deep learning, attention mechanism, imbalanced data, EEG, ECG.

I. INTRODUCTION

Intelligent systems with the adequacy of emotion estimation have a great potential to revolutionize applications in healthcare, education, marketing, entertainment, surveillance, and security. Similarly, emotions have a substantial impact on human life. The recent advancements in computing technologies and the miniaturization of physiological sensors have made it possible to acquire various physiological signals constantly during day-to-day activities. These signals including electrocardiogram (ECG), electroencephalogram (EEG), and galvanic skin response (GSR) can efficiently capture the emotion-related information that originates from autonomic nervous system activity triggered by external or internal stimuli [1]. In contrast, physical signals like facial expressions [2], speech [3], [4] and gestures [5], [6] are comparatively simple to acquire and have been extensively explored for emotion recognition [7]. However, physiological signals are more trustworthy in identifying true emotions than facial or vocal expressions since they are involuntary and cannot be controlled intentionally by an individual.

Although researchers made many efforts to recognize emotions using different channels of expression, however, physiological signals have been overlooked in emotion recognition [8]. Literature reveals significant limitations of the emotion recognition system that should be addressed. For example, the performance of the emotion classification model presented in most studies depends on a specific

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

set of hand-crafted features and thus lacks generalizability. Physiological signals demonstrate diverse morphological features that vary over time. These temporal variations of physiological signals are unique to each individual and highly correlated to the mental state and nature of the individuals [9], [10]. Similarly, the physiological signals exhibit individual-specific temporal variations triggered by the autonomic nervous system activation that results in entangled features of adjacent emotions [1]. Therefore, the excessive dependence of the classification model on conventional feature extraction methods deteriorates its generalizability. Moreover, for effective discrimination of arousal and valence levels, the extraction of emotion-specific features from physiological signals is still challenging.

Another crucial problem is the classification of adjacent emotions. Emotions elicited in the laboratory using various types of stimuli are still far away from those experienced by humans in everyday life. The emotion elicited under controlled conditions in a lab is subject to the nature and personality of each individual [11], [12]. The acquired physiological data of adjacent emotions results in an overlap feature distribution with substantial iter-class and intra-class variation and therefore deteriorates the performance of the emotion classifier. In addition, imbalanced data poses a significant challenge in deep learning, where class instances exhibit a skewed distribution [13]. In such cases, some classes, known as minority classes, are represented sparsely, while others have abundant representation and are designated as majority classes. Training a deep model with an imbalanced dataset leads the model to exhibit bias towards the majority class and thus significantly deteriorates the model performance to classify less frequent events from the minority class. However, rare events are of great importance in some real-world scenarios like surveillance, anomaly detection, and disease diagnosis. Therefore, it is imperative to design intelligent emotion recognition systems that can differentiate between adjacent emotions and address the undesirable bias introduced by imbalanced data distribution.

To alleviate the above problems, we introduce a deep multimodal emotion classification system based on physiological signals. We present a deep multimodal system for characterizing physiological signals using convolutional neural networks (CNN) and Long short-term memory networks (LSTM). The proposed model also encompasses a novel mutual attention module to learn an emotion-specific representation of multiple physiological signals. Moreover, the proposed method overcomes the challenge posed by the imbalanced distribution of class samples and improves the classification of neighboring (adjacent) emotions by enforcing a sufficient margin between class boundaries. The three key contributions are as follows.

- We introduce a deep multimodal representation learning architecture to recognize emotions using physiological signals. The proposed architecture captures crucial variations of the signals to extract spatio-temporal features.
- We design a mutual attention mechanism that effectively selects the most relevant and emotion-specific content for the target task. The proposed attention module adaptively estimates the mutually important channels and features to improve the classification performance.
- We propose a novel loss function using a feature-based Gaussian similarity measure coupled with hard sample mining strategy. This loss function is intended to improve the classification of hard samples and overcome the model's bias towards the highly represented class due to the uneven distribution of class instances.

The reminder of the paper is organized as follows. We present a brief review of the related literature on emotion recognition methods in section II. The proposed multi-modal emotion recognition methodology including dense max-margin loss function and attention mechanism is presented in section III. Section IV includes a description of datasets and signals preprocessing. In section V, the experimental findings of this study are reported. The concluding remarks of this study are given in section VI.

II. RELATED WORK

A. EMOTION MODELING

Emotion modeling is required for the description, representation, and quantitative analysis of emotion. From the perspective of emotion quantization, psychologists mostly describe emotion using two basic models of emotions; discrete emotion model and affective dimension model. The discrete emotion model includes a set of basic and instinctive emotions. The preliminary work on six basic discrete emotions was undertaken by a famous psychologist Ekman et al. [14], [15]. In addition, numerous studies have been published on differentiating different numbers of discrete emotions using physiological signals [16], [17]. In affective dimensional models, emotions are represented using multi-dimensional emotion space. The Circumplex model of affect is the most frequently used dimensional model in affective computing [18]. This model illustrates human emotions in terms of two parameters called arousal and valence. These two parameters form a 2D space for the categorization of multiple emotions as depicted in Fig.1. Several investigative studies are also conducted on valence-arousal plans for emotion categorization [19].

B. PREPROCESSING AND FEATURES EXTRACTION

Physiological signal processing for emotion recognition includes preprocessing and emotion-related information extraction. The quality of physiological signals significantly deteriorates during acquisition due to various noises such as motion artifacts, poor contact of wearable sensors with skin, and electrical interference [20]. To eradicate the erroneous information induced by these noises, bandpass filters [21], Butterworth filters [22], and moving average filters [23], [24] have been used in literature. R-peak is a key component in ECG and is used in the estimation of heart rate variability

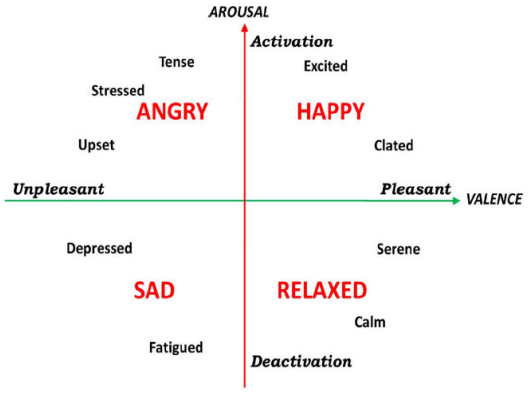

FIGURE 1. Dimensional emotion model [10].

series. Pan and Tompkin's algorithm is widely adopted to detect R-peak [21]. Similarly, for EEG, various methods have also been presented in the literature for the elimination of motion artifacts. These methods include signal-to-noise ratio, adaptive filtering, and least-mean-square(LMS) Algorithm [25]. The removal of eye artifacts is widely performed via undefined source separation technique [12]. In addition, averaging EEG channels to a common reference, downsampling, and filtering are frequently used in EEG preprocessing [24].

For ECG-based classification of emotions, various features in the time domain (RR, SDNN, RMSSD) and frequency domain (HF, LF, VLF) are extracted to discriminate between different emotions [24]. Additionally, non-linear features (SD1, SD2) have also demonstrated their efficacy in ECG-based emotion recognition models [26]. Similarly, in EEG-based emotion classification, five basic frequency bands known as delta, theta, alpha, beta, and gamma are widely used to extract features [27]. In addition, features like Differential entropy (DE), power spectral density (PSD), and differential causality (DCAU) have been broadly investigated for EEG-based emotions classification [27].

C. EMOTION RECOGNITION

Deep learning has overcome the limitations associated with traditional machine learning methods and received much attention due to its remarkable Physiological signals based emotion classification methods can be categorized broadly into two main groups; conventional methods and deep learning methods. Emotion recognition studies based on conventional methods exploit traditional methods of feature extraction and classification. For instance, traditional machine learning methods such as linear discriminant analysis [26], [28], [29], random forest [30], and support vector machines [16], [31], [32], [33], [34] are used to classify different types of emotions. However, these methods are very challenging and require great expertise to unveil the embedded emotion-related information in physiological signals due to subject specificity [27]. In addition, for traditional machine learning methods to be effective, the selection of appropriate features should be performed wisely as performance deterioration is mainly caused by irrelevant and redundant features [27].

The second group of emotion recognition studies uses deep learning methods to alleviate the problems of conventional machine learning algorithms. In computer vision and natural language processing, deep learning has overcome the limitations associated with traditional machine learning methods and received much attention due to its remarkable performance. For instance, hand-crafted feature extraction and selection is considered to be the most challenging step in pattern recognition and classification [27], [35]. However, deep learning has not only negated the issue of hand-crafted feature extraction but also improved the generalization capability of classification models [35]. In affective computing, numerous multi-modal deep learning techniques have been investigated by researchers for physiological signals-based emotion recognition. for instance, EEG signals in addition with GSR signals are used for emotion recognition using convolutional neural networks (CNN) and recurrent neural networks (RNN) [36], [37]. Similarly, ECG signals are also used in literature in combination with GSR and EEG signals to recognise human emotions using deep learning models [38], [39], [40]. Lin et al. [41] also introduced a CNN-based emotion recognition model that exploits EEG, ECG, GSR, electrooculography (EOG) and skin temperature. In this, the feature extracted from different modalities are concatenated using fully connected layer. A similar study is also carried out by Santamaria et al. [42] to classify two levels of arousal and valence. Moreover, Grapgh Neural Networks (GNN) are also used in literature to quantify the interrelationship between different physiological signals for efficient prediction of emotion state [43], [44], [45].

D. ATTENTION IN MULTIMODAL EMOTION RECOGNITION

Choosing the most appropriate and relevant channels and features is a crucial concern in EEG-based emotion identification [27]. Many channel selection and feature selection strategies have been suggested in the literature on this matter [35], [46]. Nevertheless, these traditional approaches depend on predetermined standards and hence are not applicable to real-life scenarios.To overcome this issue, many researchers employed attention mechanisms in deep models for efficient classification of emotions. For instance, bidirectional LSTM-RNNs embedded with attention mechanism are used for multi-modal emotion recognition [47]. This attention module is designed to learn most relevant temporal features. Similarly, pre-trained transformers are also employed to classify multi-modal emotion [48]. However, unlike, [47] and [48], the proposed mutual attention mechanism in this study selects the most relevant and mutually important channels and features to improve the classification performance of deep multi-modal architecture.

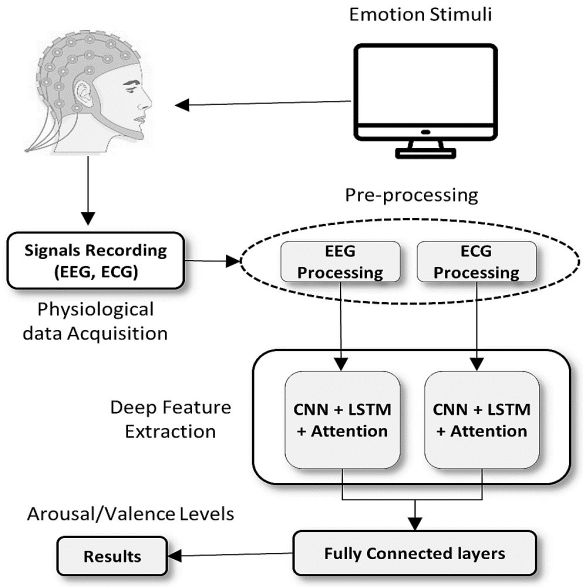

FIGURE 2. **Overview of proposed multi-modal emotion recognition system.**

III. EMOTION RECOGNITION METHODOLOGY

An overview of the proposed deep multimodel emotion classification system is given in Fig. 2. This system aims to classify complex emotions using physiological signals by learning a balanced and emotion-specific representation from an imbalanced physiological dataset. We used ECG and EEG data to classify arousal and valence levels (high/low). The proposed deep model is responsible for collecting spatio-temporal information by exploiting short and long-term variations of EEG and ECG signals triggered by emotional stimuli. We introduce an attention mechanism for the selection of emotion-specific information and assure the extraction of the most relevant and significant information associated with the target emotion. In [49] attention-based auto-encoders are presented that only focus on channels for attention estimation. However, in this work, we introduce a mutual attention mechanism that focuses on mutually important content for the target emotion. In addition, we also proposed a dense max-margin loss function that computes affective scores at the feature level and ensures the maximum margin between different classes and minimum variance inside class samples.

A. DEEP MODEL ARCHITECTURE

We designed a two-branch architecture for multi-modal emotion classification. This architecture takes EEG and ECG signals as input to recognize different levels of arousal and valence for human emotion identification as depicted in Fig.3. The first branch makes use of the ECG feature extractor to extract deep features of ECG signals. The proposed ECG feature extractor consists of CNN layers followed by LSTM layer. The attention module in the ECG model is placed after the CNN layers to emphasize the intrinsic deep features extracted by CNN layers. In CNN layers, we used the batch normalization and dropout layer to overcome the overfitting issue. For activation, we use the parametric ReLu function instead of ReLu to avoid the dying ReLu issue. To mitigate the issue of information loss, we also use skip connections in CNN layers. This architecture aims to learn an efficient representation learning by exploiting short-term and long-term variations of HRV series and thus improves classification performance with target-specific feature extraction.

Similarly, the EEG representation learning model is used in the second branch of the multi-modal framework. The proposed EEG feature extractor is composed of two CNN layers and an LSTM layer. The CNN layer is used in combination with the batch normalization and dropout layer. The parametric ReLu layers are employed for activation to mitigate the dying ReLu problem. The attention module is placed at the start of the model to select the important and most relevant channels and features for improved classification. The same attention module is integrated in both branches to refine the extracted features by highlighting the target-specific content. The extracted features from both modalities are concatenated to perform an intermediate fusion of the target-specific information. After concatenation, transformation is performed via two dense neural layers to model the extracted features for target emotion identification.

The training of all the models illustrated above is performed using the proposed supervision strategy (dense max-margin loss function). A series of comprehensive tests are conducted to evaluate the effectiveness of the proposed emotion recognition methodology. First, separate models for EEG and ECG are trained and evaluated. Afterward, the training and evaluation of multi-modal architecture is performed.

B. ATTENTION LAYER

Attention mechanisms in deep learning have gained the interest of many researchers recent years. The principle of the attention mechanism is to emphasize task-relevant information to assist the model in efficient extraction. It scales the input based on its importance. Thus the most relevant information propagates through the model. Unlike [49], the proposed mutual attention mechanism uses channel-wise as well as feature-wise attention by estimating the mutual statics of the channels and features. The key purpose of the attention module is to determine the most suitable and emotion-specific features while taking the importance of the channels into account. Although the channel mean has been used in the attention layer before; however, estimating a mutual attention mask by exploiting the importance of channels and features of each channel has not been explored yet in affective computing. The proposed attention layer is shown in Fig.4. First, an attention mask is generated and then features are scaled based on their importance for the target emotion. The formulation of the proposed attention layer

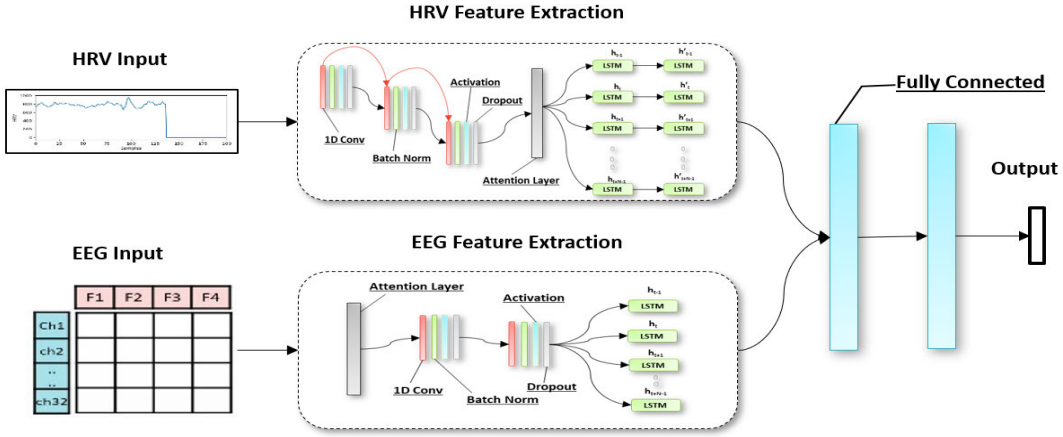

FIGURE 3. Proposed multimodal emotion recognition network.

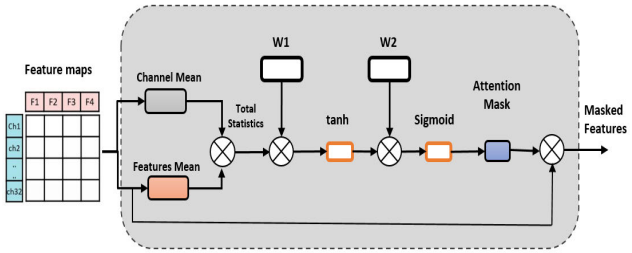

FIGURE 4. Architecture of proposed attention layer.

is given below.

$$Channel_{mean} = \frac{1}{c}\sum_{1}^{c} f_{c,s} \tag{1}$$

$$Feature_{mean} = \frac{1}{s}\sum_{1}^{s} f_{c,s} \tag{2}$$

$$Total_{stat} = Channel_{mean} \times Feature_{mean} \tag{3}$$

$$V_s = \tanh(Total_{stat} \cdot W1 + b1) \tag{4}$$

$$A_{c,s} = sigmoid(Vs \cdot W2 + b2) \tag{5}$$

$$F_{masked} = f_{c,s} \cdot A_{c,s} \tag{6}$$

Eq. (1) shows the channel-wise mean of the input matrix, while Eq. (2) shows the feature-wise mean. Both channel-wise and feature-wise operations incorporate intrinsic information along the respective dimension. The attention matrix (2D) is computed using Eq. (5) which assigns a different score to each entity based on the mutual importance of channels and their features. The attention matrix is then multiplied with the input feature map to mask the input.

C. PROPOSED DENSE MAX-MARGIN LOSS FUNCTION

We designed a novel loss functions to accomplish two objectives. 1) Enforcing compactness inside class clusters to reduce the variance of class samples. 2) To induce margin between different classes to ensure the correct classification of minority class samples. The best way to achieve these objectives is to measure class similarities using feature space representation instead of class prediction. This method provides the flexibility to manipulate feature space directly in order to induce margin and compactness at the feature level. Therefore, instead of using Euclidean similarity at the class level, we used Gaussian similarity measure [50] at the instance level that can be computed as follows.

$$d(f_i, w_j) = exp(-\frac{\| f_i - w_j \|^2}{\sigma}) \tag{7}$$

where, σ is a weighting parameter that normalizes the distance between features and their predictive class projections. Based on the Gaussian similarity measure, we employed hard sample mining strategy to improve the classification of hard samples and to address the issue of overlapped feature distributions. First, we defined hard samples based on the Gaussian similarity measure. A hard positive sample can be defined as a sample x_i of class c that is classified as a class c sample with a minimum Gaussian similarity measure. A hard negative sample can be defined as a sample x_i, not from class c but classified as a class c sample with a high Gaussian similarity measure. We define hard positive and hard negative samples mathematically as follows:

$$P_i^{ins} = \{x_i | a_i = c,\ low\ similarity\ d(f_i, wj)\} \tag{8}$$

$$N_i^{ins} = \{x_i | a_i \neq c,\ High\ similarity\ d(f_i, wj)\} \tag{9}$$

In order to maximize the margin between different class samples, we employed the Gaussian similarity measure [50] along with a hard sample mining technique to alleviate the degree of entanglement in the features associated with different classes. Enforcing sufficient margins between different class features can significantly alleviate the misclassification of the neighbor emotions and also assist the model in classifying minority class samples correctly. Therefore, we proposed a max-margin loss function, which

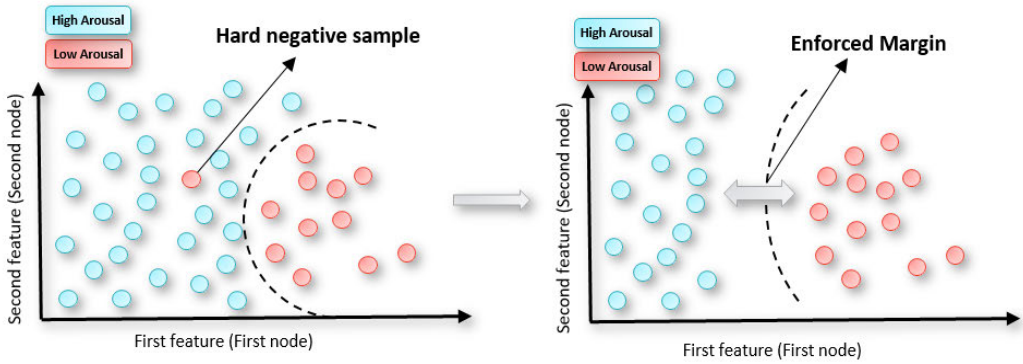

FIGURE 5. 2D feature visualization illustrating hard negative samples. The visualization is drawn considering two nodes at the final fully connected layer. The left figure shows the hard negative sample while the right figure shows the impact of margin for accurate classification of hard negative samples.

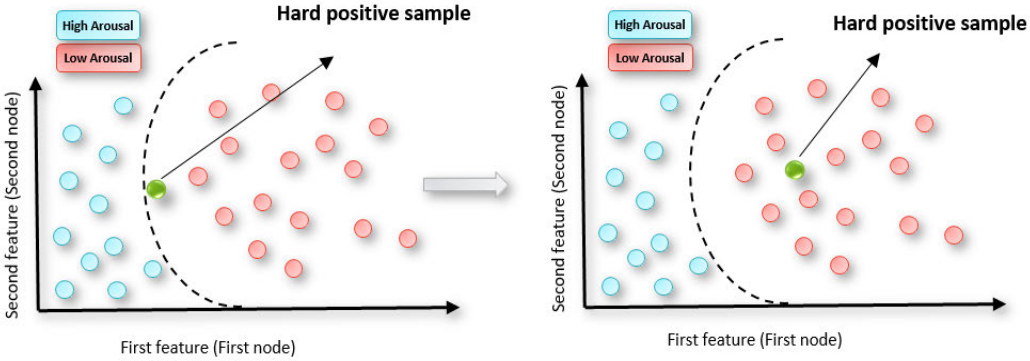

FIGURE 6. 2D feature visualization illustrating hard positive sample. The visualization is drawn considering two nodes at final fully connected layer. The left figure shows the hard positive sample while the right figure shows the impact of variance reduction for more confident classification of hard positive samples.

is formulated as

$$L^{mm} = \| \beta - d(f_i, w_f) - max\{d(f_i, w_k)\} \|^2 \quad (10)$$

where β is the hyper-parameter that enforces the margin between the two classes. In the above loss function, the first term $d(f_i, w_j)$ represents the closeness of the samples with target classes in the feature space. The second term $max\{d(f_i, w_k)\}$ represents the hard negative sample of the batch. The proposed loss function enforces the margin between class features by taking hard negative samples into account, as depicted in Fig.5.

A similar approach based on hard sample mining is also adopted to reduce the features' sparsity in intra-class samples. This part of the loss function improves the compactness of the deep features and thus facilitates the discrimination between features of different neighboring classes. We formulated the minimum variance loss function as follows.

$$L^{mv} = \| d(f_i, w_f) - min\{d(f_i, w_j)\} \|^2 \quad (11)$$

In the above loss function, the first term in the above equation $d(f_i, w_j)$ represents the anchor sample representing similarity with the target class while the second term $min\{d(f_i, w_j)\}$ represents the hard positive sample. The difference in the above equation represents the compactness of class features as illustrated in Fig.6. The depletion of the gap illustrated by Eq.11 induces compactness inside class features that results in better classification. As both parts of the loss function play their role in the classification of emotion-specific data, therefore, we took the weighted sum of the minimum variance and maximum margin losses. The cumulative dense max-margin loss function is formulated as follows.

$$L^{mvmm} = 0.5 \times (L^{mv} + L^{mm}) \quad (12)$$

The above-mentioned weighted loss function takes both hard positive and hard negative samples into account and thus assures the maximization of the margin between different classes and the minimization of the variance inside the class. To boost the classification performance, we use the aforementioned dense max-margin loss function with conventional cross-entropy loss function.

IV. EXPERIMENTAL SETUP

The proposed technique for emotion recognition is evaluated using a freely accessible AMIGOS [12] datasets. The Amigos dataset contains ECG and 14 channels EEG recordings of 40 participants [12]. These recordings are collected for

16 movie clips covering four quadrants of the circumplex model of affect. Each video clip is rated for arousal and valence on a scale from 0 to 9. Besides these publicly available datasets, we also acquired a new dataset under Young Scientist Research Program (YSRP) to evaluate the proposed methodology. Literature on affective computing reveals that very few datasets provide quality ECG signals for the analysis of emotion. Therefore, to mitigate this issue, we used YSRP data for validating the proposed models. The complete experimental setup, physiological signal acquisition, and pre-processing steps carried out in this study are explained in this chapter.

A. DATA ACQUISITION

The development of emotion recognition systems requires emotion-related physiological data. However, the acquisition of real-life emotion-related data is almost impossible. Therefore, data related to different emotions are acquired in the laboratory using different types of stimuli. These stimuli include pictures, audio clips, and video clips. However, it has been demonstrated that emotion elicitation using videos is comparatively better than audio and pictures. Therefore, for emotion elicitation, we collected 36 video clips that have been used in the DECAF [51] database. Out of the 36 movies available, We chose 20 movie scenes. Each of the four classes(high arousal, low arousal, high valence, low valence) is represented by five videos to cover all the quadrants of the two-dimensional circumplex model [18]. We presented these video clips to each subject using PsychoPy. We also designed an experiment for the collection of video clip ratings using the same PsychoPy program. The rating of each video clip for arousal and valence is recorded on a scale from 0 to 9.

We recorded Electroencephalogram (EEG) and Electrocardiogram (ECG) signals that have been used significantly for emotion recognition [9], [10], [26]. These signals are recorded from 25 students (25-34 years). For signal acquisition, we used the Mind Media Nexus-10 device. As EEG signals are considered to be the most suitable indicators for emotion recognition [24], therefore, we used two disc electrodes on the frontal lobe and significantly recorded two-channel (right frontal lobe and left frontal lobe) EEG data. We recorded ECG by adopting a wrist placement strategy.

B. PHYSIOLOGICAL SIGNAL PROCESSING

ECG Pre-processing: The quality of ECG signal significantly deteriorates because of motion artifacts and power line interference during recording. These noises suppress the emotion-specific information and result in poor classification performance. To eliminate these undesirable noises from ECG signals, we adopted a filter-based method [24]. In this method, first, a median filter with a sliding window of 200ms is applied. Following that, a second median filter with a 600ms window is used. The first filter eliminates the P peak and QRS segment while the second filter removes the T peak from the ECG waveform and produces a baseline signal. This baseline is subtracted from the original signal to remove baseline drift. Additionally, a low pass filter is utilized, with a 35Hz cut-off frequency, to alleviate high-frequency disruption caused by power lines.

EEG Pre-processing: EEG signals are heavily distorted by motion artifacts, electrooculogram artifacts, and power supply noise during acquisition. These noises also suppress the emotion-related content of the signals. Therefore pre-processing is performed before feature extraction. First, the eye blink distortions are eliminated through the undefined source separation approach. Afterward, the high-frequency noises are removed by utilizing a bandpass filter (1-45Hz) that filters out all the unwanted noises. Each EEG channel is further decomposed into five frequency bands, namely delta (0-4HZ), theta (4-8Hz), alpha (8-12Hz), beta (12-30Hz), and gamma (30+Hz) [35], [52]. These frequency bands carry vital information about emotions triggered by external stimuli.

C. INPUT DATA STRUCTURE

Differential entropy has shown significant performance in emotion recognition using EEG in recent literatures [49] and [53]. Therefore, for each EEG band, we computed the differential entropy. First, the EEG signal from all channels is decomposed into different frequency bands, namely theta, alpha, beta, and gamma [35], [52]. Afterward, for each of these bands in all channels, we computed differential entropy as follows:

$$f(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{(x-\mu)^2}{2\pi\sigma^2} \log \frac{1}{\sqrt{2\pi\sigma^2}} \tag{13}$$

$$exp \frac{(x-\mu)^2}{2\pi\sigma^2} = \frac{1}{2} \log 2\pi e\sigma^2 \tag{14}$$

For ECG signals, after the detection of R peak, we compute heart rate variability series. Prior to the heart rate variability series the ECG signal are normalized. To get equal-length segments of HRV series, we use the zero embedding technique. We append zero at the end of each sample in such a way that the length of all segments remains the same.

After preprocessing, all three datasets exhibit an imbalanced distribution of samples across different classes. The separate arousal and valence models are trained to classify two classes (High/low). Furthermore, we adopted a 10-fold cross-validation approach in all tests to investigate the efficacy of the method proposed in this study. First, the dataset is split into 10 folds. The training is carried out with nine folds, while the held-out fold is used for testing. A total of ten trials are performed, and every time a new fold is used as a test set. The final results are acquired by averaging the results of ten trials.

V. RESULTS AND DISCUSSION

This section presents the experimental results of our proposed emotion classification methodology. First, we trained and

evaluated the unimodal architectures of ECG and EEG separately. These experiments are performed to assess the effectiveness of architecture and supervision strategies for a single model based on ECG or EEG data. Second, we trained a multi-modal network using both ECG and EEG data. The experimental findings illustrate that the proposed model can effectively extract deep temporal and spatial patterns that are relevant to the target emotion.

A. ECG BASE CLASSIFICATION

To evaluate the ECG-based emotion classification model, we used ECG recordings from AMIGOS [12] dataset. We empirically selected a segment size of 100 samples of the HRV series as the input array. The 1D input array with 100 samples of HRV demonstrates optimal classification performance for ECG-based emotion classification. The classification is performed for two classes (high/low) of arousal and valence. We use L2 regularization along with dropout to address the overfitting in a better way. We maintained a weightage of 0.5 for the dense max-margin loss function. After extensive experimentation with various batch sizes, we choose an optimal batch size of 32 samples.

First, we train a baseline model to investigate the effectiveness of the mutual attention module and dense max-margin loss function. The baseline model has the same architecture illustrated in Section III-A but without a mutual attention module. Similarly, for baseline model training, only conventional cross-entropy loss function is adopted. The baseline model achieved 69% and 71% accuracy for arousal and valence levels, respectively. Subsequently, a second variant of ECG model is also trained by including a mutual attention module. For this configuration of ECG model, the arousal and valance accuracies improved by 3.45% and 6.7%, respectively. These findings depict the significance and efficacy of the attention layer that ensures the selection of suitable and emotion-specific features. Afterward, the third variant of ECG model with a mutual attention module and dense max-margin loss function is trained. Table 1 presents the performance of all three variants of the ECG model. These experiments elucidate that the dense max-margin loss function in addition to the mutual attention module significantly enhances the performance of the classifier and achieves 78.55% accuracy for Arousal and 82.35% accuracy for valence.

B. EEG BASED CLASSIFICATION

To evaluate EEG-based emotion recognition model, we use 14 channels EEG data from the AMIGOS dataset [12]. We use one-second long segments of four different frequency bands of EEG data, a similar approach as adopted in [24]. The differential entropy for each band in each channel is computed and a tensor with a dimension of 32×4 is used as model input. The proposed model with an input tensor of size 32×4 can significantly characterize the interrelation between different EEG channels. The reverse is also possible. In other words, the model can also learn and determine the correlation between different frequency bands as well if the input with a shape 4×32 is used as input to the model. The mutual attention module masks the input by assigning learned weights to each channel and its features. Unlike [49], the proposed attention layer learns mutual attention using channels mean and features mean and thus assures that suitable and emotion-specific information propagates forward in the model for better feature representation learning. We use L2 regularization along with dropout to address the overfitting. For the dense max-margin loss function, we kept the 0.5 weightage. We used a batch size of 128 for training EEG-based model.

We also trained three variants of EEG based emotion classification model. First, we train a baseline model without a mutual attention layer using conventional cross-entropy loss function and achieve 68.34% and 67.68% accuracy for arousal and valence, respectively. Second, to examine the impact of the attention module, the baseline model integrated with the mutual attention module is trained using the cross-entropy loss function. For this second configuration of EEG model, we get 72.23% accuracy for arousal and 75.36% accuracy for valence. This improvement in classification performance illustrates that the proposed attention module also significantly emphasizes the emotion-related content in EEG features. In the last, we train and evaluate the proposed EEG model with a mutual attention mechanism using a dense max-margin loss function. Table 2 illustrates the performance of three variants of the EEG model. The proposed methodology achieves 86.26% and 84.40% accuracy for arousal and valence, respectively. These findings elucidate the effectiveness of the mutual attention module and

TABLE 1. Classification performance of ECG-based model on AMIGOS dataset.

Model Variants	Accuracy (± Std.)	
	Arousal	Valence
CNNLSTM + Cross Entropy	69%(±3.66)	71%(±3.28)
CNNLSTM + Attention + Cross Entropy	72.45%(±2.80)	77.7%(±2.53)
CNNLSTM + Attention + Dense Max-margin	78.55%(±2.47)	82.35%(±2.11)

TABLE 2. Classification performance of EEG-based model on AMIGOS dataset.

Model Variants	Accuracy (± Std.)	
	Arousal	Valence
CNNLSTM + Cross Entropy	68.34% (±2.18)	67.68% (±2.34)
CNNLSTM + Attention + Cross Entropy	72.23% (±1.93)	75.36% (±2.53)
CNNLSTM + Attention + Dense Max-margin	86.26% (±1.17)	84.40% (±1.36)

TABLE 3. Multi-modal classification performance on AMIGOS dataset.

Model Variants	Accuracy (± Std.)	
	Arousal	Valence
CNNLSTM + Cross Entropy	73.45% (±1.77)	68.44% (±1.84)
CNNLSTM + Attention + Cross Entropy	87.3% (±1.35)	87.72% (±1.32)
CNNLSTM + Attention + Dense Max-margin	90.54% (±1.05)	89.16% (±1.13)

dense max-margin loss function for imbalanced EEG data classification.

C. MULTI-MODAL EMOTION CLASSIFICATION

The multi-modal network takes both EEG and ECG segments that share the exact time span to classify emotions. Therefore, we use 10-second long synchronized segments of EEG and ECG data. We extract the HRV series with zero padding from ECG segments. Similarly, for EEG, the differential entropy features are computed from the 10-second segment of EEG in each frequency band. The evaluation of multi-modal emotion classification is performed on AMIGOS [12] and YSRP dataset. Table 3 displays the potential of the proposed method to differentiate entangled and unevenly distributed features of arousal and valence classes. These findings demonstrate the significance of the proposed approach in distinguishing between complex emotions. First, the deep architecture assisted by the mutual attention module performs a key role in learning suitable and emotion-specific deep representation. Afterward, the max-margin loss function enforces the required margin between class boundaries. We also evaluated three different variants of multi-modal system for emotion recognition. Table 3 demonstrates the classification performance of the three different configurations of the multi-modal emotion recognition system. These results indicate that The proposed mutual attention mechanism and dense max-margin loss function outperform the conventional supervision method and achieve better classification performance with imbalanced data.

Table 4 illustrates the potential of the proposed emotion classification methodology to classify emotions using different datasets. Moreover, the imbalanced ratio (IR) of different datasets for two classes of arousal and valance is also given in Table 4. The deep model incorporated with a mutual attention module supervised by a dense max-margin loss function significantly elevates the classifier's confidence in discriminating complex emotions. These results also demonstrate that the proposed supervision strategy efficiently alleviates the impact of imbalanced data distribution on deep representation learning. The multi-modal classification results of YSRP dataset exhibit lower classification accuracy in comparison to the accuracy score achieved by the model on AMIGOS dataset. This reduction in classification score is due to the decrease in number of samples caused by lower number of participants. For segmentation, a window size of 10 seconds is used for both EEG and ECG data which reduces the total number of samples for training.

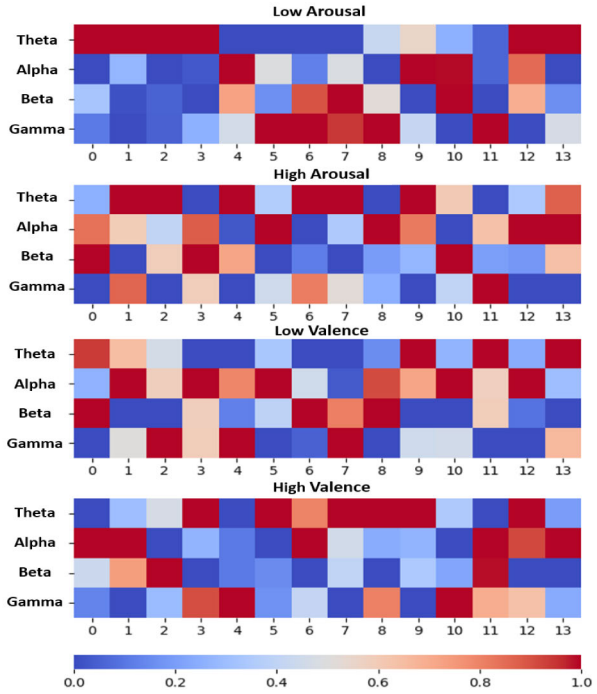

FIGURE 7. Attention visualization of EEG model.

D. IMPACT OF ATTENTION MECHANISM

To interpret the impact of the proposed attention mechanism and dense max-margin loss function, we considered EEG signals for further investigation for the following reasons. First, EEG is the most suitable and optimal signal to inspect the autonomic nervous system's activation triggered by any external or internal stimulus [7], [54]. Second, EEG has been widely investigated for emotion and therefore, the findings of this study can easily be corroborated with other theoretical studies. Third, in comparison to ECG, EEG model with 2D input (channels x frequency bands) is the most suitable candidate to evaluate the impact of the mutual attention mechanism.

The proposed mutual attention mechanism is designed to learn and determine both the efficient channels and the corresponding frequency bands in EEG that contribute to emotion recognition. It generates a mutual statistical matrix based on the importance of EEG channels and associated bands for a target emotion. During computation, both channel-wise and feature-wise mean are considered to incorporate both axes (channels and bands) to learn a mutual mask for the input tensor. We examine the attention mask learned by the EEG model for two levels of arousal and valence. The acquired attention maps of the EEG model are given in Fig.7. It shows clearly that the attention mechanism does not treat all the channels uniformly but rather allocates

TABLE 4. Classification performance on different datasets.

Model	Dataset	Description	Imbalance Ratio (IR)	Accuracy (± Std.)	
				Arousal	Valence
EEG+ECG	AMIGOS	Subjects: 39 Signals: EEG (14 channels), ECG (single channels)	Arousal: 1.08 Valence: 1.19	90.54% (±1.05)	89.16% (±1.13)
EEG+ECG	YSRP	Subjects: 25 Signals: EEG (2 channels), ECG (single channels)	Arousal: 1.14 Valence: 1.24	83.21% (±1.28)	81.46% (±1.46)

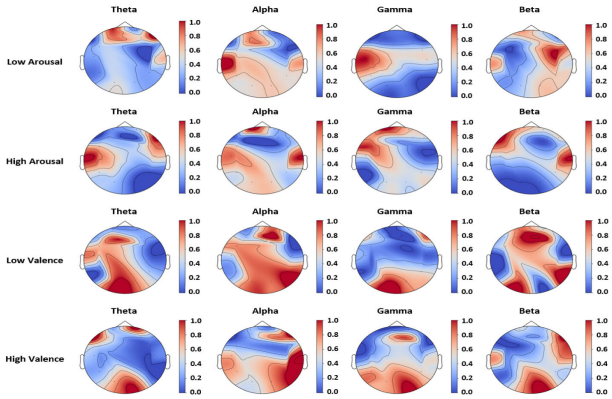

FIGURE 8. Topographic maps for attention visualization of EEG model.

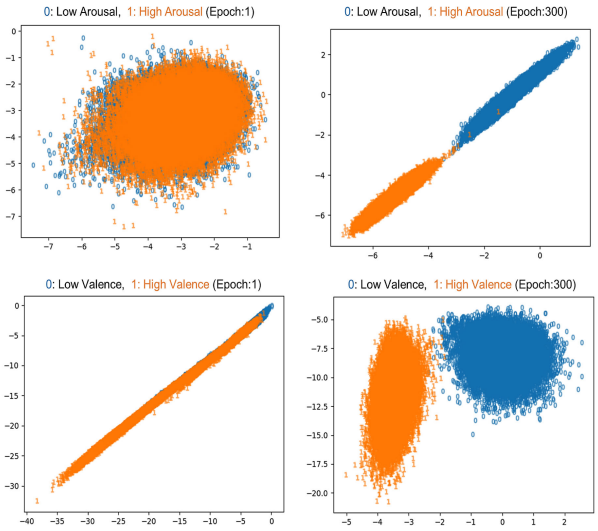

FIGURE 9. Feature visualization of arousal and valence models.

relative significance by assigning special weights to each EEG channel. Similarly, different importance score is also given to each band of EEG based on their contribution to the target task. The arousal and valance classes show clear discrimination between the selection of channels and features for processing through the model. This approach of emphasizing the most appropriate and emotion-specific features leads the model to produce remarkable results.

The improvement in classification accuracy acquired with the integration of the mutual attention module implies that the relative significance score assigned by the attention module corresponds to the autonomic nervous system's activation. Moreover, the attention mask is additionally associated with a considerable increase or decrease in EEG power for the target task. Numerous studies have investigated the correlation between EEG (bands) and emotions [24], [35], [37], [54]. For instance, an upsurge of power in low-frequency bands such as theta and alpha is inked with an elevation in valence level [24]. Similarly, an increase in activation of the left temporal lobe corresponds to sadness while a heightened activity in the right temporal lobe is associated with happiness [24]. In addition, an emotion lateralization hypothesis also suggests that the center for positive emotions exists in the left part of the brain while the negative feelings are processed in the right part of the brain. Fig.8 shows the topographic maps for attention and substantiates previous findings in the literature. The discrimination between various bands in two classes of arousal and valence significantly justify that these findings corroborate previous theories [24], [35], [37], [54] of emotion lateralization and can be seen clearly by the attention masks depicted in Fig.8.

E. IMPACT OF DENSE MAX-MARGIN LOSS FUNCTION

The proposed dense max-margin loss function aims to enhance the classifier's confidence to accurately classify adjacent and hard entangled samples of different emotions. Fig.9 shows feature visualization of arousal and valence. For the arousal and valence models, the significant impact of the proposed dens max-margin loss function can be observed by visualizing the feature space. At the start of training, the samples from different levels of arousal and valence yield an overlap and uneven feature distribution. However, the supervision of the model under dense max-margin loss function efficiently discriminates class clusters with a significant margin after 300 epochs. A clear depiction of margin induction and variance reduction can be seen in Fig.9.

In our view, the classification performance and visualization illustration emphasize the validity of the proposed dense max-margin loss function. These results offer compelling evidence to endorse the implications of three things in the proposed loss function. First, the reduction in variance inside the class cluster. Second, the induction of discrimination margin between classes, and third, the incorporation of a hard mining strategy. These three factors collectively

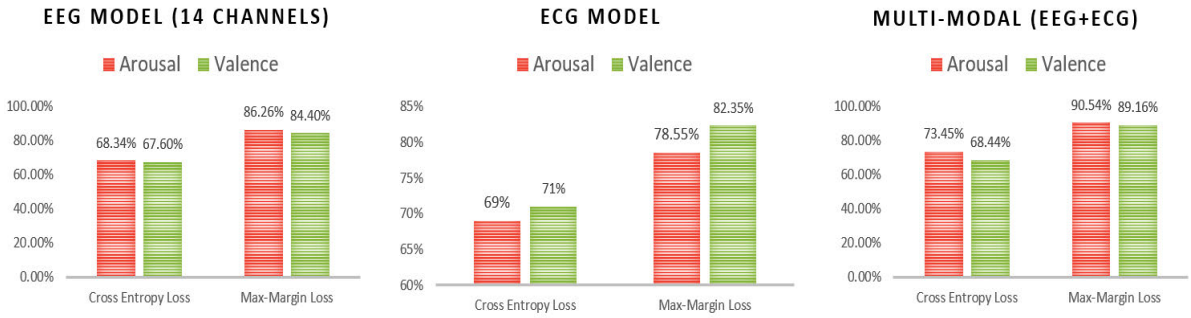

FIGURE 10. Comparison between cross entropy and dense max-margin loss function.

TABLE 5. Performance comparison.

Article	Signals(Dataset)	Classifier	Accuracy	
			Arousal	Valence
L. Santamaria et al. [42] (2018)	ECG, GSR (AMIGOS)	Convolutional Neural Network(CCN)	76.0%	75.0%
HC Yang and CC Lee [38] (2019)	EEG, ECG, GSR (AMIGOS)	Attribute-Invariant VAE	64.4%	67.1%
Chao lee et al. [47] (2020)	EEG, ECG, GSR (AMIGOS)	Attention-based bidirectional LSTM	82.5%	77.8%
Rodriguez et al. [48] (2022)	EEG, ECG (AMIGOS)	Transformer With 1D CNN	89%	85%
Bota, Patrícia, et al. [43] (2023)	EEG, ECG, GSR (AMIGOS)	1-dimensional residual temporal and channel attention network	83.07%	82.20%
Pan, Tongjie, et al. [44] (2023)	EEG, ECG, GSR (AMIGOS)	Online Multimodal Hyper Graph Learning (OMHGL)	87%	83%
Pan, Tongjie, et al. [45] (2024)	EEG, ECG, GSR (AMIGOS)	Multi-hypergraph fusion learning	80.8%	86.9%
Proposed	**EEG, ECG (AMIGOS)**	**CNN + LSTM, Dense Max-margin**	**90.54%**	**89.16%**

contribute to yielding overwhelming results. Similarly, the incorporation of a hard sampling mining strategy also boosts classification performance. The adjacent emotions induced by external stimuli are responsible for yielding hard negative samples. Moreover, the imbalanced distribution of data among different classes also contributes to hard negative samples. However, the classification results reinforce the usefulness of assimilating the hard mining strategy into a loss function that alleviates the issue of adjacent emotions and imbalanced data distribution.

Similarly, keeping the imbalance ratio (as shown in Table able 4) into consideration, the proposed loss function learned an efficient representation of ECG and EEG for two-class classification of arousal and valence as compared to the conventional cross-entropy loss function. Fig.10 illustrates the performance comparison between cross entropy and dense max-margin loss function. Unlike the cross entropy function, our proposed loss function mitigates the issue of adjacent emotions and imbalanced data by providing the flexibility to induce margin and reduce cluster variance at the feature level.

F. PERFORMANCE COMPARISON

This section presents a comparison between the experimental outcomes of the proposed emotion recognition methodology and the most relevant studies. For a fair comparison on the same basis, we chose the studies that reported multimodal results and employed AMIGOS database for model training and evaluation. Table 5 gives the summary of the most relevant studies. A most relevant study that addressed the issue of a discriminative margin between adjacent emotions by introducing the temporal margin loss function [37] achieved 79.03% accuracy for arousal and 78.72% accuracy for valence on DEAP dataset [24]. Although in [37], the experiments are performed on a different datasets, however, we compare this study based on its relevance. Unlike [37], our proposed loss function operates at the feature level to induce margin and mitigate the adverse impacts of skewed data distribution on classification performance. Similarly, the proposed method also outperforms the most recently published methods that exploit graph neural networks for emotion classification [43], [44], [45].

Moreover, the bidirectional LSTM-RNNs embedded with attention mechanism are used for multi-modal emotion recognition in [47]. However, the proposed mutual attention mechanism achieved high classification performance in comparison to [47] due to the fact that the mutual attention module significantly selects the mutually important information on the channel as well as on the feature axis. Similarly, the the proposed model with mutual attention also

outperforms the pre-trained transformers embedded model for the classification of emotion. In comparison to the relevant studies mention in Table 5, our proposed methodology demonstrated better classification performance.

G. CHALLENGES AND LIMITATIONS

In this study, the most challenging step is the efficient optimization of the model. The proposed model include LSTM layers that are prone to overfitting, especially on smaller datasets, and are sensitive to the configuration of hyperparameters. Finding the right set of hyperparameters can be time-consuming and requires extensive experimentation. Additionally, during the deep representation learning of physiological signals, the model may easily encounter vanishing gradient problem. Therefore considerable attention is needed to be paid during training.

VI. CONCLUSION

This research aims to develop a physiological signal-based emotion recognition system with a principle focus on learning an efficient deep representation for classifying complex emotions using imbalanced data. For spatio-temporal feature extraction, a deep architecture with a mutual attention module is designed that grabs the short and long-term variations of physiological signals and focuses on the most relevant features of the target class. Additionally, for efficient discrimination of adjacent emotions with imbalanced data, this study introduces a novel dens max-margin loss function that minimizes intra-class variance and maximizes inter-class margin for efficient classification.

The results of this study significantly substantiate the efficacy of the proposed emotion recognition method. The impact of the mutual attention module is considerably high which determines the relative importance of channel and channel-related features of EEG and ECG signals for a particular target. In addition, the maximum margin and minimum variance approach based on Gaussian similarity measure significantly improves the classification of adjacent emotions with entangled and unevenly distributed features in emotion recognition. Future work should concentrate on data scalability issues and integrating multiple channels of affect (facial, speech, and postures) for emotion analysis to further improve classification performance.

REFERENCES

[1] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biol. Psychol.*, vol. 84, no. 3, pp. 394–421, Jul. 2010.

[2] Y.-D. Zhang, Z.-J. Yang, H.-M. Lu, X.-X. Zhou, P. Phillips, Q.-M. Liu, and S.-H. Wang, "Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation," *IEEE Access*, vol. 4, pp. 8375–8385, 2016.

[3] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[4] L. Sun, Q. Li, S. Fu, and P. Li, "Speech emotion recognition based on genetic algorithm–decision tree fusion of deep and acoustic features," *ETRI J.*, vol. 44, no. 3, pp. 462–475, Jun. 2022.

[5] F. Noroozi, C. A. Corneanu, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 505–523, Apr. 2021.

[6] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, Nov. 2005, pp. 677–682.

[7] 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, Jan. 2009.

[8] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[9] J. Kim and E. Andre, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, Dec. 2008.

[10] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, "Recognizing emotions induced by affective sounds through heart rate variability," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 385–394, Oct. 2015.

[11] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 147–160, Apr. 2018.

[12] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Apr. 2021.

[13] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[14] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, no. 4616, pp. 1208–1210, Sep. 1983.

[15] P. Ekman, "An argument for basic emotions," *Cognition Emotion*, vol. 6, nos. 3–4, pp. 169–200, May 1992.

[16] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Med. Biol. Eng. Comput.*, vol. 42, no. 3, pp. 419–427, May 2004.

[17] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *Amer. Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[18] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, p. 1161, 1980.

[19] P. J. Lang, "The emotion probe: Studies of motivation and attention," *Amer. Psychologist*, vol. 50, no. 5, pp. 372–385, 1995.

[20] M. S. Chavan, R. Agarwala, and M. Uplane, "Suppression of baseline wander and power line interference in ECG using digital IIR filter," *Int. J. circuits, Syst. signal Process.*, vol. 2, no. 2, pp. 356–365, 2008.

[21] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985.

[22] M. Murugappan, S. Murugappan, and B. S. Zheng, "Frequency band analysis of electrocardiogram (ECG) signals for human emotional state classification using discrete wavelet transform (DWT)," *J. Phys. Therapy Sci.*, vol. 25, no. 7, pp. 753–759, 2013.

[23] 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.

[24] 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.

[25] A. GuruvaReddy and S. Narava, "Artifact removal from EEG signals," *Int. J. Comput. Appl.*, vol. 77, no. 13, pp. 17–19, Sep. 2013.

[26] G. Valenza, A. Lanata, and E. P. Scilingo, "The role of nonlinear dynamics in affective valence and arousal recognition," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 237–249, Apr. 2012.

[27] 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.

[28] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001.

[29] S. Basu, N. Jana, A. Bag, M. Mahadevappa, J. Mukherjee, S. Kumar, and R. Guha, "Emotion recognition based on physiological signals using valence-arousal model," in *Proc. 3rd Int. Conf. Image Inf. Process. (ICIIP)*, Dec. 2015, pp. 50–55.
[30] J. Liu, H. Meng, A. Nandi, and M. Li, "Emotion detection from EEG recordings," in *Proc. 12th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, Aug. 2016, pp. 1722–1727.
[31] P. Rani, C. Liu, N. Sarkar, and E. Vanman, "An empirical study of machine learning techniques for affect recognition in human–robot interaction," *Pattern Anal. Appl.*, vol. 9, no. 1, pp. 58–69, 2006.
[32] H. F. García, M. A. Álvarez, and Á. A. Orozco, "Gaussian process dynamical models for multimodal affect recognition," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 850–853.
[33] G. Valenza, L. Citi, A. Lanatá, E. P. Scilingo, and R. Barbieri, "Revealing real-time emotional responses: A personalized assessment based on heartbeat dynamics," *Sci. Rep.*, vol. 4, no. 1, pp. 1–13, May 2014.
[34] H.-W. Guo, Y.-S. Huang, C.-H. Lin, J.-C. Chien, K. Haraikawa, and J.-S. Shieh, "Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine," in *Proc. IEEE 16th Int. Conf. Bioinf. Bioengineering (BIBE)*, Oct. 2016, pp. 274–277.
[35] 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.
[36] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, "Emotion recognition from multi-channel EEG data through convolutional recurrent neural network," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 352–359.
[37] B. H. Kim and S. Jo, "Deep physiological affect network for the recognition of human emotions," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 230–243, Apr. 2020.
[38] H.-C. Yang and C.-C. Lee, "An attribute-invariant variational learning for emotion recognition using physiology," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1184–1188.
[39] Siddharth, T.-P. Jung, and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Trans. Affect. Comput.*, vol. 13, no. 1, pp. 96–107, Jan. 2022.
[40] P. Bota, T. Zhang, A. E. Ali, A. Fred, H. P. da Silva, and P. Cesar, "Group synchrony for emotion recognition using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 2614–2625, Oct. 2023.
[41] W. Lin, C. Li, and S. Sun, "Deep convolutional neural network for emotion recognition using EEG and peripheral physiological signal," in *Proc. Int. Conf. Image Graph.* Cham, Switzerland: Springer, 2017, pp. 385–394.
[42] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar, "Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS)," *IEEE Access*, vol. 7, pp. 57–67, 2019.
[43] H. Huang, M. Fan, and C.-A. Chou, "Graph-based learning of nonlinear physiological interactions for classification of emotions," *Pattern Recognit.*, vol. 143, Nov. 2023, Art. no. 109794.
[44] T. Pan, Y. Ye, H. Cai, S. Huang, Y. Yang, and G. Wang, "Multimodal physiological signals fusion for online emotion recognition," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 5879–5888.
[45] T. Pan, Y. Ye, Y. Zhang, K. Xiao, and H. Cai, "Online multi-hypergraph fusion learning for cross-subject emotion recognition," *Inf. Fusion*, vol. 108, Aug. 2024, Art. no. 102338.
[46] P. Boonthong, P. Kulkasem, S. Rasmequan, A. Rodtook, and K. Chinnasarn, "Fisher feature selection for emotion recognition," in *Proc. Int. Comput. Sci. Eng. Conf. (ICSEC)*, Nov. 2015, pp. 1–6.
[47] 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.
[48] J. Vazquez-Rodriguez, G. Lefebvre, J. Cumin, and J. L. Crowley, "Emotion recognition with pre-trained transformers using multimodal signals," in *Proc. 10th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2022, pp. 1–8.
[49] 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.
[50] M. Hayat, S. Khan, S. W. Zamir, J. Shen, and L. Shao, "Gaussian affinity for max-margin class imbalanced learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2019, pp. 6469–6479.
[51] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 209–222, Jul. 2015.
[52] W.-L. Zheng, H.-T. Guo, and B.-L. Lu, "Revealing critical channels and frequency bands for emotion recognition from EEG with deep belief network," in *Proc. 7th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Apr. 2015, pp. 154–157.
[53] 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. 2020.
[54] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 374–393, Jul. 2019.

MUHAMMAD ZUBAIR received the B.Sc. degree in electrical communication engineering from the University of Engineering and Technology (UET), Peshawar, Pakistan, in 2014, and the Ph.D. degree in information and communication technology (ICT) from Korea University of Science and Technology (UST), in 2021. He is currently a Postdoctoral Researcher with the Autonomous IoT Research Section, Electronics and Telecommunications Research Institute (ETRI), South Korea. His research interests include affective computing, deep representation learning, the Internet of Things (IoT), and AI-based healthcare applications.

SUNGPIL WOO received the B.S. and M.S. degrees from the School of Computing, Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2015 and 2017, respectively, where he is currently pursuing the Ph.D. degree. He is a Researcher with the Autonomous IoT Research Section, Electronics and Telecommunications Research Institute (ETRI), South Korea. His research interests include deep multimodal representation learning and reinforcement learning for healthcare and the Internet of Things (IoT).

SUNHWAN LIM received the B.S. and M.S. degrees in information and communication engineering from Chonbuk National University, South Korea, in 1997 and 1999, respectively, and the Ph.D. degree in information and communication engineering from Chungnam National University, in August 2011. Since September 1999, he has been with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea, where he is currently a Senior Researcher with the Autonomous IoT Research Section. His research interests include AI-data commons, the Internet of Things, and digital twin.

CHANGWOO YOON received the B.S. degree from Sogang University, the M.S. degree from Pohang University of Science and Technology (POSTECH), Pohang, South Korea, and the Ph.D. degree in computer and information science and engineering from the University of Florida, Gainesville, FL, USA, in 2005. He is currently a Principal Researcher with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. He is also a Professor with the Information and Communication Technology Department, Korea University of Science and Technology (UST), Daejeon. His current research interests include cognitive computing, artificial intelligence, wearable computing, N-screen, IPTV, cloud computing, SOA, and service creation/delivery technology.