IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 January 2023, accepted 31 January 2023, date of publication 15 February 2023, date of current version 23 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3245830

RESEARCH ARTICLE

Emotion Recognition System Based on Two-Level Ensemble of Deep-Convolutional Neural Network Models

MUHAMMAD HUSSAIN[1], (Senior Member, IEEE), EMAD-UL-HAQ QAZI[1], HATIM A. ABOALSAMH[1], (Senior Member, IEEE), AND IHSAN ULLAH[2]

[1]Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[2]Insigth SFI Research Center for Data Analytics, School of Computer Science, University of Galway, Galway, H91 TK33 Ireland

Corresponding author: Muhammad Hussain (mhussain@ksu.edu.sa)

The authors extend their appreciation to the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project no. (IFKSURG-2-109).

ABSTRACT Emotions play a crucial role in human interaction and healthcare. This study introduces an automatic emotion recognition system based on deep learning using electroencephalogram signals. A lightweight pyramidal one-dimensional convolutional neural network model is proposed that involves a small number of learnable parameters. Using the model, a two-level ensemble classifier is designed. Each channel is scanned incrementally in the first level to generate predictions, which are fused using the majority vote. The second level fuses the predictions of all the channels of a signal using a majority vote to predict the emotional state. The method was validated using the public domain challenging benchmark DEAP dataset. The electroencephalogram signals over five brain regions were analyzed. The results indicate that the frontal brain region plays a dominant role, achieving accuracies of 98.43% and 97.65% for two emotion recognition problems (distinguishing high valence vs. low valence and high arousal vs. low arousal states).

INDEX TERMS Convolutional neural networks, deep learning, electroencephalography, emotion recognition, expert systems, feature extraction, machine learning, pattern classification, psychology, signal processing.

I. INTRODUCTION

Emotion is a psychophysiological process triggered by the conscious or unconscious perception of an object or situation, and is often associated with temperament, mood, motivation, and personality. Emotions play an essential role in human healthcare, communication, and security investigations and can be expressed either verbally through emotional statements or nonverbally through cues, such as facial expressions, the intonation of voice, and bodily gestures [1]. Emotions affect decision-making, mutual interactions, and cognitive processes [2]. Most generic methods of emotion recognition include pan-cultural elements and constants across cultures that use facial expressions [63], [64], [65], [66]. The pan-cultural element in the facial displays of emotion is the association between facial muscular movements and discrete primary emotions, although cultures differ in what evokes emotion, rules for controlling the display of emotion, and behavioral consequences. Advancements in technology and emotion understanding have led to growing opportunities for automatic emotion recognition (AER) systems. Many real-world applications are based on AER systems, including security, driver fatigue monitoring, health monitoring, and interactive computer simulations and designs. AER systems can also easily detect emotions using facial expressions. Affective computing is one such domain, which aims to bridge the gap between humans and computers. The goal of affective computing is to make informed decisions based on the emotions expressed by human subjects to make personalized decisions [57].

While many studies on emotion recognition use modalities such as facial expressions, speech, text, or gestures, these

The associate editor coordinating the review of this manuscript and approving it for publication was Gerard-Andre Capolino.

modalities are based on audio and visual observations, which can be easily disguised. One alternative is to use electroencephalograms (EEGs) to capture the brain signals activated by various types of emotions [53]. EEG is a neuroimaging technique that is commonly used to analyze neural processes. From a clinical point of view, EEG signals directly capture and map brain activations associated with different emotional states, and thus cannot be disguised. Therefore, EEG brain signals are an optimal modality for detecting genuine emotions [3], [4], [5], [6], [7], [8], [9], [10].

In psychology research, two major approaches for modeling emotions are (1) the categorical approach and (2) the dimensional approach [11]. Darwin et al. pioneered the categorical approach [12], focusing on the basic emotions of happiness, sadness, surprise, disgust, anger, and fear. According to the dimensional approach, affective states are not independent; instead, they are systematically related to one another. In this approach, a model of emotion is characterized by one or more dimensions including valence and arousal [13]. Valence is the degree of aversion or attraction that an individual feels about a specific event or object, and ranges from negative to positive. Arousal is the physiological and psychological state of being reactive to stimuli, and ranges from passive (inactive) to active. The dimensional approach is assumed to be better than the categorical approach because it describes a larger set of emotions [14]. Therefore, the problem with emotion recognition at a high level is distinguishing high valence from low valence (HV vs. LV), and high arousal from low arousal (HA vs. LA).

Several studies have recently employed EEG brain signals to classify HV versus LV and HA versus LA. Most of this research has used machine learning (ML) techniques based on hand-engineered features [15], [16], [17], [18], which exhibit limited performance in emotion recognition. A pioneering study by Koelstra et al. [15] extracted power spectrum features from 32 EEG signals and classified them using a Gaussian naïve Bayes (NB) classifier into two levels of valence and arousal. Alazrai et al. [16] also employed EEG signals from the DEAP dataset, extracted quadratic time-frequency distribution (QTFD) features, and used a support vector machine (SVM) as a classifier. Huang et al. [17] introduced an asymmetric spatial pattern (ASP) as a feature extracted from EEG signals and used naïve k-nearest neighbors (K-NN), NB, and SVM for emotion classification.

Similarly, several other studies have used hand-engineered feature extraction techniques on EEG signals and performed emotion classification using various classifiers including NB, SVM, K-NN, linear discriminant analysis, and artificial neural networks [18]. The maximum reported accuracy levels for HV vs. LV and HA vs. LA classifications in the DEAP database were 85.8% and 86.6%, respectively [16], indicating that hand-engineered features cannot correctly represent the discriminative patterns from EEG signals relevant to emotions. This, in turn, means that existing methods have not reached the desired level with regard to classifying human emotions and that emotion recognition from EEGs remains a challenging problem.

Deep learning (DL) is a state-of-the-art ML technique that automatically learns the hierarchy of features and classifies them in an end-to-end fashion [19]. The features extracted by DL models are adapted to the inherent structural patterns of the data; therefore, they are more discriminative and robust than hand-engineered features [20], [54]. Unlike conventional ML techniques, DL does not require the design of feature descriptors, selection of the most discriminative features, or adaptation of a suitable classifier [21]. The most effective DL architecture is the convolutional neural network (CNN). Various two- and three-dimensional CNN models, such as AlexNet, VGG, 3DCNN, and C3D [22], have shown excellent performance in many fields. Recently, one-dimensional (1D) CNN models have been successfully used for music generation, epilepsy detection, text understanding, and other time-series data [23], [24].

Motivated by the outstanding performance of DL in many recognition tasks, we used DL to develop a robust and effective AER system based on EEG signals. Because DL models involve a large number of learnable parameters, their training requires a large dataset of EEG signals, which is difficult to acquire for the AER problem. To overcome this problem, we propose a lightweight pyramidal 1D CNN (LP-1D-CNN) model containing a smaller number of learnable parameters. An EEG signal consists of several channels, each of which is 1D. To predict emotions from an EEG signal, each channel must be analyzed. The temporal length of each channel is normally large; for example, the temporal length of each channel in the DEAP is 1 min and consists of 8,064 samples. For the analysis of each channel, in the case of a 1D CNN topology, a large number of input parameters directly influence the model complexity and proneness to overfitting.

To overcome this problem, we first segmented each channel into small windows and trained an LP-1D-CNN model on these windows. Because the size of each input window was small, the complexity of the LP-1D-CNN model was low, and it was robust to overfitting. At the test time, the predictions of all the windows of a channel by the LP-1D-CNN model were fused, and the decisions from each channel were used to predict the emotional state from the EEG signal. Using LP-1D-CNN, we built a deep automatic emotion recognition (Deep-AER) system, which is a two-level ensemble model for emotion classification.

We validated the effectiveness and robustness of Deep-AER by using the benchmark DEAP dataset. We focused on two emotion classification problems: HV versus LV, and HA versus LA. To identify the brain regions that play a dominant role in AER, we analyzed EEG signals in five brain regions: *FRONT*, *CENT*, *PERI*, *OCCIP*, and *ALL*. These results confirm that *FRONT* plays a dominant role in the AER. The main contributions of this study are as follows.

1) A new lightweight 1D CNN (LP-1D-CNN) model and a data augmentation technique for its training;

2) The Deep-AER system for the classification of emotions from EEG signals based on the two-level ensemble of LP-1D-CNN models; and

3) The analysis of brain regions to authenticate the effectiveness of the proposed method and the region which plays a dominant role in emotion recognition.

The remainder of this paper is organized as follows. Section II presents an overview of related work. Section III describes the Deep-AER system framework based on DL in detail. Section IV presents the experimental protocol and evaluation criteria. Section V presents the results and discussion, and Section VI concludes the study.

II. RELATED WORK

Several EEG-based AER systems have emerged in recent years. The categorization of emotions is a classification problem that involves the extraction and subsequent classification of discriminatory features from EEG signals. In the following paragraphs, we review state-of-the-art techniques proposed to solve this problem.

The pioneering work by Koelstra et al. [15] extracted power spectrum density (PSD) features from EEG signals and classified them using a Gaussian NB classifier into two levels of valence and arousal. This method achieved accuracy levels of 57.6% for HV versus LV, and 62HA versus LA in the DEAP database.

Alazrai et al. [16] employed EEG signals from the DEAP dataset, extracted QTFD-based features, and used an SVM as a classifier. The authors classified valence and arousal into two states (high and low) and achieved accuracies of 85.8% and 86.6% for HV versus LV, and HA versus LA, respectively.

Huang et al. [17] introduced ASP as a feature extracted from EEG signals, and used K-NN, NB, and SVM for emotion classification. The average accuracies achieved using this method for valence (HV versus LV) and arousal (HA versus LA) were 66.05% and 82.46%, respectively, for DEAP.

Chung and Yoon [25] classified valence and arousal using statistical and shallow learning methods such as Bayesian classification. The authors extracted power spectral features from the EEG signals and classified them using the Bayes classifier. They divided valence and arousal into two classes (HV vs. LV and HA vs. LA). On the DEAP dataset, they achieved accuracies of 66.6% and 66.4%for valence and arousal, respectively.

Candra et al. [26] used a discrete wavelet transform to extract time-frequency domain features from EEG signal. They computed the entropy of the detail coefficients corresponding to the alpha, beta, and gamma bands, and used an SVM to classify valence and arousal as high or low. This method achieved accuracies of 65.13% for valence (HV versus LV) and 65.33% for arousal (HA versus LA.

Rozgic et al. [27] introduced a method that extracts discriminative features using PSD. The authors used SVM, NB nearest neighbors, and nearest-neighbor voting for classification. The SVM had the best classification accuracy at 76.9% and 69.1for HV vs. LV and HA vs. LA, respectively.

Abeer et al. [28] proposed a method for classifying the valence and arousal dimensions into two classes (HV vs. LV and HA vs. LA). The authors extracted PSD and prefrontal asymmetry features from EEG signals and used a deep neural network (DNN) as a classifier. The proposed technique achieved an accuracy of 82% for HV versus LV, and HA versus LA on the DEAP dataset.

The method proposed by Zhang et al. [29] extracts power spectral and statistical features from EEG signals and classifies them using the J48 classifier. They divided the data into two classes: HV versus LV, and HA versus LA. The authors used an ontological model for the integration and representation of EEG data. On the DEAP dataset, this method achieved 75.19% and 81.74% accuracy for HV versus LV, and HA versus LA, respectively.

Liu et al [30] proposed an approach based on deep belief networks (DBNs) for the classification of valence and arousal dimensions into two classes (HV vs. LV and HA vs. LA). The accuracies obtained were 85.2% and 80.5% for HV versus L, and HA versus LA, respectively.

Atkinson and Campos [31] extracted a set of features, such as Hjorth parameters, fractal dimension, statistical features, and band power, for various frequency bands. They also used the mRMR algorithm to select a subset of discriminative features from the set of extracted features. This method achieved accuracies of 73.1% and 73.0% for HV versus L, and HA versus LA, respectively.

Tripathi et al. [32] extracted statistical time-domain features from EEG signals and used two types of neural networks (DNN and CNN) as classifiers to differentiate EEG signals into two classes: HV vs. LV and HA vs. LA. The accuracies obtained for the DEAP dataset when using the DNN to distinguish HV vs. LV and HA vs. LA classes were 75.78% and 73.12%, respectively. Similarly, the accuraciesof CNN in classifying HV vs. LV and HA vs. LA classes were 81.4 and 73.3%, respectively.

Yin et al [33] used a stacked autoencoder to classify EEG signals into HV vs. LV and HA vs. LA classes. The authors extracted power spectral and statistical features fromthe EEG signals. The accuracies achieved for HV vs. LV and HA vs. LA on the DEAP dataset were 83.04% and 84.18%, respectively.

Zhuang et al. [34] used empirical mode decomposition (EMD) for emotion recognition. In this approach, EMD-based features were extracted from EEG signals, and an SVM was used as a classifier to discriminate between HV vs. LV and HA vs. LA classes. The results showed that the accuracies attained using this method were 69.1% and 71.9% on the DEAP dataset for HV versus LV, and HA versus LA, respectively.

Li et al. [35] extractedthe nonlinear dynamic domain, frequency domain, and time-domain features from EEG signals and used an SVM as a classifier. The accuracies obtained for classifying HV vs. LV and HA vs. LA classes were 80.7% and 83.7respectively, on the DEAP dataset.

Menezes et al [36] proposed an approach that extracts the PSD, higher-order crossings, and statistical feature, and uses different classifiers, such as random forest and SVM. The highest accuracies achieved on the DEAP dataset forthe HV vs. LV and HA vs. LA classes using the SVM classifier were 88.4% and 74.0%, respectively.

Li et al. [37] introduced an R2G-STNN method, proposing spatial and temporal neural network models with regional to global hierarchical feature-learning processes to learn discriminative spatial-temporal EEG features. The authors used a bidirectional long short-term memory (LSTM) network to learn both regional and global spatiotemporal features. The SEED database was used to validate the proposed method. They achieved an accuracy of 93.38% in an emotion recognition experiment that identified positive versus neutral versus negative classes.

Shen et al. [38] proposed a parallel sequence-channel projection CNN that includes a temporal stream subnetwork, a spatial stream subnetwork, and a fusion classification block. The authors used the temporal stream to extract temporal continuity via the sequence projection layer, whereas the spatial stream was used to capture spatial correlation via the channel projection layer. They achieved accuracies of 96.16% and 95.89% forvalence and arousal, respectively, on the DEAP dataset.

The above methods follow a conventional approach, that is, they first extract features using different techniques and then perform classification using various methods. End-to-end DL methods have been proposed, as employed for other 1D signals such as ECG [39]. Zhang et al. [40] proposed an automatic epilepsy and seizure classification approach based on multispike liquid state machines (LSMs). The authors used spiking neural networks (SNNs) to identify spatiotemporal data and effectively identify normal, intermittent, and seizure EEGs. This method achieves an accuracy of 71.23% using a multispike LSM model.

Cui et al. [48] proposed an end-to-end regionally asymmetric CNN (RA-CNN) for emotion recognition consisting of temporal, regional, and asymmetric feature extractors. The authors used continuous 1D convolutional layers in the temporal feature extractor to learn the time-frequency representations. Subsequently, a regional feature extractor consisting of two 2D convolutional layers is used to capturethe regional information among the physically adjacent channels. Next, an asymmetric feature extractor captures discriminative information between the left and right hemispheres of the brain. This method achieved accuracies of 96.65% for valence and 97.11% for arousal in the DEAP dataset.

Yang et al. [49] proposed an end-to-end system for emotion recognition based on a parallel convolutional recurrent neural network (PCRNN). The results showed that The best accuracies achieved for the HV vs. LV and HA vs. LA classes were 90.8% and 91.03%, respectively, on the DEAP dataset.

Islam et al. [50] introduced a convolutional neural-network-based method. The accuracies obtained for classifying HV vs. LV and HA vs. LA classes on the DEAP dataset were 81.51% and 79.42%, respectively. Yang et al. [51] used a multi-column CNN (MC-CNN) method to obtain accuracies of 81.4% and 80.5% in classifying HV vs. LV and HA vs. LA classes, respectively.

Pandey et al. [46] used a DNN for emotion recognition and divided EEG signals into HV vs. LV and HA vs. LA classes. They obtained accuracies of 62.25% and 61.25% for the HV vs. LV and HA vs. LA classes, respectively. Yin et al. [55] proposed a method based on a graph CNN (GCNN) and LSTM. The accuracies obtained in classifying HV versus LV and HA versus LA classes were 90.45% and 90.60%, respectively.

Tan et al. [56] employed an SNN to model spatiotemporal EEG patterns. The accuracies obtained for classifying HV vs. LV and HA vs. LA classes were 67.76% and 78.97%, respectively, on the DEAP dataset.

Giuntini et al. [58], [59] assessed and modeled the temporal behavior of emotional and depressive user interactions on social networks. They modeled user interactions using complex networks and grouped them using the Clauset-Newman-Moore greedy modularity maximization. Furthermore, they combined the Empath framework and VADER lexicon to obtain the feature set for accessing the mood of the user using the standard deviation. The authors of separate research also proposed the use of tracing the roadmap of depressive users (TROAD) framework to recognize sequential patterns from social media users [59].

Zhang et al. [60] proposed a novel methodology for recognizing emotions using speech. The authors used autoencoders along with emotion embedding to extract the deep emotion features. Furthermore, they underwent data augmentation to further refine the results using the IEMOCAP and EMODB publicly available datasets.

Ahmed et al. [61] proposed a technique based on body movement for emotion recognition. They used a two-layer feature-selection process to recognize five emotions: happiness, sadness, fear, anger, and neutral. After the feature-selection process, score and rank-level fusion were applied to further improve the results. They obtained accuracies of 90.0% during walking, 96.0% during sitting, and 86.66% in an action-independent scenario.

Zhang et al. [62] proposed a deep autoencoder-based system for an EEG-based collaborative multimodal emotion recognition. The authors first used decision trees on the facial expression features, and the resultant vector was analyzed for facial expressions, followed by the bimodal deep automatic encoder (BDAE) for the fusion of EEG and facial expression signals, and then the LIBSVM classifier to perform the classification. Furthermore, the proposed approach achieved an average accuracy of 85.71%.

An overview of state-of-the-art methods indicates that most existing methods do not perform well in emotion recognition tasks. Methods based on handcrafted features do not learn discriminative information from EEG signals

and their performance depends on the tuning of various parameters. These techniques do not generalize well because hand-engineered features are typically not learned from the data under study and do not encode their structural patterns. Although DL-based methods show better performance because of their end-to-end learning approach, existing DL methods do not analyze information at the channel level. Motivated by this, we designed a deep ensemble model that decoded emotions using channel-level analysis.

III. DEEP-AUTOMATIC EMOTION RECOGNITION SYSTEM BASED ON DEEP LEARNING

The proposed deep automatic emotion recognition (Deep-AER) system was based on a two-level ensemble of deep LP-1D-CNN models. Fig. 1 shows the first level of an ensemble of non-overlapping windows of the EEG channel X_i that are passed to the LP-1D-CNN model M^i, and their predictions are fused to get the ith prediction. And Fig. 2 shows the second level of the ensemble containing predictions of all channels computed in the first level are fused for getting the final prediction of the Deep-AER system. The detailed architecture of the LP-ID-CNN model and its training and testing steps are described in the following subsections. We represented an EEG signal captured by C electrodes over a time interval with T timestamps $t_1, t_2, \ldots, t_T$ as a $C \times T$ matrix, as shown in Eq. (1):

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_c \end{bmatrix} = \begin{bmatrix} x_1(t_1) & x_1(t_2) \ldots\ldots & x_1(t_T) \\ x_2(t_1) & x_2(t_2) \ldots\ldots & x_2(t_T) \\ . & . & . \\ . & . & . \\ . & . & . \\ x_c(t_c) & x_c(t_2) \ldots\ldots & x_c(t_T) \end{bmatrix} \tag{1}$$

where $X_i = [x_i(t_1)\, x_i(t_2) \ldots x_i(t_T)] \in R^T$ is the i^{th} channel of $\boldsymbol{X}$ captured from the i^{th} electrode.

The first-level ensemble is designed to determine the state of each channel X_i. It consists of three main modules: (i) splitting the input channel X_i into K non-overlapping sub-signals using a window of fixed temporal length T_w: $X_i = \{S_1^i S_2^i \ldots S_k^i\}$ where $S_k^i = [x_i(t_{j_k})\, x_i(t_{j_k+1}) \ldots x_i(t_{j_k+T_w})] \in R^{T_w}$ and $k = 1, 2, \ldots, K$; (ii) classification of each sub-signal S_k^i with the LP-1D-CNN model M^i corresponding to channel X_i: $O_k^i = M^i(S_k^i)$ where $O_k^i \in \{0, 1\}$ is the predicted label of $S_k^i, k = 1, 2, \ldots, K$; and (iii) fusing the predictions of all sub-signals $S_k^i, k = 1, 2, \ldots, K$ using the majority vote: $O^i = majority\{O_1^i, O_2^i, \ldots, O_K^i\}$, where O^i is the predicted label of i^{th} channel X_i.

Using the predictions of all channels, the second-level ensemble predicts the final state of EEG signal $\boldsymbol{X}$ using majority vote fusion: $O = majority\{O^1, O^2, \ldots, O^c\}$, where O is the class label of EEG signal $\boldsymbol{X}$. The first level of the ensemble considers the dependencies among different segments of the same channel and the second level determines the dependencies between different channels.

A brain signal evoked by a particular task originates from a specific location in the brain. However, because of volume conduction, it is superimposed with other signals and captured from different brain locations [41]. This implies that different channels capture the signal evoked by a particular emotion in different quantities. Therefore, we trained a different CNN model for each channel to learn which part of the brain activity evoked by a particular emotion was captured by the channel. We hypothesized that the fusion of the predictions of the CNN models corresponding to different channels (i.e., the second-level ensemble) would predict the emotional state represented by the EEG signal. Furthermore, taking the complete channel corresponding to an emotional state raises specific issues: (i) the limited available data are not sufficient for a CNN model and (ii) the length of each channel is usually long (e.g., in DEAP, the length of each channel is 8,064). If the entire channel is used as input, the depth and, thus, the number of learnable parameters of the CNN model will increase and overfitting will be unavoidable. To overcome these difficulties, we segmented a channel into sub-signals by employing a fixed-size window. Using these signals, we trained a convolutional neural network (CNN) model for the channel. This approach solves these two problems. At the test time, the CNN model locally analyzes a channel, and fusion provides a global decision regarding the state of the channel. With these considerations, the designed LP-1D-CNN model for each channel has a very low complexity (i.e., only 8,462 learnable parameters), and the data generated by windowing each channel are sufficient for training and avoiding overfitting. In the following sections, we discuss our proposed LP-1D-CNN model and data augmentation, training, and testing schemes.

A. LIGHTWEIGHT PYRAMIDAL 1D CONVOLUTIONAL NEURAL NETWORK MODEL

We used a 1D-CNN to develop the Deep-AER system. The proposed architecture of the LP-1D-CNN model for the Deep-AER System., shown in Fig. 3, consists of an input layer, convolutional (CONV) blocks, and fully connected (FC) layers. The input layer takes the 1D channel of an EEG signal as input and passes it to a series of CONV blocks, which extract a hierarchy of features from the input signal. These features are passed to the first FC layer, which further processes them to extract discriminative information, and the second FC layer, which together with the softmax layer, predicts the class label of the input signal.

We used z-score normalization to normalize the input signals to unit variance and zero mean. This normalization helps avoid local minima and enables faster convergence. The normalized input is processed by four CONV blocks, where each block consists of three layers: a *Conv* layer, batch normalization layer (*bN*), and nonlinear activation layer (*Relu*). The number of kernels for *Conv-1* was 32; the receptive field of each kernel was 1×5; the number of kernels for *Conv-2* was 24; and the receptive field and depth of each kernel were 1×3 and 32, respectively. The number of kernels for *Conv-3* is 16, and the receptive field and depth of each

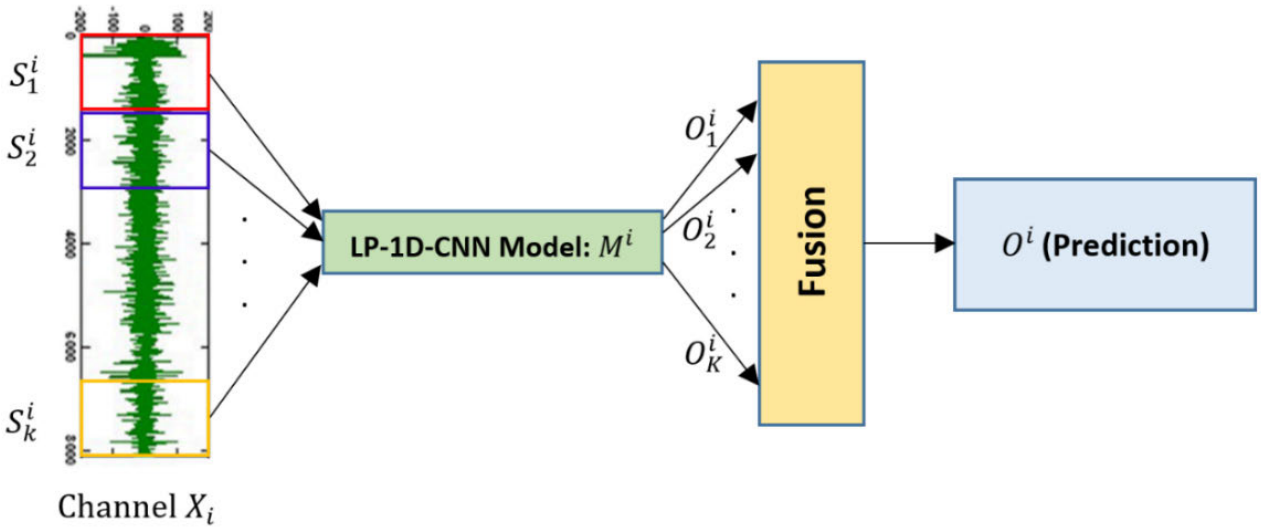

FIGURE 1. First level of ensemble: non-overlapping windows of the EEG channel X_i are passed to LP-1D-CNN model M^i, and their predictions are fused to get ith prediction.

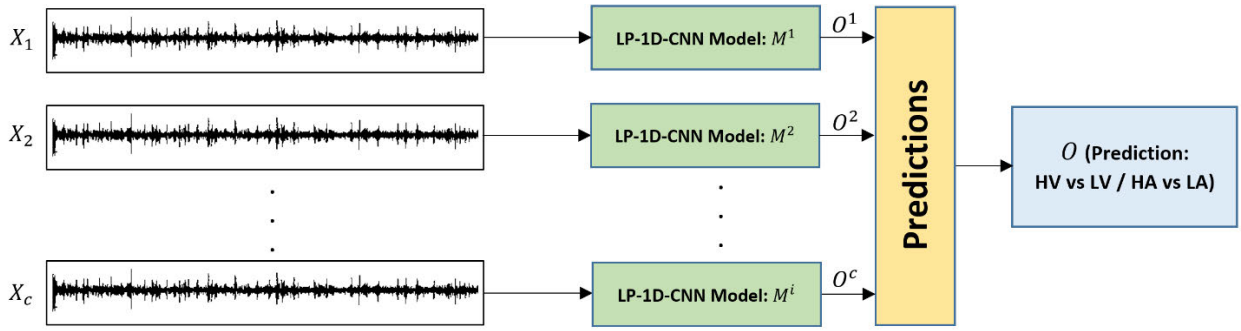

FIGURE 2. Second level of ensemble: Predictions of all channels computed in the first level are fused for getting the final prediction of Deep-AER system.

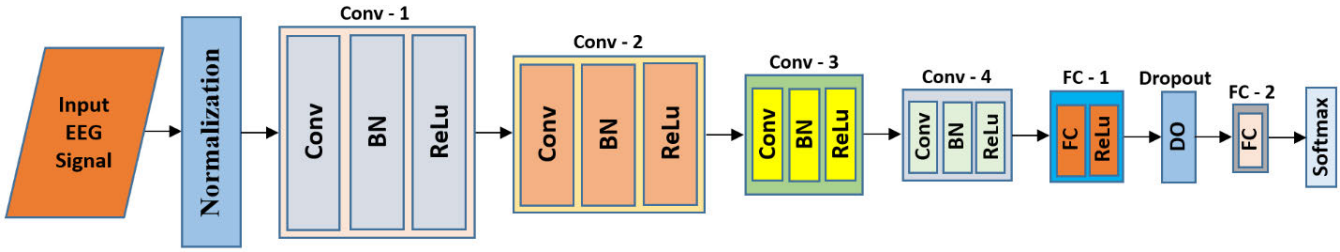

FIGURE 3. Proposed architecture of LP-1D-CNN model for Deep-AER System. It consists of four convolutional (Conv) blocks, and two FC layers.

kernel are 1 × 3 and 24, respectively. The number of kernels for *Conv-4* was eight, and the receptive field and depth of each kernel were 1 × 3 and 16, respectively. Unnecessary or redundant features are reduced using larger strides in the *Conv* layers; the strides are 3, 2, 2, and 2 in *Conv-1*, *Conv-2*, *Conv-3*, and *Conv-4*, respectively.

The output of the fourth block is passed to the first FC layer (*Fc*1), followed by the *Relu* layer, and another FC layer (*Fc*2). We examined two options for the number of neurons in *Fc*1:20 and 40. Dropout was used before *Fc*2 to avoid the risk of overfitting. The output of *Fc*2 is then passed to a softmax layer, which serves as a classifier and predicts the class of the input signal. There were two neurons in *Fc*2. The specifications of the model and its variants are presented in Table 1.

The proposed deep LP-1D-CNN model automatically learns EEG signal structures from the data and performs classification in an end-to-end manner. The proposed approach is opposite to the traditional hand-engineered approach, in which features are extracted first, and a subset of the extracted features is then selected and finally passed to a classifier for classification. The convolutional layer, which consists of a plane of many 1D channels or feature maps, is the primary component of the CNN model. This layer performs convolution by sliding the kernel over the input to obtain a convolved output (feature map).

TABLE 1. Specifications of four LP-1D-CNN models with Conv ($1 \times r$, /*str*, *nic*, *noc*) [where $1 \times r$ is the receptive field, *str* is stride, and *nic* and *noc* are number of input and output feature maps (channels)].

Layer	M_1^i	M_2^i	M_3^i	M_4^i
Conv - 1		(1 × 5, /3, 1, 32) *batchNorm* *Relu*		
Conv - 2		(1 × 3, /2, 32, 24) *batchNorm* *Relu*		
Conv - 3		(1 × 3, /2, 24, 16) *batchNorm* *Relu*		
Conv - 4		(1 × 5, /3, 16, 8) *batchNorm* *Relu*		
FC1	$Fc1 = 20$		$Fc1 = 40$	
Dropout	-	0.5	-	0.5
FC2		$Fc2 = 2$		
Classifier		Softmax		
No. of Parameters	8,462		12,842	

Let $\boldsymbol{a}_{l-1} \in R^{N_{l-1} \times d_{l-1}}$ be the activation of the $(l-1)^{th}$ layer, where d_{l-1} is the number of channels (feature maps) in the $(l-1)^{th}$ layer and N_{l-1} is the number of neurons in each channel. Let $w_l^i \in R^{r_l \times d_{l-1}}$ be the i^{th} kernel of l^{th} layer, where r_l is the receptive field of the kernel. The preactivation of the i^{th} channel of l^{th}*Conv* layer was calculated using Eq. (2):

$$c_l^i = w_l^i *_s \boldsymbol{a}_{l-1} + b_l^i \tag{2}$$

where b_l^i is the bias of the kernel w_l^i and $*_s$ is the convolution operation with stride s. Channel activation was computed using the ReLU nonlinear activation function, as shown in Eq. (3):

$$a_l^i = ReLU\left(c_l^i\right). \tag{3}$$

For the first *Conv* layer, the input is $S_k^i = \boldsymbol{a}_0 \in R^{N_0 \times 1}$, where N_0 is the number of sample points in the input 1D signal, S_k^i. In Fig. 3, the operation defined in (2) is represented as a *Conv* layer and the operation defined in (3) is represented as a ReLU layer.

After the *Conv* blocks, each model M^i had two *FC* layers. All neurons in the C*onv4* layer were connected to neurons in the first FC layer $FC1$. The activation of *Conv*4 is $a_4 \in R^{N_4 \times d_4}$ or $a_4 \in R^{N_4 d_4}$ after vectorization. Let $W_1 \in R^{N_4 d_4 \times N_5}$ be the weight matrix of *FC1* and $b_1 \in R^{N_5}$ be the bias vector of $FC1$. The pre-activation was computed using Eq. (4):

$$z_1 = W_1^T a_4 + b_1 \tag{4}$$

and its activation after applying the ReLU nonlinearity function is calculated using Eq. (5):

$$a_5 = ReLU\left(z_1\right). \tag{5}$$

The activation of $FC2$ is calculated in a similar manner. Note that the operations defined in (4) and (5) are represented as the $FC1$ and ReLU layers in Fig. 3.

The number of neurons in $FC1$ differ among the different models (see Table 1 for details). The second FC layer, $FC2$ has two neurons, HV versus LV or HA versus LA, and is a two-class problem. Furthermore, the outputs from the last FC layer were fed into the softmax function to predict the class probability of the input EEG channel X_i. Further details regarding the *Conv*, batch normalization, ReLU, and *FC* layers can be found in [24]. The 1D-CNN model analyzes a signal to learn a hierarchy of discriminative information and predict its class. In a CNN, the kernels are learned from data, unlike in the hand-engineered approach, in which the kernels are predefined (e.g., wavelet transform). With the novel idea of shared weights, a CNN has the advantage of significantly reducing the number of parameters.

Normally, a CNN model has a small number of kernels in low-level layers and a large number of kernels in high-level layers. However, the complexity of this type of structure is high, owing to a large number of learnable parameters. The size of the weight matrix W_1 in (4) depends on the number of neurons in the layer before the *FC1* layer. If the neurons in the *Conv*4 block are large, then the size of W_1 is large (i.e., it will cause a drastic increase in the number of learnable weights, leading to the problem of overfitting).

Instead, we used a pyramid architecture, where the number of kernels was large in the low-level layers and small in the higher-level layers. This architecture minimizes the risk of overfitting by significantly reducing the number of learnable parameters. A large number of kernels are used in the *Conv*1 layer, which is reduced by a constant number in the *Conv*2, *Conv*3, and *Conv*4 layers (e.g., models M^1 to M^4, specified in Table 1, contain *Conv*1, *Conv*2, *Conv*3, and *Conv*4 layers with 32, 24, 16, and 8 kernels, respectively).

The low-level layers extract a large number of microstructures composed of higher-level layers into higher-level features, which are small in number but discriminative. In other words, they implicitly perform feature reduction and

selection, which are essential parts of most methods based on hand-engineered features. In this study, we considered four models based on pyramid architecture to demonstrate the effectiveness of the LP-1D-CNN model. Table 1 lists the detailed specifications of these models, and provides the number of learnable parameters for each model. Using these models, we demonstrate how a properly designed model can result in better performance despite having fewer parameters, thereby reducing the risk of overfitting. Models with pyramid architectures have significantly fewer learnable parameters (Table 1).

B. DATA AUGMENTATION

In our approach, the problem of predicting the state of an EEG signal $\boldsymbol{X}$ is decomposed into smaller problems in predicting the classes of channels X_i of the signal. If the entire channel X_i is used as an input instance for the CNN model, it is difficult to train the CNN model for two reasons. First, the complexity of the model is very high because of the input size (e.g., the length of each X_i in DAEP is 8,064 sample points). Second, the available data were insufficient for training. To overcome this problem, X_i was divided into segments S_k^i of temporal length T_w, which were then passed to a CNN model to predict their states. The Xi class of X_i can be predicted by fusing their states. We used a window of T_w to create segment S_k^i. Consequently, we must train only one small CNN model for each X_i. The training instances of X_i were segmented using a window of T_w to create training instances to train the corresponding LP-1D-CNN model M^i to provide sufficient training instances to train the model. The available instances of X_i were divided into disjoint training and testing sets consisting of 90% and 10% of the total signals, respectively. Only the training set was used to create training data for M^i.

In the DEAP dataset, the total number of EEG signal instances was 1,280, for 32 subjects. We divided these instances into training and testing sets such that 90% (i.e., 1,152) were used for training and 10% (i.e., 128) for testing, enabling us to use 10-fold cross-validation for performance evaluation. As the number of each channel X_i in the training set of EEG signals was 1,152 and its length was 8,064 samples, the total number of instances of sub-signals S_k^i corresponding to X_i was 13,824 if $T_w = 5$. That is, 13,824 patterns are available to train the LP-ID-CNN model M^i, which is sufficient because it involves 8,462 learnable parameters (see Table 1).

C. TRAINING OF LIGHTWEIGHT PYRAMIDAL 1D CONVOLUTIONAL NEURAL NETWORK MODEL

Each LP-1D-CNN model M^i is trained using the training data created for the channel X_i (details given in the previous section). To train the model, we used a cross-entropy loss function, stochastic gradient descent with Adam (SGDA) as an optimizer [42], and backpropagation for gradient calculations. Using SGDA, the learnable parameters $\theta = (\mathrm{W,b})$ were updated using the following iterative procedure, as shown in Eq. (6):

$$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}, \tag{6}$$

where

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \tag{7}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}, \tag{8}$$

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t \tag{9}$$

$$v_t = \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g_t^2, \tag{10}$$

where m, v, t, and g_t are the first-moment vector, second-moment vector, time step, and gradient of the loss function, respectively, as described in Eq. (7)–(10) as follows: This algorithm has four hyperparameters: learning rate (α), β_1, β_2, and epsilon (ε), where β_1 and β_2 represent the exponential decay rates. The algorithm updates the exponential moving averages of the gradient (m_t) and the squared gradient (v_t), where the hyperparameters β_1, $\beta_2 \in [0, 1]$ control the exponential decay rates of these moving averages. The moving averages are estimates of the gradient's first moment (mean) and second raw moment (uncentered variance). However, these moving averages are initialized as (vectors of) zeros, leading to moment estimates that are biased toward zero, particularly during the initial time steps, t, especially when the decay rates are low. Following Kingma and Jimmy [42], we set $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\alpha = 1 \times 10^{-4}$, and $\varepsilon = 10^{-8}$. This enables the network to converge at a fast rate, thereby improving the efficiency of the training process. To improve generalization and avoid overfitting, the dropout technique was applied to $FC1$ using a probability value of 0.5.

D. TESTING

After training the model M^i corresponding to each channel X_i, an unknown or test EEG signal $\boldsymbol{X}$ is classified using a two-level ensemble of deep LP-1D-CNN models. The architectures of the first- and second-level ensembles are shown in Figs. 1 and 2, respectively. The first-level ensemble was designed to decide the state of each channel X_i. First, the trained LP-1D-CNN model M^i is used to predict the label of the channel X_i by classifying each of its sub-signals S_k^i and fusing their predictions. The temporal length of the channel X_i in the DEAP was 1 min and consisted of 8,064 samples. It is divided into K non-overlapping sub-signals S_k^i using a window of fixed temporal length T_w, which is used to create training patterns for learning M^i. These sub-signals S_k^i are treated as independent signal instances and passed to the LP-1D-CNN model M^i, which predicts its class label (that is, $O_k^i = M^i(S_k^i)$ where $O_k^i \in \{0, 1\}$ is the predicted label of $S_k^i, k = 1, 2, \ldots, K$. The class label of X_i is predicted by fusing the predictions of all sub-signals S_k^i using the majority vote: $O^i = majority\{O_1^i, O_2^i, \ldots, O_K^i\}$, where O^i is the predicted label of the channel X_i. Using the predictions of all channels obtained in the first-level ensemble, the second-level ensemble predicts the final state of EEG signal $\boldsymbol{X}$

using majority vote fusion: $O = majority\left\{O^1, O^2, \ldots, O^c\right\}$, where O is the class label of the EEG Signal $\boldsymbol{X}$.

IV. DATASET, EXPERIMENTAL PROTOCOL, AND EVALUATION CRITERIA

In this section, we present details of the dataset used for the experiments, experimental protocols, and evaluation criteria.

A. DATASET

DEAP is a benchmark EEG database for the analysis of spontaneous emotions prepared by the Queen Mary University of London [15]. It was generated to build an adaptive music video recommendation system based on user emotions. The database was recorded by using music clips to evoke emotions. The database consists of physiological signals of 32 participants (16 men, 16 women; age range:19 to 37 years, mean age:26.9 years) recorded while watching 40 one-minute music videos. The dataset contains 32 channel EEG signals and eight peripheral physiological signals. The details of the EEG signal recordings, pre-processing, and stimulus materials can be found in [15]. At the end of each music clip, participants assessed their emotional states in terms of valence, arousal, dominance, and liking. A self-assessment manikin (SAM) was used to visualize valence and arousal scales. Participants rated valence and arousal on a continuous nine-point scale. The valence scale ranged from unhappy to happy or joyful. Participants with valence ratings below five were assumed to have negative emotions, whereas those who did not were considered to have positive emotions. Similarly, the arousal scale ranges from inactive or passive to active. Participants with arousal ratings below five were considered inactive, whereas the others were assumed to be active. The sample EEG signals related to the HV and LV states measured from the Fp1 channel are shown in Fig. 4.

In this study, we considered two dimensions (valence and arousal) and 32-channel EEG signals, and treated the problem of emotion recognition as two classification problems: HV vs. LV and HA vs. LA. Valence scales from 1 to 5 (excluding 5) were mapped to the LV and 5 to 9 were mapped tothe HV. Similarly, the arousal scale from 1 to 5 (excluding 5) was mapped to the LA, and 5 to 9 were mapped tothe HA. In addition, we considered the following regions (se Fig. 5 for specification of brain regions) to identify their roles in emotion recognitio:

i. *FRONT* : Frontal-right (FR) and frontal-left (FL):12 channels
ii. *CENT* : Central-right (CR) and central-left (CL); four channels.
iii. *PERI* : parietal right (PR) and parietal left (PL); six channels
iv. *OCCIP* : *O*ccipital − right (OR) occipital-left (OL); four channels
v. *ALL* : All regions (AR): 32 channels.

B. EVALUATION CRITERIA

To evaluate the performance of the proposed system, we used a 10-fold cross-validation to test the system over different data variations. The signals for each class were divided into 10 folds: one fold (10%) was used for testing, whereas the remaining nine (90%) were used to train the model. The average performance was then computed ten times. Well-known metrics were used to evaluate performance: accuracy, sensitivity, specificity, geometric mean (g-mean), F-measure, and precision. The definitions of these metrics are given in Eq. (11)–(16), respectively:

$$Accuracy(Acc) = \frac{TP + TN}{Total\ Samples} \tag{11}$$

$$Specificity\,(Spec) = \frac{TN}{TN + FP} \tag{12}$$

$$Sensitivity\,(Sens) = \frac{TP}{FN + TP} \tag{13}$$

$$Precision\,(Prec) = \frac{TP}{TP + FP} \tag{14}$$

$$F - Measure\,(FM)) = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} \tag{15}$$

$$G - Mean\,(GM) = \sqrt{Specificity * Sensitivity} \tag{16}$$

where *TP is* true positives, or the number of LV (LA) that were identified as LV (LA); *FN is* false negatives, or the number of LV (LA) predicted as HV (HA); *TN is* true negatives, or the number of HV (HA) identified as HV (HA); and *FP is* false positives, or the number of HV (HA) predicted as LV (LA).

TensorFlow, a freely available DL library from Google, was used to implement the LP-1D-CNN model in Python [43]. We trained the LP-1D-CNN model M^i on a system with an Intel®Xeon®processor E5-2670 v2 (2.5 ∼ GHz, 20 CPUs), 32 GB RAM, and an NVIDIA® Quadro®K4000 3 GB graphics card.

V. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents and discusses the results of the two emotion classification problems (HV vs. LV and HA vs. LA). We analyzed the potential of the proposed Deep-AER system for emotion recognition in five brain regions: *the FRONT*, *CENT*, *PERI*, *OCCIP*, and *ALL*. The best LP-1D-CNN model was selected by analyzing the results. A 10-fold cross-validation technique was used to perform all the experiments. In this section, we present comparisons with state-of-the-art studies and discuss directions for future research. In this study, we tested three different sizes of T_w:5, 10, and 15 s. Using these sizes for T_w, we divided each channel of 8,064 samples into 12 (672 samples each), six (1,344 samples each), and four (2,016 samples each) sub-signals S_k^i, respectively. Of the three choices, $T_w = 5$ s yielded the best result (98.43%) for the HV vs. LV case in the *FRONT* brain region. The system achieved accuracies of 96.8% and 93.7% when using 10- and 15-second window sizes, respectively.

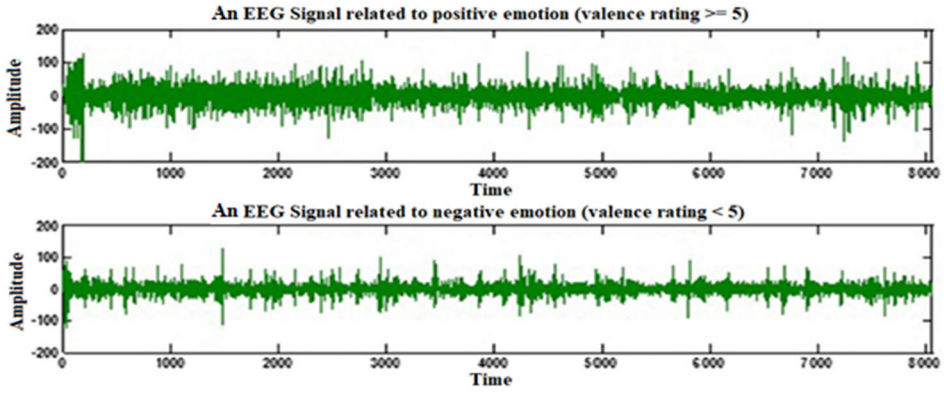

FIGURE 4. Two samples of EEG signals related to high and low valence, captured from channel Fp1.

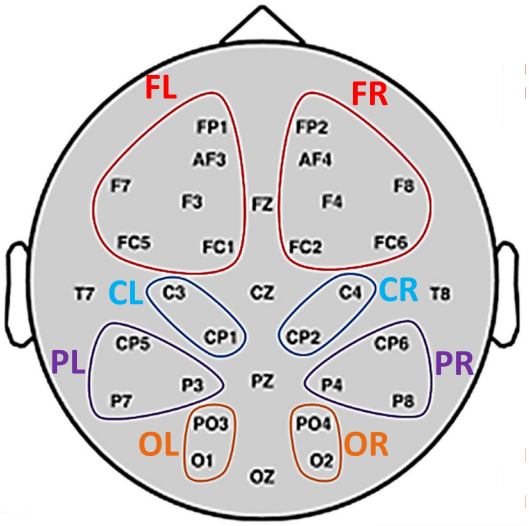

FIGURE 5. Specification of brain regions - ***ALL***: all regions, representing 32 channels; ***FRONT***: frontal right (FR) and frontal left (FL); ***CENT***: central right (CR) and central left (CL); ***PERI***: parietal right (PR) and parietal left (PL); ***OCCIP***: occipital right (OR) and occipital left (OL).

This shows that each small window of 5 s in the ensemble contained more relevant information, and the system analyzed a local part of the signal minutely. Therefore, in all other experiments using different brain regions (i.e., *CENT*, *PERI*, *OCCIP*, *and ALL*), we used a window size of 5 s.

A. SPECIFICATIONS OF LIGHTWEIGHT PYRAMIDAL 1D CONVOLUTIONAL NEURAL NETWORK MODEL MODELS

To determine the best model M^i for each channel X_i, we considered four models M_j^i, $j = 1, 2, 3, 4$, as shown in Table 1, and performed experiments on the five brain regions. We performed all experiments using 10-fold cross-validation with all four models for two problems: HV vs. LV and HA vs. LA. These experiments led us to select the best LP-1D-CNN model for further analysis. The DEAP dataset is used to train and test the models. Models M_1^i through M_4^i (pyramid models) were designed by reducing the number of kernels or filters w_l^i by 25% as the network deepened. Pyramid models contain fewer learnable parameters than traditional models; therefore, they generalize well and are less prone to overfitting.

B. ANALYSIS OF BRAIN REGIONS

To analyze the brain region in which our system was the most effective for emotion recognition, we performed five experiments considering EEG signals captured from four different brain regions and the whole brain.

1) EXPERIMENT 1: FRONT REGION

There are 12 channels X_i in EEG signal $\boldsymbol{X}$ recorded from this region, as shown in Fig. 5. The average performance results obtained with the four different models (M_1^i to M_4^i) in the *FRONT* region are listed in Table 2 . Among the four models, when M_2^i was used in the Deep-AER system, it yielded the best mean accuracy of 98.43% for the HV vs. LV problem, with a mean specificity and sensitivity of 97.8% and 98.7%, respectively. Similarly, for the problem of HA versus LA, the same model resulted in the best mean accuracy of 97.65%, with a mean specificity and sensitivity of 97.9% and 97.5%, respectively. The other performance measures for this model were better than those of other models.

The mean accuracies of the four models when sub-signals S_k^i of channel X_i, (i = 1,2,..., 12) of the FRONT region were used as training and testing instances, as shown in Figs. 6 and 9 for HV vs. LV and HA vs. LA, respectively. Further analysis of the mean training and testing accuracies of the four models on each channel of the *FRONT* region (i.e., first-level ensemble) and all channels of the region (i.e., second-level ensemble) is shown in Figs. 7 and 8 for HV vs. LV and Figs. 10 and 11 for HA vs. LA. These results have several implications. First, the model M_2^i outperformed the other models in all cases. This is likely because it uses dropout and fewer neurons in *FC1*, which implicitly performs feature selection. Second, none of the models suffered from overfitting and the differences between the mean training and testing accuracies were small. Third, the ensemble enhances the model's performance when the decision is made using only the sub-signals S_k^i. The mean accuracy was less than that when the first ensemble was used, and the second-level

TABLE 2. Comparison of different LP-1D-CNN models over *FRON* region.

	Model	M_1^i	M_2^i	M_3^i	M_4^i
HV vs. LV	*Acc*	97.65±0.66	**98.43±0.60**	96.87±0.63	97.65±0.64
	Sens	0.975	**0.987**	0.975	0.975
	Spec	0.956	**0.978**	0.956	0.978
	Prec	0.975	**0.988**	0.975	0.987
	GM	0.965	**0.982**	0.965	0.976
	FM	0.975	**0.987**	0.975	0.981
HA vs. LA	*Acc*	96.87±0.68	**97.65±0.63**	96.1±0.66	96.87±0.70
	Sens	0.962	**0.975**	0.975	0.974
	Spec	0.959	**0.979**	0.94	0.959
	Prec	0.974	**0.987**	0.963	0.974
	GM	0.96	**0.976**	0.957	0.966
	FM	0.968	**0.981**	0.969	0.974

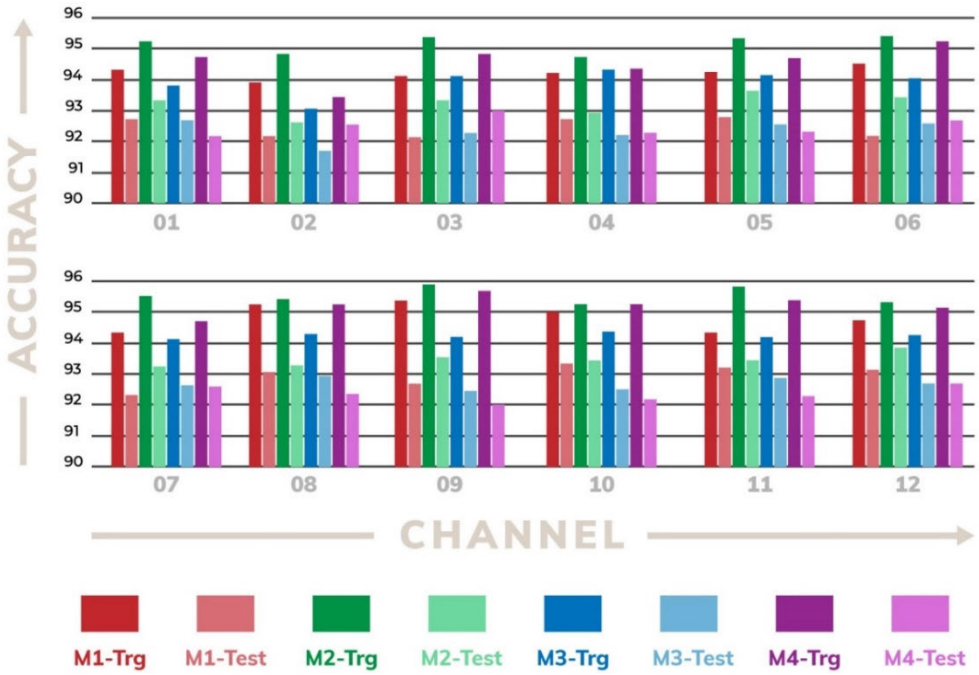

FIGURE 6. Single model channel-wise accuracies for HV vs. LV with models M_1^i through M_4^i on the FRONT region using the DEAP dataset.

ensemble provided better mean accuracy than the first-level ensemble, as shown in Figs. 6, 7, 9, and 10.

Based on the overall results shown in Table 2 and Figs. 6–11 and the above discussion, we conclude that model M_2^i with a dropout layer and 20 neurons in the $FC1$ layer gives the best results when only the EEG signals from the *FRONT* region are considered. This model has a similar or slightly higher performance than the other models, but less complexity (i.e., it involves fewer learnable parameters than M_3^i and M_4^i). Further, to provide insight into the performance of this model, the 10-fold cross-validation results for the *FRONT* region are shown in Table 3; the standard deviation is 0.60 for the HV vs. LV problem and 0.63 for the HA vs. LA problem, respectively, indicating the robustness of the model. Given the superior performance of M_2^i, we used only this model in all experiments.

2) EXPERIMENT 2: CENTRAL REGION

EEG signal $\boldsymbol{X}$ recorded from this region has four channels,as shown in Fig. 5. In this case, when M_2^i is used in the Deep-AER system, it provides a mean accuracy of 92.3% (see Table 3) for the problem of HV vs. LV and a mean specificity and sensitivity of 91.8% and 92.7%, respectively. The mean *Prec*, *GM*, and *FM* are 92.7%, 91.9%, and 93.2%, respectively. For HA versus LA, M_2^i yields a mean accuracy of 93.8%, mean specificity of 94.5%, and mean sensitivity of 93.2%. The mean *Prec*, *GM*, and *FM* are 94.2%, 93.3%, and 94.1%, respectively.

3) EXPERIMENT 3: PARIETAL REGION

There are six channels X_i in EEG signal $\boldsymbol{X}$ recorded from this brain region, as shown in Fig. 5. The accuracies obtained

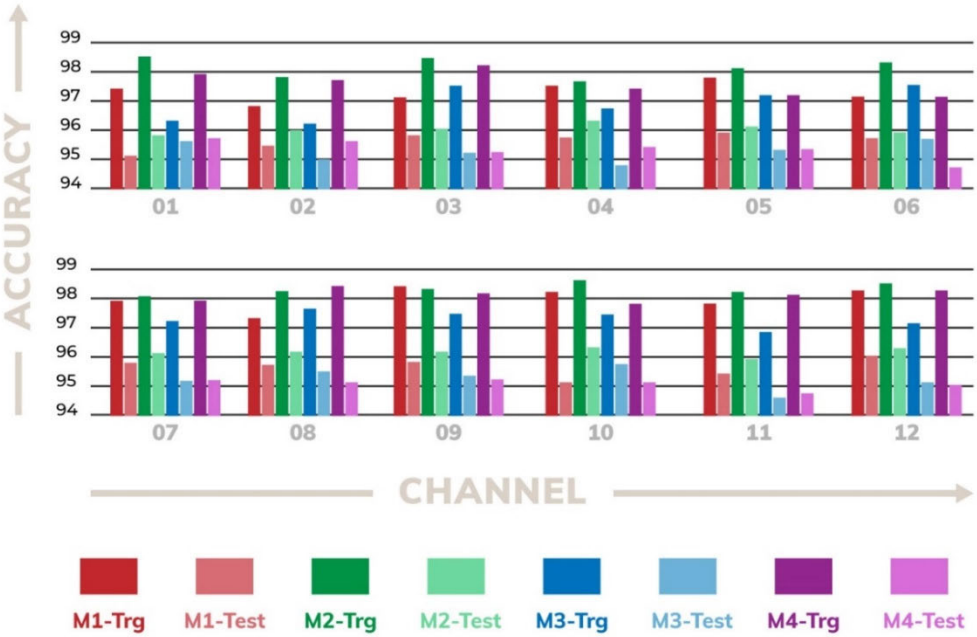

FIGURE 7. First-level ensemble channel-wise accuracies for HV vs. LV with models M_1^i through M_4^i on the FRONT region using the DEAP dataset.

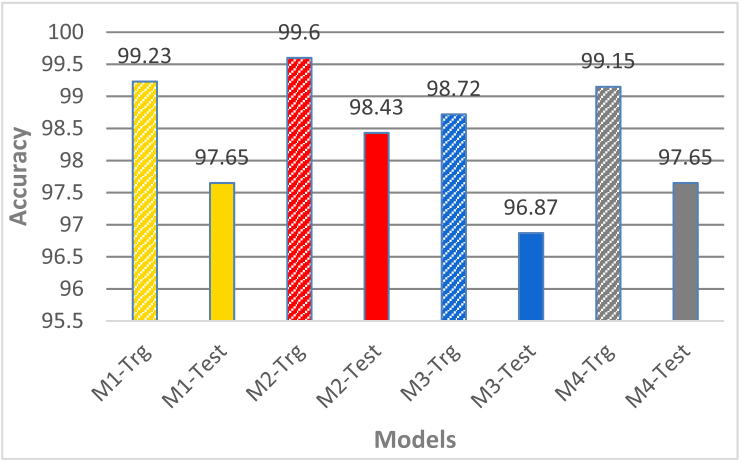

FIGURE 8. Second-level ensemble accuracies of HV vs. LV with models M_1^i through M_4^i on the FRONT region using the DEAP dataset.

using M_2^i on the *PERI* region are listed in Table 3. M_2^i provided a mean accuracy of 94.6% for the problem of HV vs. LV; the mean specificity and sensitivity were 93.8% and 95.7%, respectively. The mean *Prec*, *GM*, and *FM* are 95.1%, 94.2%, and 94.9%, respectively. In the case of HA vs. LA, M_2^i yielded a mean accuracy of 93.2% and mean specificity and sensitivity of 94.1% and 92.4%, respectively. The mean *Prec*, *GM*, and *FM* values were 93.7, 92.6, and 93.5%, respectively.

4) EXPERIMENT 4: OCCIPITAL REGION

EEG signal $\boldsymbol{X}$ recorded from this region has four channels, as shown in Fig. 5. The accuracies obtained using M_2^i on this region are presented in Table 3 . The model yielded a mean accuracy of 91.4% for HV vs. LV and a mean specificity and sensitivity of 90.8% and 92.7%, respectively. The mean *Prec*, *GM*, and *FM* are 91.8%, 90.6%, and 91.7%, respectively. For the HA vs. LA problem, M_2^i resulted in a mean accuracy of 92.7% and mean specificity and sensitivity of 93.4% and 92.1%, respectively. The mean *Prec*, *GM*, and *FM* are 93.6%, 91.4%, and 93.2%, respectively.

5) EXPERIMENT 5: ALL REGIONS

In this experiment, EEG signal $\boldsymbol{X}$ consisted of 32 channels covering all brain regions, as shown in Fig. 5. The accuracies obtained using M_2^i are presented in Table 3 . When M_2^i was used in the Deep-AER system, it yielded a mean accuracy of 91.7% for the problem of HV vs. LV, and a mean specificity and sensitivity of 90.4% and 92.8%, respectively. The mean *Prec*, *GM*, and *FM* are 92.6%, 91.1%, and 92.2%,

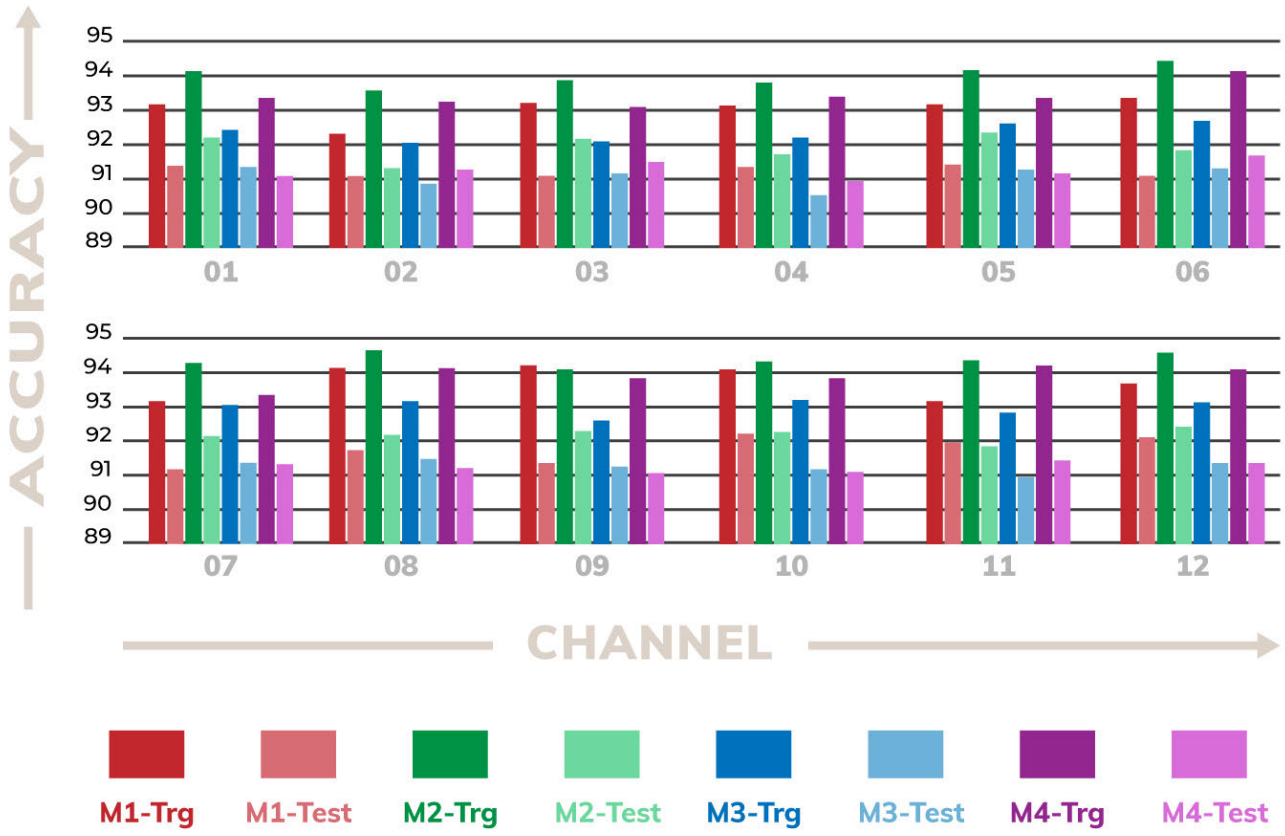

FIGURE 9. Single model channel-wise accuracies for HA vs. LA with models M_1^i through M_4^i on the FRONT region using the DEAP dataset.

TABLE 3. 10-fold cross-validation accuracies (%) of the Deep-AER system on different brain regions using model M_2^i for two-class problems.

	HV vs. LV					HA vs. LA				
Fold	*FRONT*	*CENT*	*PERI*	*OCCIP*	*ALL*	*FRONT*	*CENT*	*PERI*	*OCCIP*	*ALL*
*K*1	98.43	92.9	95.3	91.4	89.8	96.8	94.5	93.7	93.7	91.4
*K*2	99.2	91.4	93.7	92.1	93.7	98.43	92.9	92.9	91.4	90.6
*K*3	98.43	92.1	95.3	91.4	92.1	97.65	93.7	93.7	93.7	89.8
*K*4	97.65	92.9	94.5	90.6	89.8	97.65	94.5	92.9	92.1	90.6
*K*5	98.43	92.1	95.3	92.1	92.1	98.43	93.7	93.7	91.4	89.8
*K*6	99.2	93.7	94.5	91.4	91.4	96.8	94.5	92.1	93.7	90.6
*K*7	97.65	90.6	93.7	92.1	90.6	97.65	93.7	93.7	91.4	89.1
*K*8	99.2	91.4	94.5	90.6	93.7	98.43	92.9	91.4	92.9	89.8
*K*9	98.43	93.7	95.3	92.1	92.1	97.65	94.5	94.5	93.7	90.6
*K*10	97.65	92.1	94.4	90.6	91.4	96.8	93.7	93.7	92.9	91.4
Mean	**98.43**	92.3	94.6	91.4	91.7	**97.65**	93.8	93.2	92.7	90.3
SD	**0.60**	0.96	0.60	0.62	1.30	**0.63**	0.59	0.86	0.97	0.70

respectively. In the case of HA vs. LA, M_2^i yielded a mean accuracy of 90.3% and mean specificity and sensitivity of 91.8% and 90.1%, respectively. The mean *Prec*, *GM*, and *FM* are 91.2%, 89.6%, and 90.9%, respectively.

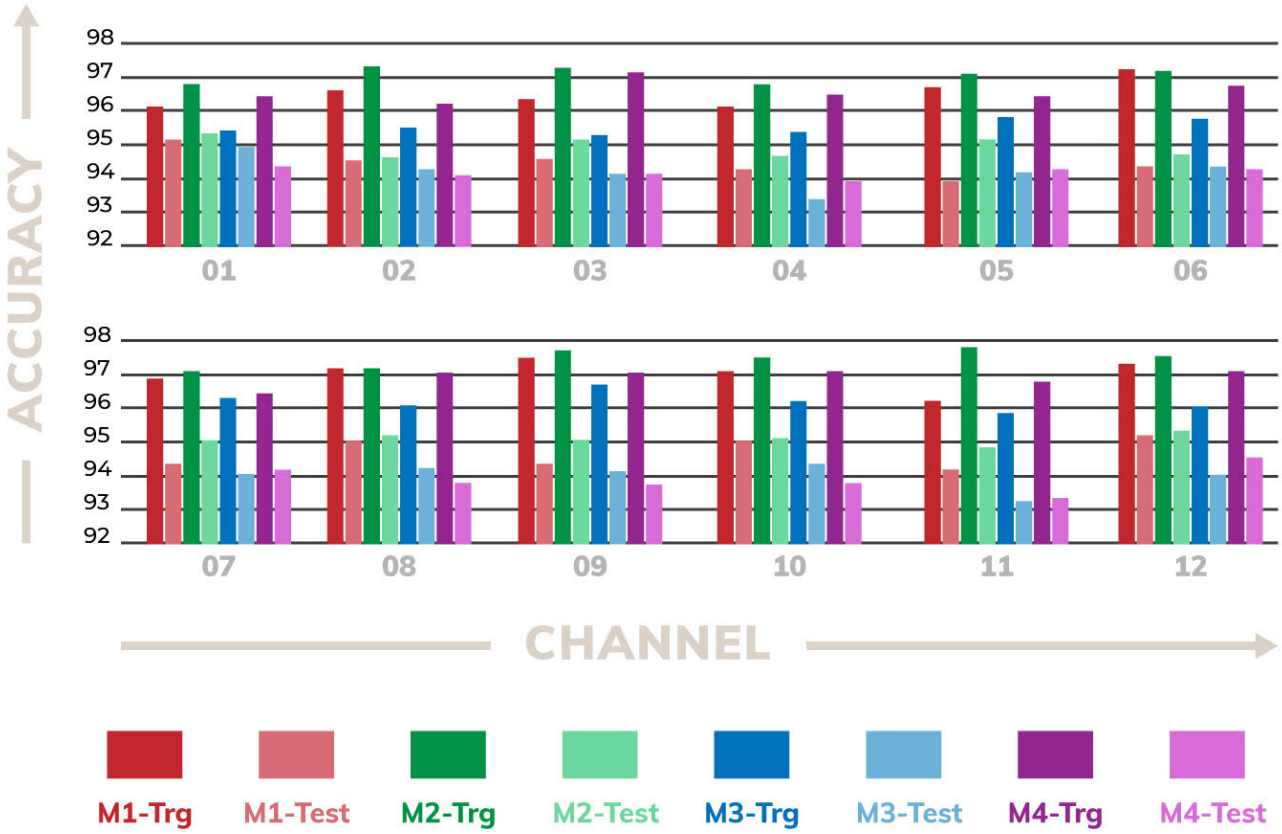

FIGURE 10. First-level ensemble channel-wise accuracies for HA vs. LA with models M_1^i through M_4^i on the FRONT region using the DEAP dataset.

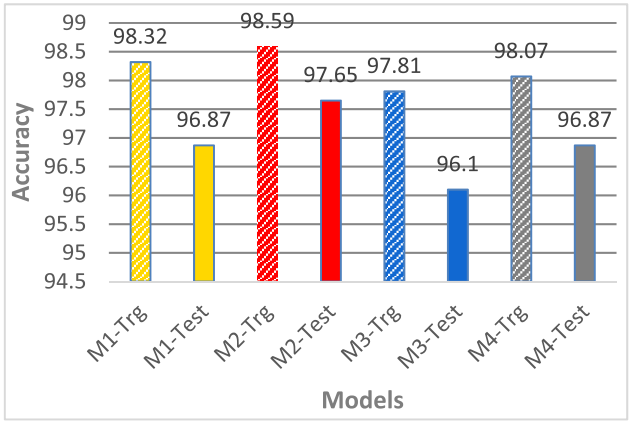

FIGURE 11. Second-level ensemble accuracies for HV vs. LV with models M_1^i through M_4^i on the FRONT region using the DEAP dataset.

To assess the performance of the Deep-AER system, we conducted five experiments corresponding to the five brain regions, as shown in Fig. 5. The numbers of channels in the EEG signals captured from *FRONT, CENT, PERI, OCCIP,* and *ALL* were 12, 4, 6, 4, and 32, respectively. In each experiment, after training the LP-1D-CNN models M_2^i for each channel X_i of the corresponding region, we designed a Deep-AER system as a two-level ensemble, employing a majority vote strategy to fuse the local decisions for the HV versus LV and HA versus LA problems. Different regions lead to different results, as shown in Table 3 . The Deep-AER system provided the best performance in the *FRONT* region with model M_2^i; the accuracies for all other regions were below 95%. A comparison of different brain regions with regard to accuracy is shown in Fig. 12. The results indicate that the *FRONT* region plays a dominant role in emotion recognition, with accuracies of 98.43% and 97.65% for HV versus LV and HA versus LA, respectively]. Detailed results for the *FRONT* region with model M_2^i are shown in Table 3 and Figs. 6 through 11. The 10-fold cross-validation results in Table 3 show that the standard deviations in the case of *FRONT* for the two problems were 0.60 and 0.63. In contrast, the standard deviations for the other regions were higher, with the exception of *CENT* for the HA versus LA problems. This indicates that Deep-AER's emotion recognition performance is robust for the *FRONT* brain region; in other words, the system provides very similar results across variations of the training and testing datasets.

The channel-wise and first-level ensemble results are depicted in Figs. 6, 7, 9, and 10 indicate that the ensemble performed better than the single model. This is because, in the ensemble, a model simulates experts analyzing local parts of the signal and fusing their local decisions using the majority vote to make the final decision. Thus, the first-level ensemble combines local decisions with the global context and outperforms a single model. Furthermore, Figs. 7, 8, 10, and 11 show that the second-level ensemble outperformed the

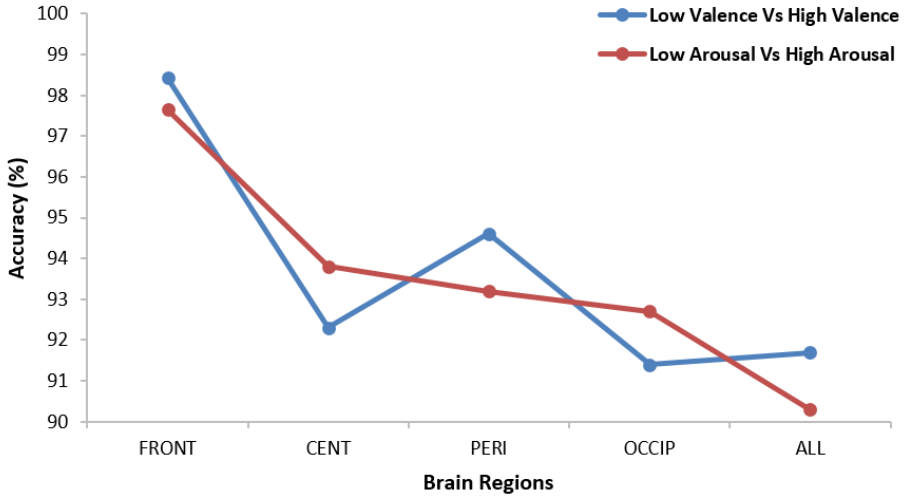

FIGURE 12. Performance comparison of brain regions *(FRONT, CENT, PERI, OCCIP, and ALL)* w.r.t accuracy.

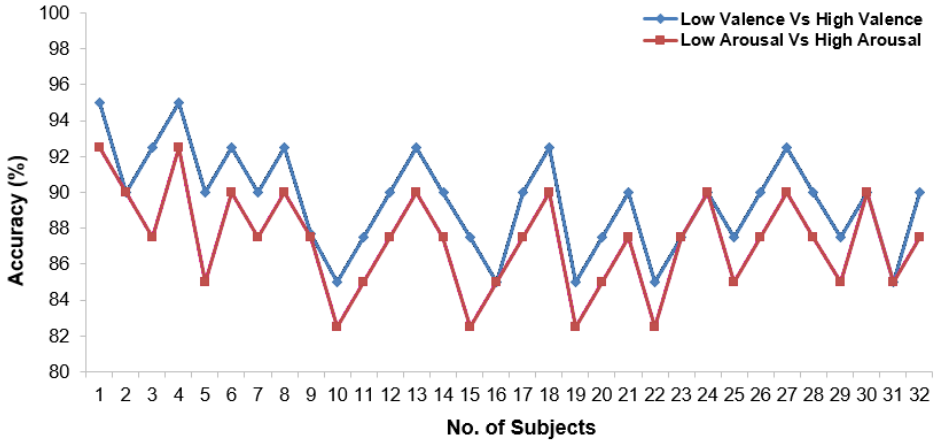

FIGURE 13. Experimental results for HV vs. LV and HA vs. LA w.r.t accuracy for independent subjects using the DEAP dataset.

first-level ensemble because the second-level ensemble fused local decisions based on individual channels with a global context defined by all channels in a specific brain region. Moreover, the Deep-AER system is based on end-to-end LP-1D-CNN models (i.e., each model takes an input signal and provides a decision). Thus, there is no need to perform signal pre-processing, manual feature extraction, or laborious parameter selection and tuning. Rather, the system learns discriminative information from the data in a fully automatic process.

Notably, our design of the LP-1D-CNN model requires minimal memory space, and its architecture of the LP-1D-CNN model is based on a pyramid design that uses the least possible number of learnable parameters. The best pyramid-based LP-1D-CNN model, M_2^i contained 8,462 parameters. A small number of learnable parameters implies a less complex model, which results in less memory overhead and ensures improved generalization. This implies that the proposed Deep-AER system does not heavily depend on data, is robust, and generalizes better than state-of-the-art methods.

TABLE 4. Confusion matrix for HV vs. LV problem on FRONT region using model M_2^i.

		Predicted Class	
		Low	High
Actual Class	Low	81 (98.78%)	1 (1.22%)
	High	1 (2.17%)	45 (97.83%)

The mean accuracies of the Deep-AER system in the *FRONT* region with model M_2^i were 98.43% and 97.65% for HV vs. LV and HA vs. LA, respectively, which validates the generalization power of the proposed system. Tables 4 and 5 show the confusion matrices forHV vs. LV and HA vs. LA problems in the*FRONT* region.

C. PERFORMANCE USING DIFFERENT SUBJECTS FOR TRAINING AND TESTING DATA

In our evaluation, we used all trials from all subjects and divided them into training and testing data using 10-fold

TABLE 5. Confusion matrix for HA vs. LA problem on FRONT region using model M_2^i.

		Predicted Class	
		Low	High
Actual Class	Low	77 (97.47%)	2 (2.53%)
	High	1 (2.04%)	48 (97.96%)

cross-validation to test the performance of the system over different data variations. To test the system's generalizability, we performed experiments using testing data from subjects whose data were not used for training (i.e., we used training and testing data from different subjects).

For this purpose, we investigated the subject-independent EEG emotion recognition problem, in which training and testing EEG data samples were obtained from different subjects. To this end, we adopted a leave-one-subject-out cross-validation strategy in conducting the experiment, in which we circularly took one subject's EEG signals as the testing data and the EEG signals of all other subjects as the training data. The average result of all the recognition accuracies was then calculated after each subject was used once as the testing data. The experimental results of the proposed method are shown in Table 6, from which we can observe that Deep-AER achieves accuracy levels of 89.46% and 87.27% for HV versus LV, and HA versus LA, respectively. This experiment shows that our model captures discriminative features despite individual differences, indicating the generalization power of the Deep-AER system. This also indicates that the performance of the Deep-AER system is stable for both the dependent and independent subjects. Table 6 shows the average accuracy of the *FRONT* brain region using the M_2^i. The experimental results for HV vs. LV and HA vs. LA with regard to accuracy for independent subjects are shown in Fig. 13. Well-known metrics were used to evaluate performance: accuracy, sensitivity, specificity, g-mean, F-measure, and precision.

The accuracy varies among subjects; it is due to the reason that different subjects give widely differing responses to stimuli (internalized and externalized). One possible direction to overcome the subjectivity issue is to analyze different frequency bands and employ frequency bands that encode emotional states and are invariant across subjects.

D. ANALYSIS OF DOMINANT BRAIN REGION IN DEEP-AER SYSTEM

In previous studies [44], [45], researchers have highlighted the importance and association of the *FRONT* brain region with emotion. In this study, we also observed the dominance of the *FRONT* region over the *CENT*, *PERI*, *OCCIP*, and *ALL* brain regions for HV vs. LV and HA vs. LA classification problems. Our findings validate previous research regarding the involvement of the *FRONT* region in positive and negative emotions [44], [45] and achieved the best accuracy rates on the *FRONT* region for both the HV vs. LV and HA vs. LA problems. All other brain regions (i.e., *CENT*, *PERI*, *OCCIP*, and *ALL*) yielded accuracies below 95% for HV versus LV and HA versus LA problems. The question arises as to whether the channels of EEG signals captured from the *FRONT* region are correlated. Table 7 shows the Pearson correlation coefficients (r) between the channels in the *FRONT* region for the HV versus LV problem. A few of these correlation coefficients showed a moderately positive or negative correlation, but most were in the range of weak and negligible positive or negative correlation. There were no strong or very strong correlation coefficients between the channels over the *FRONT* region when using model M_2^i. The maximum correlation between the channels was negligible to weak. This indicates that all channels in the *FRONT* region have discriminative information and must be considered in the design of the Deep-AER system.

E. DOES THE MODEL OVERFIT?

In this study, we segmented each channel into small windows and trained one LP-1D-CNN model on these windows. As the size of each input window is small, the complexity of the LP-1D-CNN model is low, and it is robust against overfitting. Moreover, the best pyramid-based LP-1D-CNN model, $M2M_2$ contains 8,462 parameters. A small number of learnable parameters implies a less complex model, which results in less memory overhead and ensures improved generalization. The results shown on ten different folds and over different regions in Table 3 are consistent despite the different training and test sets. This indicated that the model did not overfit.

F. COMPARISONS

To evaluate the effectiveness of the proposed Deep-AER system, we compared our experimental results with those of previous emotion recognition studies that used EEG signals. Previous methods, as shown in Table 8, have used hand-engineered features such as PSD, power asymmetry, band power, statistical features, QTFD- and EMD-based features, fractal dimension, Hjorth parameters, wavelet statistical features, EEG spectral power, and wavelet entropy for HV versus LV and HA versus LA problems. From Table 8, it is clear that the proposed Deep-AER system has significantly better classification performance than other state-of-the-art approaches.

The Deep-AER system outperformed existing methods for three reasons. First, it is based on a DL approach, which has shown outstanding performance in many problems compared with approaches using hand-engineered features [19], [20], [21], [22]. Secondly, it employs pyramid architectures for the design of CNN models, which are less complex and do not require large amounts of data for learning. Third, unlike other methods based on deep learning, it uses an ensemble strategy that combines local decisions with global context [46], [48], [49], [50], [51], [55], [56].

To train the LP-1D-CNN model, we used 13,824 EEG signals, each consisting of 672 samples; one epoch (training and validation) took 4.12 s. Each model was trained on

TABLE 6. Results over *FRONT* region using M_2^i for subject-independent experiment.

HV vs. LV	***Acc***	**89.46±2.81**
	Sens	89.8
	Spec	88.7
	Prec	89.9
	GM	89.4
	FM	89.8
HA vs. LA	*Acc*	**87.27±2.79**
	Sens	87.2
	Spec	87.6
	Prec	88.1
	GM	87.3
	FM	88.4

TABLE 7. Pearson correlation coefficients (r) between channels for HV vs. LV over FRONT region using model M_2^i.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}
C_1	1.0000	-0.0192	-0.0562	0.1272	0.1877	-0.0238	-0.0780	-0.1905	-0.2614	-0.3589	-0.0185	0.1165
C_2	-0.0192	1.0000	-0.2524	-0.0187	-0.0650	0.2223	0.0578	0.0877	0.1995	0.3217	0.0467	0.3609
C_3	-0.0562	-0.2524	1.0000	-0.1931	-0.1407	-0.0268	0.1307	0.0040	0.0990	0.1861	-0.0825	0.2768
C_4	0.1272	-0.0187	-0.1931	1.0000	-0.1670	-0.0973	-0.1024	0.0452	0.1458	-0.2279	-0.1670	-0.1383
C_5	0.1877	-0.0650	-0.1407	-0.1670	1.0000	-0.0142	0.2765	-0.1571	-0.2156	-0.0056	0.0032	0.2417
C_6	-0.0238	0.2223	-0.0268	-0.0973	-0.0142	1.0000	-0.1408	-0.1079	-0.0197	0.1154	0.0126	0.1887
C_7	-0.0780	0.0578	0.1307	-0.1024	0.2765	-0.1408	1.0000	-0.2224	-0.1141	0.0442	0.1300	-0.3138
C_8	-0.1905	0.0877	0.0040	0.0452	-0.1571	-0.1079	-0.2224	1.0000	-0.0430	0.1044	0.0768	0.1995
C_9	-0.2614	0.1995	0.0990	0.1458	-0.2156	-0.0197	-0.1141	-0.0430	1.0000	0.0439	-0.1042	0.2928
C_{10}	-0.3589	0.3217	0.1861	-0.2279	-0.0056	0.1154	0.0442	0.1044	0.0439	1.0000	-0.1330	0.1814
C_{11}	-0.0185	0.0467	-0.0825	-0.1670	0.0032	0.0126	0.1300	0.0768	-0.1042	-0.1330	1.0000	0.2261
C_{12}	0.1165	0.3609	0.2768	-0.1383	0.2417	0.1887	-0.3138	0.1995	0.2928	0.1814	0.2261	1.0000

100 epochs; therefore, each model took 6.86 minutes approximately for training.

G. FUTURE WORK

Emotion recognition from EEG brain signals is a challenging task, and many problems in this area are yet to be resolved. Therefore, there are several directions for future research related to this study. The most crucial challenge for future research is to improve the accuracy of the proposed system further. To this end, we plan to design a more generalized and powerful model by increasing the depth of the model to determine how this affects accuracy. To develop such models, state-of-the-art and sophisticated DL techniques were used. We employed only one fusion technique, and explored other state-of-the-art fusion techniques. In the current study, we used a benchmark dataset for emotion recognition research (i.e., DEAP). A future direction is to record our dataset for human emotion recognition.

In the future, it would also be interesting to investigate the identification of categorical and individual emotions such as joy, anger, and fear for emotion recognition. Although the proposed system performed well on a benchmark dataset, its deployment in a real-time environment for the healthcare sector and security domains is a worthwhile avenue for future work. It would also be interesting to observe whether including more subjects in the experiments has any positive or negative impact on the results as the amount of data for the classifier increases. Future research should consider human knowledge and emotions. In the future, we plan to develop DL-based models that can effectively recognize unknown and emerging emotion categories using continuous DL [52]. We intend to research brain–computer interfaces in the context of real-time sensing, which should be particularly useful in caring for physically disabled and older people.

A future application of AER in ML and AI is the detection of human emotions using wireless EEG recording devices. Such systems will reduce the difficulties faced by the subjects during the recording of EEG signals. Another important future application of AER is the construction of an intelligent emotion detection and recognition application for law enforcement agencies that use EEG signals. Law enforcement agencies can use such applications to detect and recognize suspects by analyzing their brain signals. Another future direction could involve the development of an automotive AI

AQ:4 **TABLE 8. Performance comparison of proposed system with previous work for HV vs. LV and HA vs. LA problems.**

Research	Features	Classifier	Accuracy	
			HV vs. LV	HA vs. LA
Koelstra et al. [15], 2012	PSD	Gaussian naïve Bayes	57.6%	62.0%
Chung and Yoon [25], 2012	PSD	Naïve Bayes	66.6 %	66.4 %
Haung et al. [17], 2012	ASP	KNN, naïve Bayes, SVM	66.05%	82.46%
Zhang et al. [29], 2013	PSD, statistical features	Ontological model	75.19%	81.74%
Rozgic et al. [27], 2013	PSD	SVM	76.9%	69.1%
Candra et al. [26], 2015	Wavelet entropy	SVM	65.13%	65.33%
Atkinson and Campos [31], 2016	Band power, statistical features, fractal dimension, Hjorth parameters	SVM	73.1%	73.0%
Liu et al. [30], 2016	Deep belief network (DBN)–based features	SVM	85.2%	80.5%
Abeer et al. [28], 2017	PSD, frontal asymmetry	DNN	82.0%	82.0%
Tripathi et al. [32], 2017	Statistical time-domain features	Neural networks (NNs)	81.4%	73.3%
Zhuang et al. [34], 2017	EMD-based features	SVM	69.1%	71.9%
Li et al. [35], 2017	Frequency-domain, nonlinear-dynamic domain, and time-domain features	SVM	80.7%	83.7%
Yin et al. [33], 2017	Power spectral and statistical features	Neural networks (NNs)	83.04%	84.18%
Menezes et al. [36], 2017	PSD, HOC, and statistical features	SVM	88.4%	74.0%
Alazrai et al. [16], 2018	QTFD-based features	SVM	85.8%	86.6%
Yang et al. [49], 2018	Parallel convolutional recurrent neural network (PCRNN)		90.8%	91.03%
Pandey et al. [46], 2019	Variational mode decomposition (VMD)	Deep neural network (DNN)	62.5%	61.25%
Islam et al. [50], 2019	CNN		81.51%	79.42%
Yang et al. [51], 2019	Multi-column convolutional neural network (MC-CNN)		81.4%	80.5%
Pandey et al. [46], 2019	DNN		62.25%	61.25%
Li et al. [37], 2019	Spatial-temporal features	R2G-STNN	93.38% (positive versus neutral versus negative classes)	
Shen et al. [38], 2020	Spatial-temporal features	Parallel sequence-channel projection CNN	96.16%	95.89%
Sharma et al. [47], 2020	HOS + LSTM	Softmax	84.16%	85.21%
Shen et al. [28], 2020	PSCP-Net (Parallel sequence-channel projection)	Softmax	96.16%	95.89%
Ahmed et al. [61], 2020	Two-layer feature-selection process	Score and rank-level fusion	90.0% during walking, 96.0% during sitting, and 86.66% in an action-independent scenario to recognize five emotions: happiness, sadness, fear, anger, and neutral	
Zhang et al. [62], 2020	Bimodal deep automatic encoder (BDAE)	LIBSVM	85.71% for multimodal emotion recognition	
Cui et al. [48], 2020	Regional-asymmetric convolutional neural network (RA-CNN)		96.65%	97.11%
Yin et al. [55], 2021	Graph convolutional neural networks (GCNN) and LSTM		90.45%	90.60%
Tan et al. [56], 2021	Spiking neural network (SNN)		67.76%	78.97%
Current work	Features extracted using LP-1D-CNN model	Softmax	**98.43%**	**97.65%**

system for driver monitoring using EEG signals. Such a system can detect the driver's status, for example, their emotions, tiredness, attention level, distraction, and drowsiness. Other types of emotions were also considered. It is also preferable to have more balanced datasets for each class. Recently, DL techniques have been widely used to improve emotion recognition performance. Deep neural networks can also be used to investigate brain functional connectivity patterns during different emotions by using graph CNNs (GCNN).

VI. CONCLUSION

We addressed the problem of emotion recognition by using EEG brain signals. We modeled this problem as two binary classification problems (i.e., HV vs. LV and HA vs. LA) based on a dimensional approach to emotion modeling [11]. For the HV vs. LV and HA vs. LA problems, we developed a Deep-AER system based on deep LP-1D-CNN models and validated it using the benchmark DEAP dataset. Most existing studies on this problem have used hand-engineered features, which involve laborious manual parameter tuning. Their performance depends heavily on the selection of hyperparameters and they do not learn the internal structure of the data. As such, they do not generalize well across cases. In addition, they involve laborious designs in which the first features are extracted, selected, and then passed to a classifier. All these stages involve hyperparameters, whose joint manual tuning is laborious and time-consuming.

In contrast, we proposed a deep LP-1D-CNN model that contains a small number of learnable parameters learned in an end-to-end fashion, automatically and implicitly extracted and selected features, and finally classified them. Using LP-1D-CNN, we built a two-level ensemble model. In the first level of the ensemble, each channel is scanned incrementally using LP-1D-CNN to generate predictions that are fused using a majority vote. The second level of the ensemble combines the predictions of all the channels of an EEG signal using a majority vote to detect the emotional state.

To identify the brain regions that play a dominant role in AER, we analyzed EEG signals in five brain regions: *FRONT*, *CENT*, *PERI*, *OCCIP*, and *ALL*. The results indicate that *FRONT* plays a dominant role in AER. Over this region, Deep-AER achieved accuracies of 98.43% and 97.65% for two AER problems (HV vs. LV and HA vs. LA, respectively). The Deep-AER system makes substantial improvements over previous systems for emotion recognition based on EEG signals, as it significantly outperforms state-of-the-art techniques by a large margin. The proposed system outperformed the existing methods for three reasons. First, it is based on a DL approach, which has shown outstanding performance in many problems compared with hand-engineered features. Second, it employs a pyramid architecture to design CNN models with less complexity, and does not require large amounts of data for learning. Third, it uses an ensemble strategy that combines local decisions with the global context. More importantly, it shows that a DL-based system for classifying brain signals outperforms traditional ML techniques. The results show that our DL-based method demonstrates a better classification performance than other state-of-the-art approaches, suggesting that this method can be successfully applied to develop other EEG-based expert systems.

The focus of this study was on the recognition of emotions, such as valence and arousal. The proposed method predicts the HV vs. LV and HA vs. LA states with high accuracy; it can be helpful in many mental health problems, including obsessive-compulsive disorder.

REFERENCES

[1] Y. Liu and O. Sourina, "EEG-based subject-dependent emotion recognition algorithm using fractal dimension," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2014, pp. 3166–3171.

[2] 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, nos. 3–4, pp. 60–73, 2016.

[3] M. Khateeb, S. M. Anwar, and M. Alnowami, "Multi-domain feature fusion for emotion classification using DEAP dataset," *IEEE Access*, vol. 9, pp. 12134–12142, 2021.

[4] Y. Zhang, C. Cheng, and Y. Zhang, "Multimodal emotion recognition using a hierarchical fusion convolutional neural network," *IEEE Access*, vol. 9, pp. 7943–7951, 2021, doi: 10.1109/ACCESS.2021.3049516.

[5] M. R. Islam, M. A. Moni, M. M. Islam, M. Rashed-Al-Mahfuz, M. S. Islam, M. K. Hasan, M. S. Hossain, M. Ahmad, S. Uddin, A. Azad, S. A. Alyami, M. A. R. Ahad, and P. Lio, "Emotion recognition from EEG signal focusing on deep learning and shallow learning techniques," *IEEE Access*, vol. 9, pp. 94601–94624, 2021.

[6] S. Mohsen and A. G. Alharbi, "EEG-based human emotion prediction using an LSTM model," in *Proc. IEEE Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2021, pp. 458–461, doi: 10.1109/MWSCAS47672.2021.9531707.

[7] 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, doi: 10.1109/TAFFC.2020.3013711.

[8] R. M. Mehmood, R. Du, and H. J. Lee, "Optimal feature selection and deep learning ensembles method for emotion recognition from human brain EEG sensors," *IEEE Access*, vol. 5, pp. 14797–14806, 2017, doi: 10.1109/ACCESS.2017.2724555.

[9] J. X. Chen, P. W. Zhang, Z. J. Mao, Y. F. Huang, D. M. Jiang, and Y. N. Zhang, "Accurate EEG-based emotion recognition on combined features using deep convolutional neural networks," *IEEE Access*, vol. 7, pp. 44317–44328, 2019, doi: 10.1109/ACCESS.2019.2908285.

[10] C. Qing, R. Qiao, X. Xu, and Y. Cheng, "Interpretable emotion recognition using EEG signals," *IEEE Access*, vol. 7, pp. 94160–94170, 2019, doi: 10.1109/ACCESS.2019.2928691.

[11] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness Cogn.*, vol. 17, no. 2, pp. 484–495, 2008, doi: 10.1016/j.concog.2008.03.019.

[12] C. Darwin and P. Ekman, *The Expression of the Emotions in Man and Animals*, 3rd ed. New York, NY, USA: Oxford Univ. Press, 1998.

[13] J. R. Davitz, "Auditory correlates of vocal expression of emotional feeling," in *The Communication of Emotional Meaning*, J. Davitz, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 101–112.

[14] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emotions*, vol. 1, no. 1, pp. 68–99, Jan. 2010.

[15] 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, Jun. 2012.

[16] R. Alazrai, R. Homoud, H. Alwanni, and M. I. Daoud, "EEG-based emotion recognition using quadratic time-frequency distribution," *Sensors*, vol. 18, no. 8, pp. 27–39, 2018, doi: 10.3390/s18082739.

[17] D. Huang, C. Guan, K. K. Ang, H. Zhang, and Y. Pan, "Asymmetric spatial pattern for EEG-based emotion detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Brisbane, QLD, Australia, Jun. 2012, pp. 1–7.

[18] G. Chanel, J. J. M. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *Int. J. Hum.-Comput. Stud.*, vol. 67, no. 8, pp. 607–627, 2009.

[19] E. Albilali, H. Aboalsamh, and A. Al-Wabil, "Comparing brain-computer interaction and eye tracking as input modalities: An exploratory study," in *Proc. Int. Conf. Current Trends Inf. Technol.*, Dubai, United Arab Emirates, Dec. 2013, pp. 232–236.

[20] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, vol. 3361, no. 10. Cambridge, MA, USA: MIT Press, 1995.

[21] T. Zhang, W. Chen, and M. Li, "AR based quadratic feature extraction in the VMD domain for the automated seizure detection of EEG using random forest classifier," *Biomed. Signal Process. Control*, vol. 31, pp. 550–559, Jan. 2017.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[24] I. Ullah, M. Hussain, E.-U.-H. Qazi, and H. Aboalsamh, "An automated system for epilepsy detection using EEG brain signals based on deep learning approach," *Expert Syst. Appl.*, vol. 107, pp. 61–71, Oct. 2018.

[25] S. Y. Chung and H. J. Yoon, "Affective classification using Bayesian classifier and supervised learning," in *Proc. 12th Int. Conf. Control, Automat. Syst. (ICCAS)*, 2012, pp. 1768–1771.

[26] H. Candra, M. Yuwono, R. Chai, A. Handojoseno, I. Elamvazuthi, H. T. Nguyen, and S. Su, "Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 7250–7253.

[27] V. Rozgic, S. N. Vitaladevuni, and R. Prasad, "Robust EEG emotion classification using segment level decision fusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 1286-1290.

[28] A. Al-Nafjan, M. Hosny, A. Al-Wabil, and Y. Al-Ohali, "Classification of human emotions from electroencephalogram (EEG) signal using deep neural network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 9, pp. 419–425, 2017.

[29] X. Zhang, B. Hu, J. Chen, and P. Moore, "Ontology-based context modeling for emotion recognition in an intelligent web," *World Wide Web*, vol. 16, no. 4, pp. 497–513, Jul. 2013.

[30] W. Liu, W. L. Zheng, and B. L. Lu, "Emotion recognition using multimodal deep learning," in *Proc. Int. Conf. Neural Inf. Process.* Cham, Switzerland: Springer, 2016, pp. 521–529.

[31] J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Expert Syst. Appl.*, vol. 47, pp. 35–41, Apr. 2016.

[32] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya, "Using deep and convolutional neural networks for accurate emotion classification on DEAP data," in *Proc. 29th Innov. Appl. Artif. Intell. (IAAI)*, 2017, pp. 4746–4752.

[33] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Comput. Methods Programs Biomed.*, vol. 140, pp. 93–110, Mar. 2017.

[34] N. Zhuang, Y. Zeng, L. Tong, C. Zhang, H. Zhang, and B. Yan, "Emotion recognition from EEG signals using multidimensional information in EMD domain," *BioMed Res. Int.*, vol. 2017, pp. 1–9, Aug. 2017.

[35] X. Li, J. Z. Yan, and J. H. Chen, "Channel division based multiple classifiers fusion for emotion recognition using EEG signals," in *Proc. Int. Conf. Inf. Sci. Technol.*, vol. 11, Wuhan, China, Mar. 2017, p. 7006.

[36] M. L. R. Menezes, A. Samara, L. Galway, A. Sant'Anna, A. Verikas, F. Alonso-Fernandez, H. Wang, and R. Bond, "Towards emotion recognition for virtual environments: An evaluation of EEG features on benchmark dataset," *Pers. Ubiquitous Comput.*, vol. 21, no. 6, pp. 1003–1013, Dec. 2017.

[37] Y. Li, W. Zheng, L. Wang, Y. Zong, and Z. Cui, "From regional to global brain: A novel hierarchical spatial–temporal neural network model for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 568–578, Apr. 2022, doi: 10.1109/TAFFC.2019.2922912.

[38] L. Shen, W. Zhao, Y. Shi, T. Qin, and B. Liu, "Parallel sequence-channel projection convolutional neural network for EEG-based emotion recognition," *IEEE Access*, vol. 8, pp. 222966–222976, 2020, doi: 10.1109/ACCESS.2020.3039542.

[39] B. M. Mathunjwa, Y.-T. Lin, C.-H. Lin, M. F. Abbod, and J.-S. Shieh, "ECG arrhythmia classification by using a recurrence plot and convolutional neural network," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102262, doi: 10.1016/j.bspc.2020.102262.

[40] M. Zhang, X. Lin, and P. Du, "An epilepsy and seizure classification approach based on multi-spike liquid state machines," in *Proc. 15th Int. Conf. Comput. Intell. Secur. (CIS)*, Dec. 2019, pp. 103–107, doi: 10.1109/CIS.2019.00030.

[41] C. Brunner, M. Billinger, M. Seeber, T. R. Mullen, and S. Makeig, "Volume conduction influences scalp-based connectivity estimates," *Frontiers Comput. Neurosci.*, vol. 10, Nov. 2016, Art. no. 121, doi: 10.3389/fncom.2016.00121.

[42] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.

[43] (Jun. 15, 2017). *TensorFlow*. Accessed: Jun. 25, 2017. [Online]. Available: https://www.tensorflow.org/

[44] R. J. Davidson, J. Schwartz, G. E. Saron, J. C. Bennett, and D. J. Goleman, "Frontal versus parietal EEG asymmetry during positive and negative affect," *Psychophysiology*, vol. 16, no. 2, pp. 202–203, 1979.

[45] R. J. Davidson, P. Ekman, C. D. Saron, J. A. Senulis, and W. V. Friesen, "Approach-withdrawal and cerebral asymmetry: Emotional expression and brain physiology: I," *J. Pers. Social Psychol.*, vol. 58, no. 2, pp. 330–341, 1990, doi: 10.1037/0022-3514.58.2.330.

[46] P. Pandey and K. R. Seeja, "Subject independent emotion recognition from EEG using VMD and deep learning," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 5, pp. 1730–1738, May 2022, doi: 10.1016/j.jksuci.2019.11.003.

[47] R. Sharma, R. B. Pachori, and P. Sircar, "Automated emotion recognition based on higher order statistics and deep learning algorithm," *Biomed. Signal Process. Control*, vol. 58, Apr. 2020, Art. no. 101867.

[48] 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, doi: 10.1016/j.knosys.2020.106243.

[49] 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, doi: 10.1109/IJCNN.2018.8489331.

[50] M. R. Islam and M. Ahmad, "Virtual image from EEG to recognize appropriate emotion using convolutional neural network," in *Proc. 1st Int. Conf. Adv. Sci., Eng. Robot. Technol. (ICASERT)*, May 2019, p. 14, doi: 10.1109/ICASERT.2019.8934760.

[51] H. Yang, J. Han, and K. Min, "A multi-column CNN model for emotion recognition from EEG signals," *Sensors*, vol. 19, no. 21, p. 4736, Oct. 2019, doi: 10.3390/s19214736.

[52] S. Thuseethan, S. Rajasegarar, and J. Yearwood, "Deep continual learning for emerging emotion recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 4367–4380, 2022, doi: 10.1109/TMM.2021.3116434.

[53] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 417–429, Jul./Sep. 2017, doi: 10.1109/TAFFC.2017.2712143.

[54] Y. Tang, F. Ren, and W. Pedrycz, "Fuzzy C-means clustering through SSIM and patch for image segmentation," *Appl. Soft Comput.*, vol. 87, Feb. 2020, Art. no. 105928, doi: 10.1016/j.asoc.2019.105928.

[55] 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.

[56] C. Tan, M. Šarlija, and N. Kasabov, "NeuroSense: Short-term emotion recognition and understanding based on spiking neural network modelling of spatio–temporal EEG patterns," *Neurocomputing*, vol. 434, pp. 137–148, Apr. 2021.

[57] L. Cing, F. Wu, Z. L. Yu, and F. Hu, "A real-time speech emotion recognition system and its application in online learning," in *Emotions, Technology, Design, and Learning*. New York, NY, USA: Academic, 2016, pp. 27–46.

[58] F. T. Giuntini, K. L. de Moraes, M. T. Cazzolato, L. de Fatima Kirchner, M. de Jesus D. D. Reis, A. J. M. Traina, A. T. Campbell, and J. Ueyama, "Modeling and assessing the temporal behavior of emotional and depressive user interactions on social networks," *IEEE Access*, vol. 9, pp. 93182–93194, 2021, doi: 10.1109/ACCESS.2021.3091801.

[59] F. T. Giuntini, K. L. P. de Moraes, M. T. Cazzolato, L. D. F. Kirchner, M. D. J. D. D. Reis, A. J. M. Traina, A. T. Campbell, and J. Ueyama, "Tracing the emotional roadmap of depressive users on social media through sequential pattern mining," *IEEE Access*, vol. 9, pp. 97621–97635, 2021, doi: 10.1109/ACCESS.2021.3095759.

[60] C. Zhang and L. Xue, "Autoencoder with emotion embedding for speech emotion recognition," *IEEE Access*, vol. 9, pp. 51231–51241, 2021, doi: 10.1109/ACCESS.2021.3069818.

[61] F. Ahmed, A. S. M. H. Bari, and M. L. Gavrilova, "Emotion recognition from body movement," *IEEE Access*, vol. 8, pp. 11761–11781, 2020, doi: 10.1109/ACCESS.2019.2963113.

[62] H. Zhang, "Expression-EEG based collaborative multimodal emotion recognition using deep autoencoder," *IEEE Access*, vol. 8, pp. 164130–164143, 2020, doi: 10.1109/ACCESS.2020.3021994.

[63] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[64] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Pers. Social Psychol.*, vol. 17, no. 2, pp. 124–129, 1971, doi: 10.1037/h0030377.

[65] P. Ekman, "Facial expression and emotion," *Amer. Psychol.*, vol. 48, no. 4, pp. 384–392, 1993, doi: 10.1037/0003-066X.48.4.384.

[66] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Sci. Inf.*, vol. 44, no. 4, pp. 695–729, 2005, doi: 10.1177/0539018405058216.

MUHAMMAD HUSSAIN (Senior Member, IEEE) received the M.Sc. and M.Phil. degrees from the University of the Punjab, Lahore, Pakistan, in 1990 and 1993, respectively, and the Ph.D. degree in computer science from Kyushu University, Fukuoka, Japan, in 2003. He is currently a Professor with the Department of Computer Science, King Saud University (KSU), Saudi Arabia. He worked as a Postdoctoral Researcher with Kyushu University, from April 2003 to September 2005, and received funding from Japan Science and Technology Agency (JST). He joined the Department of Computer Science, KSU, as an Assistant Professor, in September 2005, and was promoted to an Associate Professor, in December 2009, and a Professor, in December 2013. His current research interests include deep learning, image forensics, digital watermarking, medical imaging (mammograms, diabetic retinopathy, and EEG brain signals), and biometrics (face recognition and fingerprint recognition). He has received several research grants from the Japan Science and Technology Agency (JST), the National Science Technology and Innovation Plan (NSTIP) of Saudi Arabia, and the Research Center of the College of Computer and Information Sciences, KSU. He has published more than 160 research papers in ISI-indexed journals and refereed international conference proceedings in these research areas. He is a member of an editorial board, an advisor, and a reviewer of many famous ISI journals, international conferences, and funding agencies. He is currently an Editor of the journal *Applied Intelligence* (Springer). He was an Editor of the *Journal of Computer and Information Sciences* and *Journal of King Saud University* (Elsevier) and has served on the program committees of various international conferences.

EMAD-UL-HAQ QAZI received the Ph.D. degree from King Saud University (KSU), Riyadh, Saudi Arabia. He worked as a Postdoctoral Researcher with the Centre of Excellence in Information Assurance (COEIA), KSU, from September 2020 to October 2021, where he is currently a Researcher with the Department of Computer Science, College of Computer and Information Sciences (CCIS). His research interests include security analytics, deep learning, pattern recognition, biometrics, and image and signal processing.

HATIM A. ABOALSAMH (Senior Member, IEEE) received the Ph.D. degree in computer engineering and science from the University of Miami, USA, in 1987. He was the Vice Rector of Development and Quality with King Saud University (KSU), Riyadh, Saudi Arabia, from 2006 to 2009, and the Dean of the College of Computer and Information Sciences. He is currently a Professor and the Chairperson of the Department of Computer Science, KSU. His research interests include pattern recognition, biometrics, probability modeling, and machine learning. He is also a Fellow Member of the British Computer Society and a Senior Member of the Association of Computing Machinery, USA, and the International Association of Computer Science and Information Technology. He was the Vice President of the Saudi Computer Society. He was the Editor-in-Chief of the *Journal KSU—Computer Sciences*. He is also the Editor-in-Chief of the Saudi Computer Society.

IHSAN ULLAH received the Ph.D. degree from the University of Milan, Italy, in 2017. Currently, he is an Assistant Professor with the School of Computer Science, University of Galway, Ireland. He is also a funded Investigator with the Insight Research Center for Data Analytics, Galway. His research interests include computer vision, deep learning, biometrics, and social data.