Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

Tech Science Press

https://doi.org/10.32604/rimni.2026.075907

ARTICLE

Hierarchical Convolutional Neural Network for Emotion Recognition Using EEG and Facial Expressions

Muhammad Shafiq[1,2,*], **Salman Afsar Awan**[2] **and Tahir Kamal**[3]

[1]School of Computer Science, Shandong Xiehe University, Jinan, China
[2]Department of Computer Science, University of Agriculture, Faisalabad, Pakistan
[3]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China

*Corresponding Author: Muhammad Shafiq. Email: shafiqk786@hotmail.com

Received: 11 November 2025; Accepted: 19 February 2026; Published: 06 November 2026

ABSTRACT: Emotion recognition is crucial for advancing human–computer interaction (HCI) by enabling systems to interpret complex affective states. While Electroencephalogram (EEG) signals provide direct insights into neural activity, facial expressions offer external emotional cues. However, unimodal systems often struggle with robustness and generalization across diverse subjects. This study presents a Hierarchical Convolutional Neural Network (HCNN) framework that integrates EEG and facial expressions through multi-level convolutional feature extraction and feature-level fusion. The proposed model combines deep hierarchical representations with handcrafted temporal–frequency and texture-based descriptors to form a unified feature vector. Experiments on the MAHNOB-HCI and DEAP datasets show that the HCNN achieves accuracies of 91.40% and 88.09%, outperforming CNN-, LSTM-, and SVM-based methods. The results demonstrate the model's ability to effectively capture complementary cross-modal correlations while reducing feature redundancy and computational complexity. The HCNN framework shows great promise for real-time emotion recognition applications, offering a scalable, interpretable, and data-efficient solution for multimodal emotion recognition in next-generation HCI systems.

KEYWORDS: Human-computer interaction; emotion recognition; electroencephalogram (EEG); hierarchical convolutional neural network; multimodal fusion; facial expression analysis; affective computing

1 Introduction

Emotion recognition has become increasingly critical for advancing human–computer interaction (HCI) systems across diverse application domains. Emotions represent complex mental states shaped by both internal and external stimuli, manifesting through intricate psycho-physiological processes. Understanding these emotional states enables deeper insights into human behavior, which is essential for designing adaptive and intelligent systems. Affective computing has therefore emerged as a key discipline for recognizing emotions using multimodal data in computational environments [1,2].

Early research demonstrated the limitations of unimodal emotion recognition, as reliance on a single information source often fails to capture the multidimensionality nature of affective states. Consequently, bimodal and multimodal approaches, particularly those combining facial expressions with electroencephalogram (EEG) signals, have gained increasing attention [3–5]. These frameworks typically involve two core stages: extracting informative representations from raw data and applying fusion strategies to integrate complementary modalities [6].

Copyright © 2027 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

EEG-based emotion recognition remains one of the most actively studied areas due to EEG's ability to reflect neural activity associated with emotional changes. Recent studies have focused on optimizing feature extraction techniques and selecting effective classifiers for EEG signals [1,7–9]. For example, Reference [10] investigated early- and late-fusion strategies by integrating EEG with blood volume pulse signals, while [11] explored deep learning methods for learning feature correlations in facial expression recognition. Hierarchical feature learning has also been widely explored, including stacked autoencoders applied to EEG power spectral density, modular network designs to mitigate fusion limitations [12], and multi-column CNN architectures to improve recognition accuracy [13]. Despite these advances, many existing approaches continue to suffer from premature fusion, feature redundancy, and increased computational overhead when processing large-scale EEG data [14].

Beyond handcrafted feature-based approaches, deep learning has substantially advanced emotion recognition by enabling end-to-end learning from raw or minimally processed signals. Architectures such as CNN autoencoders [15], convolutional LSTMs [16], recurrent neural networks [17], and hybrid deep models have demonstrated strong performance across datasets [18,19]. CNNs are effective in capturing spatial abstractions, whereas RNNs and LSTMs model temporal dependencies. More recently, attention-based Transformer architectures have been applied to multimodal emotion recognition to enhance cross-modal alignment [20–22]. Nevertheless, their models often involve high computational complexity and may face challenges related to robustness across subjects and real-time deployment [9].

Although traditional emotion models such as valence–arousal framework provide a practical foundation, they may oversimplify the representation of affective states. Alternative emotion models, including Plutchik's Wheel of Emotions or Ekman's six basic emotions, as well as extended dimensions incorporating dominance, suggest richer representational possibilities for multimodal systems. Empirical evidence consistently shows that integrating EEG with facial expressions improves recognition performance compared to unimodal systems [23]; however, challenges related to feature redundancy, fusion strategy design, and computational scalability remain unresolved [24].

Prior EEG–facial fusion studies commonly combine modalities at a single stage or rely primarily on final-layer deep representations, which can suppress intermediate complementary cues and increase redundancy. In contrast, we propose a hierarchical CNN-based framework that preserves and integrates multi-level representations while also leveraging compact handcrafted descriptors for EEG and facial texture. This hybrid design is intended to improve robustness under noise and modality uncertainty while keeping the architecture lightweight compared with attention-heavy alternatives.

To address these challenges, this study introduces a hierarchical convolutional neural network (HCNN) framework that explicitly extracts and preserves multi-level features representations from both EEG and facial modalities. Unlike prior hierarchical fusion models that typically aggregate features at a single convolutional depth or at the decision level, the proposed HCNN retains intermediate representations at multiple abstraction levels, enabling complementary modeling of spatial, temporal, and cross-modal characteristics. In addition, the framework integrates deep hierarchical features with hand-crafted time- and frequency-domain EEG descriptors and facial texture features to enhance robustness against EEG non-stationarity and facial variability while reducing feature redundancy through feature-level fusion.

In summary, this work presents a hierarchical multimodal learning framework that combines multi-level convolutional feature preservation with adaptive feature-level fusion for bimodal emotion recognition. The proposed approach is evaluated on benchmark datasets and demonstrates competitive performance while maintaining lower architectural complexity compared with attention-based multimodal models.

2 Literature Review

Emotions are commonly defined as the nervous system's organic response to internal or external stimuli [25]. Classical theories proposed by [26] and [27] provide foundational frameworks for understanding emotional states. Ekman identified six basic emotions; joy, surprise, anger, sadness, disgust, and fear, while Russell introduced the circumplex model, in which emotions arise from varying combinations of two neurophysiological dimensions: valence and arousal. For instance, joy is associated with positive valence and moderate arousal. Based on these models, emotion recognition problems are frequently formulated as binary or multi-class classification tasks defined along valence-arousal dimensions.

Among the available modalities, electroencephalogram (EEG) signals have gained prominence in HCI-driven emotion recognition due to their affordability, high temporal resolution, and direct reflection of neural activity. However, EEG signals are inherently non-stationary and susceptible to artifacts, posing challenges for robust emotion classification across sessions and subjects. Despite these challenges, EEG-based emotion recognition has been successfully applied in domains such as e-learning, healthcare, virtual reality, entertainment, and therapeutic monitoring [28–30].

Facial expressions constitute another critical modality, offering visually interpretable cues for emotional states. Facial expressions are commonly categorized into basic emotions, compound emotions formed by blending basic expressions, and micro-expressions, which are subtle and involuntary. Existing facial emotion recognition approaches can be broadly divided into geometric methods, which analyze facial landmarks [24], and texture-based methods, which rely on descriptors such as Gabor filters and local binary patterns [31]. The Facial Action Coding System (FACS) [32] remains a widely adopted framework for decomposing facial expressions into action units and has served as the basis for numerous recognition models [33,34].

Human emotion recognition methods generally follow two complementary strategies: behavioral analysis and physiological signal analysis. While behavioral approaches capture external manifestations of emotion, they may fail to reliably identify internal affective states. Physiological signals, including EEG, blood volume pulse, and peripheral biosignals, provide richer internal representations, motivating the transition from unimodal to multimodal emotion recognition frameworks that exploit complementary modality strengths [35,36].

Early multimodal emotion recognition studies primarily relied on traditional machine learning techniques. For Example, Reference [37] combined EEG and facial expressions using support vector machines, achieving approximately 70% accuracy. Reference [38] employed differential entropy features from multi-channel EEG with linear classifiers, reporting around 74% accuracy. Reference [17] examined gender-based effects using EEG and eye movement signals, observing improved performance only under same-gender conditions. Although these studies demonstrated the feasibility of multimodal emotion recognition, classical machine learning approaches struggled with high-dimensional feature spaces, feature redundancy, and computational scalability.

The advent of deep learning has significantly mitigated these limitations. Reference [39] introduced 3D and 1D CNN architectures with decision-level fusion for multimodal emotion recognition, achieving notable performance gains. Reference [29] integrated CNN and LSTM models for EEG-based emotion classification, outperforming unimodal baselines. Reference [24] further combined EEG with peripheral physiological signals, reporting accuracy improvements approaching 89%. These works demonstrate the effectiveness of deep multimodal fusion, though challenges related to redundant feature learning and increased computational cost persist.

Recent research (2023–2025) has focused on advanced EEG-specific modeling strategies. Spectral analysis techniques such as variational mode decomposition (VMD) and entropy-based representations (e.g.,

differential/feature entropy) have been employed to capture discriminative EEG frequency components and time–frequency patterns [40–42]. Deep learning models such as NeuroFeat and ERTNet have been proposed to improve EEG feature learning and interpretability through attention mechanisms and transformer-based architectures [43,44]. Graph neural networks have also been explored for cross-modal correlation learning by preserving spatial electrode relationships [21]. While these approaches demonstrate strong performance, they often require large training datasets and incur high computational overhead, limiting real-time deployment in HCI systems.

EEG preprocessing and artifact removal remain critical challenges in emotion recognition pipelines. EEG signals are commonly contaminated by artifacts arising from eye blinks, muscle activity, and head movements. Traditional artifact removal techniques include independent component analysis and filtering, while recent studies have explored machine learning-based approaches such as *k*-nearest neighbors and recurrent neural networks for automated artifact detection and removal. Despite these advances, residual noise and baseline variability continue to affect EEG-based emotion classification robustness.

Building upon these findings, this study proposes a Hierarchical Convolutional Neural Network (HCNN) framework to address feature redundancy and computational inefficiency in multimodal emotion recognition. By preserving multi-level convolutional representations and integrating hierarchical feature extraction with optimized feature-level fusion, the proposed approach enhances cross-modal dependency modelling. In addition, handcrafted time- and frequency-domain EEG features are combined with deep representations to improve robustness against EEG non-stationarity and facial variability. A comparative summary of prior approaches and their limitations is presented in Table 1.

Table 1: Comparison of past approaches and their limitations.

Study	Modality	Feature Extraction	Model Used	Dataset	Accuracy (%)	Limitations
Almanza-Conejo, Almanza-Ojeda [8]	EEG	Channel selection + CNN	CNN	DEAP	85.2	Limited to unimodal EEG; lacks cross-modal generalization
Zhang et al. [24]	Facial Expressions	LBP + SVM	SVM	FER-2013	82.7	Handcrafted features only; poor adaptability to real-world data
Wu et al. [6]	EEG + Facial	DL-based feature fusion	LSTM	MAHNOB-HCI	88.3	Fusion inefficient; struggles with redundancy
Liu et al. [43]	EEG + Facial	Multimodal feature selection	Hybrid CNN–LSTM	SEED-V	89.5	Lacks attention-based alignment; dataset-specific
Mutawa et al. [9]	EEG + Facial	Real-time multimodal fusion	CNN–BiLSTM	Custom clinical dataset	90.1	High complexity; limited scalability in real-world
Sharma et al. [23]	EEG + Facial	Transfer learning + fine-tuned features	Transformer-based Fusion	DREAMER	92.3	Computationally expensive; limited subject independence

(Continued)

Table 1 (continued)

Study	Modality	Feature Extraction	Model Used	Dataset	Accuracy (%)	Limitations
Chen et al. [45]	EEG + Facial	Cross-modal correlation learning	Graph Neural Network	AMIGOS	91.7	Sensitive to noise; requires large training data
Proposed (Ours)	EEG + Facial	Hierarchical CNN feature fusion + handcrafted features	HCNN	MAHNOB-HCI, DEAP	91.4	Balances accuracy and complexity; yet to be tested on large-scale datasets

3 Materials and Methods

Fig. 1 illustrates the proposed Hierarchical Convolutional Neural Network (HCNN) framework. Emotion recognition requires processing complex multimodal data, where facial video and EEG signals provide complementary affective cues. The HCNN is designed to address limitations of prior multimodal approaches, such as inadequate cross-modality integration and feature redundancy, by employing hierarchical convolutional layers that extract multi-level representations. These layers enable the joint modeling of spatial, temporal, and frequency-domain characteristics from both modalities. This section details the datasets, preprocessing steps, and input representation strategy used in the proposed framework.

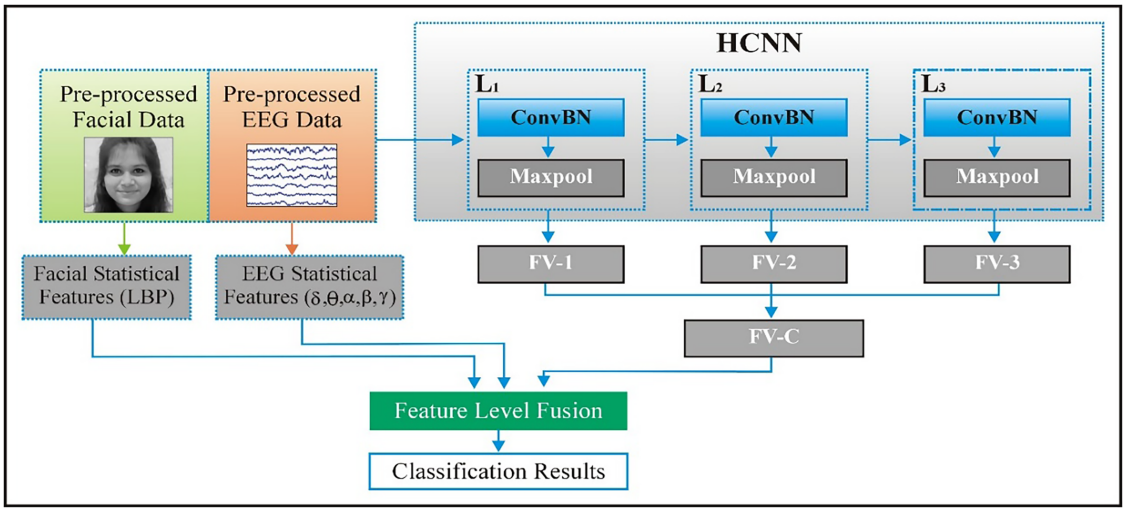


Figure 1: HCNN-based bimodal emotion recognition model [6].

3.1 Data Preprocessing

The MAHNOB-HCI Dataset [31] consists of emotional responses elicited by 20 audiovisual stimuli. EEG and facial recordings were obtained from 22 participants (10 males and 12 females), each of whom provided self-assessed valence and arousal ratings on a 1–9 scale. Following common practice, a threshold of 5 was applied to binarize emotional states into high and low classes. EEG signals were recorded from 32 electrodes placed according to the international 10–20 System with a Biosemi Active II device at a sampling rate of 256 Hz. Synchronized facial videos were captured at a resolution of 720 × 580 pixels and 60 frames per

second. In addition to self-reports, facial valence annotations were provided by five independent observers using FEELTRACE [46].

The DEAP dataset [47] includes multimodal recordings from 32 participants, each exposed to 40 one-minute music video clips. EEG signals were recorded form 32 channels at 512 Hz and downsampled to 128 Hz. Facial video recordings are available for 22 participants only; therefore, experiments involving facial expressions and bimodal fusion were restricted to subjects with complete EEG-facial data. Participants rated arousal and valence on a 1–9 scale, and a threshold of 5 was applied for binary classification. Trials with missing or incomplete facial recordings were excluded rather than imputed to avoid introducing bias.

To address class imbalance after thresholding valence and arousal labels, class distributions were examined for each dataset, and random undersampling of the majority class was applied within the training set only. No resampling was performed on validation or test sets.

EEG signals were bandpass filtered between 4–43 Hz to eliminate low-frequency drift and high-frequency noise. Spectral features were extracted using standard frequency bands: theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–43 Hz). Ocular and muscle artifacts were mitigated using ICA (Infomax/FastICA) with the number of independent components equal to the number of EEG channels. Artifact-related components were identified using ICLabel, correlation with EOG channels, and kurtosis/spectral slope criteria, and were removed before feature extraction.

$$\text{PDF} = \frac{1}{2\pi\sqrt{\sigma}} e^{\frac{(s-\mu)^2}{2\sigma^2}} \tag{1}$$

3.2 Preserving 3D Data Structure for HCNN Input

To preserve spatiotemporal dependencies inherent in multimodal emotion data, the HCNN processes both EEG and facial inputs in their native three-dimensional formats rather than flattening them into one-dimensional vectors.

3.2.1 Video Data

Facial video data were represented in a 3D format *(H × W × T)*, where *H* and *W* denote spatial dimensions and *T* represents the temporal axis. Convolutional layers extracted spatial features from individual frames, while deeper layer modeled temporal dynamics across frame sequences. Facial regions were cropped and centered to reduce background interference [48], resized to 227 × 227 pixels, and augmented using horizontal and vertical flips as well as small random rotations (±10°) to increase training diversity without altering emotion-related cues.

3.2.2 EEG Signals

EEG inputs were represented in an $E \times T \times R$ structure, where E denotes the number of electrodes, T corresponds to time samples, and R represents trials. Convolutional filters were applied across electrode dimensions to learn spatial dependencies, while subsequent layers captured temporal dynamics and trial-level variability.

To avoid data leakage, all temporal windows generated from the same trial were assigned exclusively to either the training, validation, or test set. No overlapping windows from the same trial appeared across different data partitions, ensuring that performance metrics were not artificially inflated by temporal correlation.

This three-dimensional input strategy preserves intrinsic multimodal patterns and enables the HCNN to effectively learn complementary representations from EEG and facial data.

4 Proposed Model

4.1 Traditional CNN Model

With advances in computational power, deep learning has become a dominant tool for classification. Convolutional neural networks (CNNs) are particularly effective at extracting hierarchical features representations directly from raw data by jointly learning features and classifiers within a unified framework [49–51]. Convolutional layers progressively learn features of increasing abstraction, while fully connected layers and softmax classifiers enable final decision-making.

In a standard CNN, each layer performs nonlinear transformations on the output of the preceding layers apply trainable kernels to generate activation maps, whereas pooling layers reduce dimensionality and improve robustness against translation and local distortion. The convolutional operation at layer l is expressed as:

$$O_n^{(l)}(x,y) = \sigma\left(\sum_{m=1}^{M}\sum_{p=0}^{F-1}\sum_{q=0}^{F-1} k_{nm}^{(l)}(p,q) * O_m^{(l-1)}(x-p,y-q) + b_n^{(l)}\right) \tag{2}$$

Here, m and n denote input and output feature maps, $k_{nm}^{(l)}$ is the convolution kernel, $\sigma(.)$ is the nonlinear activation function, and $b_n^{(l)}$ is bias.

Pooling operations reduce computational complexity and enforce invariance. For max pooling:

$$O_n^{(l+1)}(x,y) = \max_{1<i<h,1<j<w}\left(O_m^{(l)}(x+i,y+j)\right) \tag{3}$$

where h, w define the pooling filter size.

At the output layer, softmax produces posterior probabilities:

$$P(y=k|z) = \frac{e^{zk}}{\sum_{j=1}^{k} e^{zj}} \tag{4}$$

where z is the logit vector. Training occurs through forward propagation and error-driven backpropagation, adjusting weights to minimize classification loss [48].

4.2 Hierarchical CNN Model

The proposed Hierarchical Convolutional Neural Network (HCNN) extends traditional CNN by explicitly preserving and integrating feature representations extracted at multiple convolutional depths. Unlike conventional CNN architectures which rely solely on the final convolutional layer, the HCNN retains intermediate feature vectors, denoted as FV-1, FV-2, FV-3, to capture complementary information at different abstraction levels.

The architecture (Fig. 2) employs three convolutional blocks with kernel sizes 7 × 7, 5 × 5, and 3 × 3. Each block is followed by ConvBN (convolution + batch normalization), ReLU activation, and 2 × 2 max pooling. Dropout is applied before fusion to mitigate overfitting. Residual connections are incorporated to improve stability and gradient flow, consistent with modern architectures such as ResNets [52]. The hierarchical feature representations serve distinct roles:

- **FV-1** captures low-level spatial and spectral patterns, including facial edges, textures, and EEG rhythmic activity.
- **FV-2** encodes mid-level temporal dynamics and inter-channel EEG correlations.
- **FV-3** models high-level semantic and cross-modal dependencies.

These feature vectors are concatenated to form a combined hierarchical feature representation denoted as FV-C.

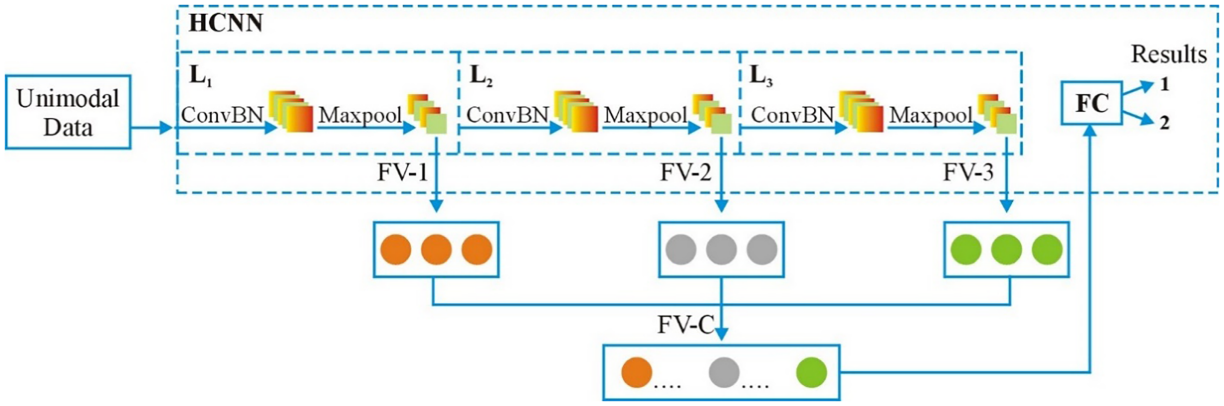


Figure 2: Proposed model of Hierarchical CNN to recognize emotions using EEG and facial data.

4.3 Handcrafted Feature Extraction

To complement deep hierarchical representations, handcrafted descriptors are extracted to enhance robustness, particularly for non-stationary EEG signals.

4.3.1 Hjorth Parameters

Hjorth parameters capture time-domain EEG characteristics and are defined as follows [53]:

$$Activity = var(y(t)) \tag{5}$$

$$Mobility = \sqrt{\frac{var(y'(t))}{var(y(t))}} \tag{6}$$

$$Complexity = \frac{Mobility(y'(t))}{Mobility(y(t))} \tag{7}$$

4.3.2 Sample Entropy

Sample entropy measures signal irregularity and complexity [54]:

$$SampEn\,(m, r, N) = -ln\frac{A\,(m, r, N)}{A\,(m+1, r, N)} \tag{8}$$

where (m) is embedding dimension, (r) is tolerance, and (N) is data length.

For facial images, Local Binary Patterns (LBP) are applied to 3 × 3 neighborhoods [55].

$$LBP\,(x_c, y_c) = \sum_{n=0}^{7} 2^n s(i_n - i_c) \tag{9}$$

where i_c and i_n are gray values at the center and neighbor pixels. Uniform LBP reduces feature dimensionality and ensures rotation invariance. The resultant the LBP operator on a 3 × 3 image can be seen in Fig. 3a, while the LBP features of a face image are shown in Fig. 3b.

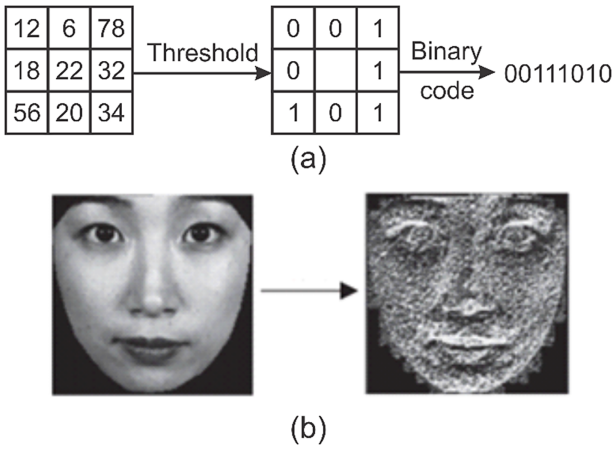


Figure 3: Feature extraction process for facial emotion representation: (**a**) LBP operation; (**b**) LBP features of a facial image.

4.4 Matching Score-Based Fusion

Let $\boldsymbol{F} = \{F_{i,1}, F_{i,2}, \ldots, F_{i,n}\}$ be facial features and $\boldsymbol{E} = \{E_{i,1}, E_{i,2}, \ldots, E_{i,m}\}$ EEG features. After normalization via Median Absolute Deviation (MAD):

$$MAD = Median(|y_i - \tilde{y}|) \tag{10}$$

the normalized feature vectors are concatenated:

$$FV - C = [F_{norm}, E_{norm}] \tag{11}$$

To reduce feature redundancy and emphasize informative cross-modal relationships, a matching score–based fusion strategy is applied:

$$FV - C = \sum_{i=1}^{N} w_i f_i \tag{12}$$

where w_i is the learned weight. When S_{fuse} falls into overlapping regions, Mahalanobis distance is applied for decision refinement.

Adaptive fusion weights. Let $f_i \in R^d$ denote the modality-specific feature vector (e.g., EEG, face, or hierarchical level). Fusion weights are computed by a lightweight gating module:

$$a_i = \frac{exp\,(u^{\top} f_i)}{\sum_{j=1}^{M} exp(u^{\top} f_i)}, \sum_i a_i = 1 \tag{13}$$

where u is trainable. The fused representation is then

$$f_{fuse} = \sum_{i=1}^{M} a_i f_i. \tag{14}$$

Matching score. We compute a similarity-based matching score between modalities using cosine similarity:

$$s\left(f_a, f_b\right) = \frac{f_a \top f_b}{\left\|f_a\right\|, \left\|f_b\right\|}, \tag{15}$$

and use it as an auxiliary consistency signal during training (higher similarity indicates stronger cross-modal agreement).

Mahalanobis refinement. For classification, we optionally refine the decision by measuring the Mahalanobis distance between the fused embedding and class prototypes:

$$D_c(X) = (x - \mu_c)^{\top} \sum -1 \left(X - \mu_c\right), \tag{16}$$

where μ_c and Σ are estimated from training embeddings of class c. The predicted class is $\arg min_c D_c(x)$. This refinement reduces sensitivity to feature-scale imbalance and improves separation when class distributions overlap.

Finally, handcrafted features $\boldsymbol{H} = \{H_1, H_2, \ldots, H_{d_2}\}$ are appended to FV-C:

$$\boldsymbol{X_F} = \{X_1, \ldots, X_{d_1}, H_1, \ldots, H_{d_2}\} \tag{17}$$

This enriched representation integrates deep hierarchical features with handcrafted descriptors for classification via softmax.

4.5 Training Configuration

All networks were implemented in PyTorch and trained using the AdamW optimizer with an initial learning rate of 1e−3 and a cosine learning-rate schedule. The batch size was 32, and training was run for up to 100 epochs with early stopping (monitor: validation loss; patience: 10 epochs; min_delta: 1e−4). Dropout was applied after each convolutional block with a probability of 0.3, and L2 regularization (weight decay) was set to 1e−4. We used binary cross-entropy loss for valence/arousal classification, and all experiments were repeated with a fixed random seed (42) to ensure reproducibility.

5 Results and Discussion

The performance of the proposed deep learning-based emotion recognition framework was evaluated using EEG and facial sequences from the DEAP and MAHNOB-HCI datasets. Since frontal brain regions are strongly associated with emotion processing [56], six frontal EEG channels were selected for analysis. EEG signals were bandpass filtered between 4–43 Hz to eliminate low-frequency drift and high-frequency noise. Spectral features were extracted using standard frequency bands: theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–43 Hz). Ocular and muscle artifacts were mitigated using ICA (Infomax/FastICA) with the number of independent components equal to the number of EEG channels. Artifact-related components were identified using ICLabel, correlation with EOG channels, and kurtosis/spectral slope criteria, and were removed before feature extraction.

A sliding-window strategy with a window length of 6 s and 50% overlap was applied, producing nine segments per trial and yielding 360 samples per subject across 40 trials. As described in Section 3, all windows originating from the same trial were assigned exclusively to either the training, validation, or test set to avoid temporal leakage. The resulting feature dimensionality per sample was 23,040 for EEG ($6 \times 5 \times 128 \times 6$) and 9,275,220 for video data ($227 \times 227 \times 30 \times 6$).

Model performance was evaluated for unimodal (EEG-only and facial-only) and bimodal configurations. Emotion recognition was formulated along the valence (pleasure/satisfaction) and arousal (activation–de-activation) dimensions, as illustrated in Fig. 4. Results were reported in terms of accuracy, precision, recall, and F1-score.

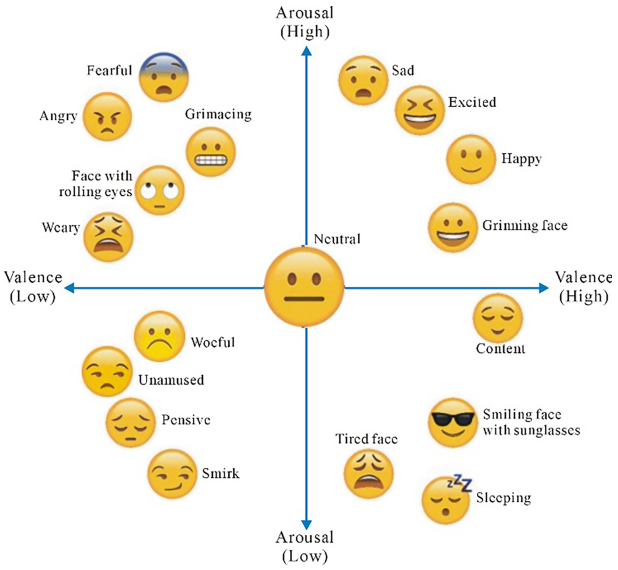


Figure 4: Valence-Arousal representation of emotional states (self-created).

5.1 Statistical Testing

Statistical significance was evaluated at $\alpha = 0.05$. When assumptions of normality were satisfied, we used a paired t-test; otherwise, we used the Wilcoxon signed-rank test. In addition to p-values, we report effect sizes (Cohen's d/Hedges' g) to quantify the magnitude of improvements. Results with $p \geq 0.05$ are described as not statistically significant.

For facial expression data, Levene's test yielded $p = 0.068$, satisfying homogeneity, so a standard t-test was applied. For EEG data, Levene's test gave $p = 0.052$, indicating a borderline violation; therefore, Welch's t-test was used.

5.2 Facial Expression Results

Using MAHNOB-HCI, the model achieved 82.43% (arousal) and 81.51% (valence). On DEAP, accuracies were 80.08% (arousal) and 80.74% (valence). Precision, recall, and F1-scores are presented in Table 2.

Table 2: Facial-expression-only emotion recognition results on MAHNOB-HCI and DEAP datasets.

Dataset	Emotion	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)	*p*-Value
MAHNOB-HCI	Arousal	82.43	83.12	83.34	85.56	0.018
	Valence	81.51	83.22	82.71	83.91	<0.001
DEAP	Arousal	80.08	80.97	79.73	80.13	0.068
	Valence	80.74	80.59	80.36	81.05	0.025

For MAHNOB-HCI, both valence and arousal results were statistically significant ($p < 0.05$). For DEAP, valence recognition was statistically significant ($p = 0.025$), while arousal recognition did not reach statistical significance ($p = 0.068$). A cross-dataset comparison of facial-expression-based results is shown in Fig. 5.

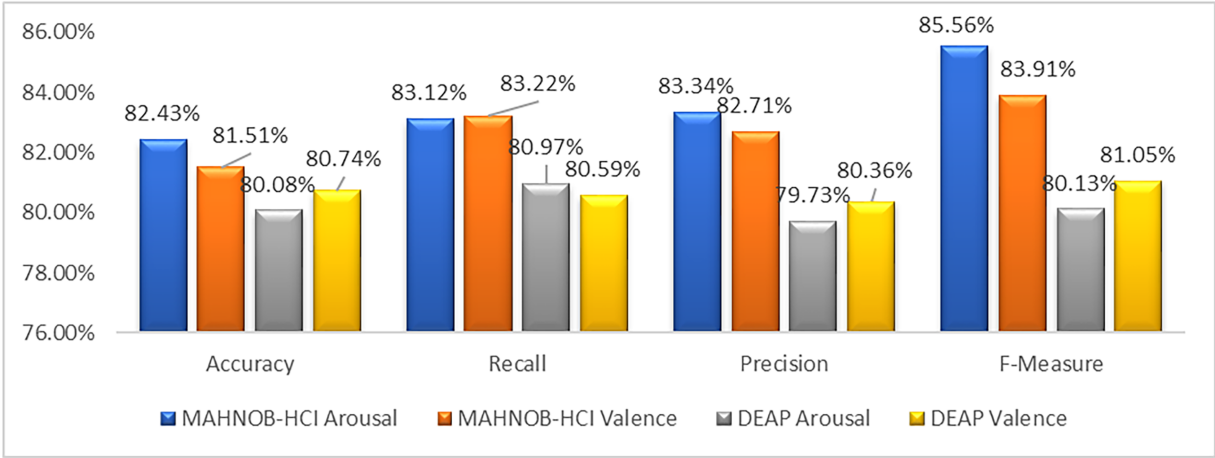


Figure 5: Comparison of valence and arousal recognition using facial expression data.

5.3 EEG Results

With EEG data alone, HCNN achieved 83.27% (arousal) and 83.62% (valence) on MAHNOB-HCI, and 80.15% (arousal) and 79.82% (valence) on DEAP. Precision, recall, and F1-scores are provided in Table 3.

Table 3: EEG-only emotion recognition results on MAHNOB-HCI and DEAP datasets.

Dataset	Emotion	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)	*p*-Value
MAHNOB-HCI	Arousal	83.27	82.29	82.75	82.96	<0.001
	Valence	83.62	82.18	82.64	82.05	<0.001
DEAP	Arousal	80.15	79.37	79.77	80.41	0.052
	Valence	79.82	79.95	80.11	80.52	<0.001

For MAHNOB-HCI, all results were significant ($p < 0.05$). On DEAP, arousal was borderline ($p = 0.052$), while valence remained highly significant ($p < 0.001$). Combined results are shown in Fig. 6.

5.4 Fusion Results

Bimodal fusion consistently yielded the highest performance across both datasets. On MAHNOB-HCI, accuracies of 90.72% (arousal) and 91.40% (valence) were achieved. On DEAP, fusion results reached 88.09% (arousal) and 87.84% (valence). In both datasets, valence recognition outperformed arousal recognition, indicating stronger cross-modal reinforcement for affective valence. Fusion results are summarized in Table 4 and illustrated in Fig. 7.

All fusion results were statistically significant ($p < 0.05$). Performance improvements can be attributed to hierarchical multiscale feature extraction and adaptive feature-level fusion, which effectively capture complementary information from EEG and facial modalities.

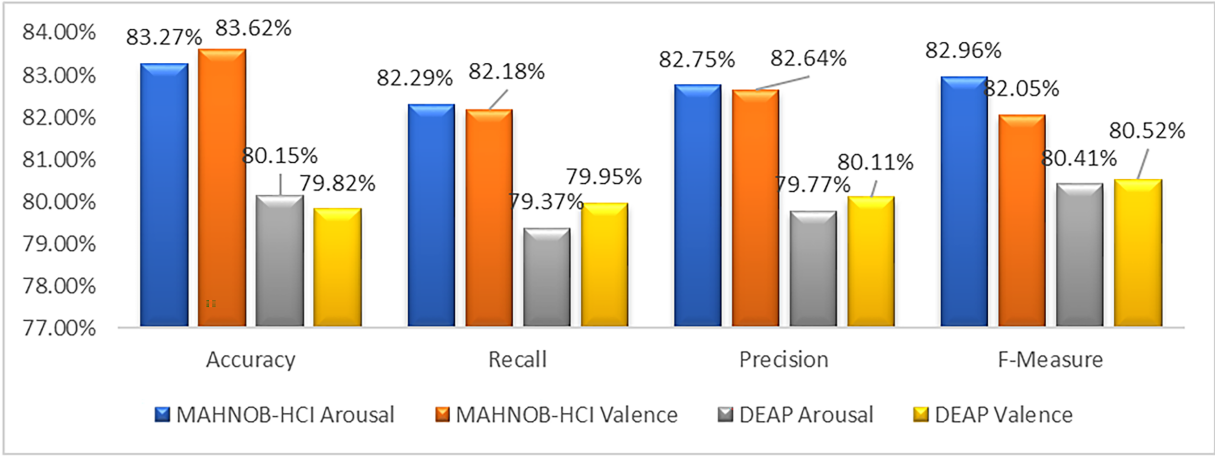


Figure 6: Comparison of valence and arousal recognition using EEG data.

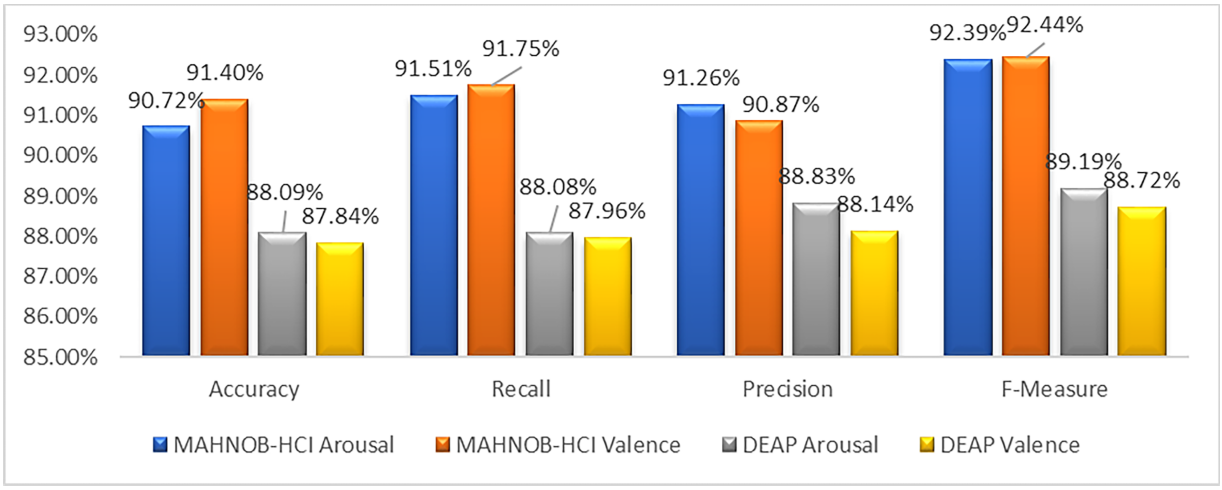


Figure 7: Comparison of fusion-based valence and arousal recognition results.

Table 4: Bimodal EEG–facial fusion results on MAHNOB-HCI and DEAP datasets.

Dataset	Emotion	Accuracy (%)	Recall (%)	Precision (%)	F1-Score (%)	*p*-Value
MAHNOB-HCI	Arousal	90.72	91.51	91.26	92.39	0.005
	Valence	91.40	91.75	90.87	92.44	<0.001
DEAP	Arousal	88.09	88.08	88.83	89.19	<0.001
	Valence	87.84	87.96	88.14	88.72	0.006

5.5 Comparative Analysis

To benchmark performance, the HCNN was compared with state-of-the-art CNN- and SVM-based emotion recognition models (Table 5). The proposed method surpassed existing techniques, with a 3.2% accuracy improvement over the best multimodal fusion approach [6]. The hierarchical fusion strategy was the primary driver of this gain.

Table 5: Comparative performance of emotion recognition models.

Method	Accuracy (%)	Precision	Recall	F1-Score
CNN-LSTM [6]	88.3	86.9	85.5	86.2
SVM + Handcrafted Features	82.7	81.2	80.9	81.0
Transformer-based Fusion Network [57]	89.6	88.4	87.5	88.0
Graph Neural Network for Multimodal EEG-Facial Fusion [45]	90.2	89.5	89.0	89.3
Proposed HCNN Model	91.4	90.8	90.2	90.5

5.6 Computational Complexity

We report parameter count, FLOPs, and inference latency (batch = 1) to support practical deployability. The proposed HCNN contains 3.2 million parameters and requires 1.2 GFLOPs per inference. On an NVIDIA Tesla V100 GPU, average inference time was 25 ms per sample. These results indicate that the proposed model is computationally feasible for near real-time HCI settings relative to attention-heavy multimodal designs.

6 Discussion

The proposed HCNN demonstrates competitive performance when compared with representative multimodal emotion recognition approaches reported in recent literature. Table 6 summarizes comparative results across studies employing EEG, facial expressions, and hybrid fusion strategies. For example, Reference [6] introduced a hierarchical LSTM-based framework that integrates EEG and facial expressions through temporal feature aggregation. Reference [58] proposed a multimodal approach combining CNNs and decision trees to fuse EEG, facial expressions, and galvanic skin response signals, reporting accuracies of 81.2% and 91.5% on the LUMED-2 and DEAP datasets, respectively. Reference [59] employed LIBSVM with fuzzy logic–based decision fusion of EEG and facial features, achieving 85.7% accuracy, while [60] explored decision-level fusion methods using enumerated weighting and adaptive boosting, reporting relatively lower online recognition performance.

Table 6: Comparison of the proposed method with state-of-the-art methods.

Study	Methodology	Dataset	Accuracy (%)	Advantages	Limitations
Huang et al. [60]	SVM for fusion	DEAP	70.0	Computational efficiency	Low accuracy; feature redundancy
Ramzan and Dawn [29]	CNN-LSTM	DEAP	85.2	Temporal feature extraction	Weak cross-modality integration
Wu et al. [6]	Hierarchical LSTM	MAHNOB-HCI	85.5	Improved modality fusion	High computational cost; sequential nature
Cimtay and Ekmekcioglu [28]	Pretrained CNN (transfer learning)	DEAP	79.2	Efficient cross-subject classification	Modest accuracy improvement
Zhang et al. [24]	Hierarchical Fusion CNN	DEAP	85.7	Strong multimodal fusion; good accuracy	Limited scalability across datasets

(Continued)

Table 6 (continued)

Study	Methodology	Dataset	Accuracy (%)	Advantages	Limitations
Li et al. [57]	Transformer-based multimodal fusion	DEAP, SEED	88.6	Captures long-range dependencies; robust fusion	Requires large datasets; high GPU memory
Hassan et al. [61]	Graph Attention Networks (GAT) on EEG topography + CNN on video	DEAP	89.8	Preserves spatial electrode relations	Computational complexity; slower training
Sharma et al. [23]	Cross-attention Multimodal ViT	DEAP, DREAMER	90.2	Strong global context learning; high cross-subject accuracy	Limited interpretability
Proposed HCNN Model	Hierarchical CNN with feature-level fusion	MAHNOB-HCI, DEAP	91.4 (MAHNOB-HCI)/ 88.1 (DEAP)	Multilevel feature extraction and fusion	Complex training pipeline

While prior studies often rely on similar emotional state categorizations, our research focuses specifically on valence and arousal to facilitate direct comparison. As shown in Table 6, the proposed model significantly outperforms other multimodal approaches. By adopting a hierarchical structure, the model enhances multilevel feature representation of bimodal data and reduces variations from physical parameter adjustments, thereby improving classification performance.

ROC analysis further validates the method. Fig. 8 presents ROC curves for valence and Fig. 9 for arousal. In both cases, performance on the MAHNOB-HCI dataset was superior, although results across both datasets were consistent with accuracy, precision, recall, and F1 metrics.

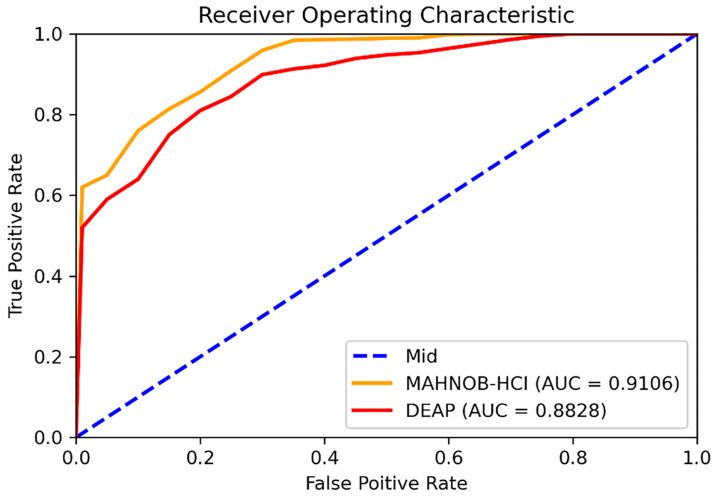


Figure 8: Comparison of results between datasets using the ROC curve on the valence space.

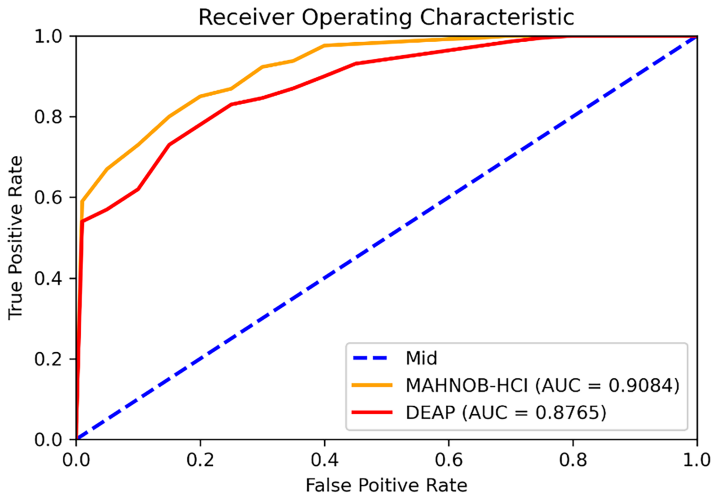


Figure 9: Comparison of results between datasets using the ROC curve on the arousal space.

Topographical analysis provides additional insight into the neurophysiological relevance of the extracted features. As shown in Fig. 10, frontal EEG electrodes exhibit stronger correlations with both valence and arousal across multiple frequency bands, highlighting the importance of frontal brain activity in emotion processing. Fig. 11 shows that valence-related correlations are more broadly distributed across electrodes, whereas arousal-related correlations are predominantly concentrated in frontal regions. These observations are consistent with prior findings reported by [56], reinforcing the physiological plausibility of the proposed approach.

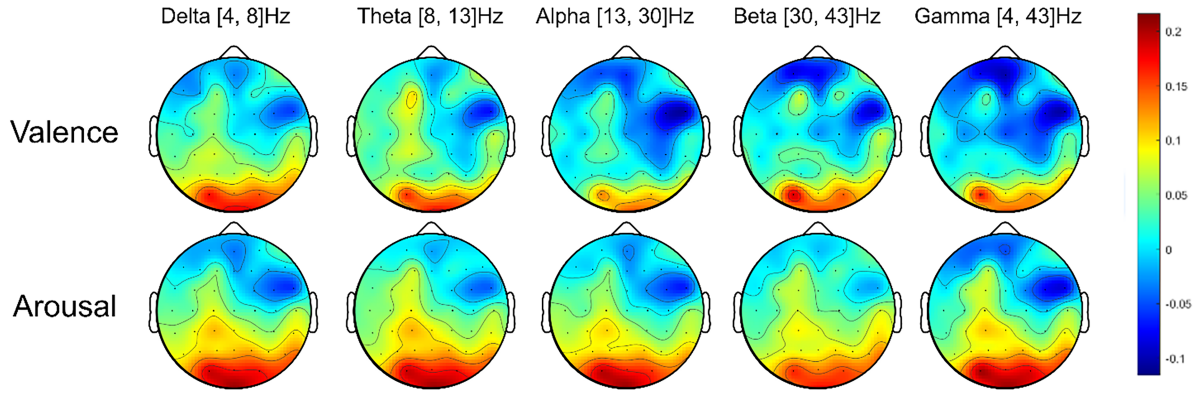


Figure 10: The correlation maps between PSD values, averaged over all sequences, using MAHNOB-HCI dataset and continuous valence for delta, theta, alpha, beta, and gamma bands.

Experiments with varying train-test split ratios (Tables 7 and 8) confirm the robustness of the proposed approach. Accuracy improved with larger training proportions, while p-values remained stable, indicating statistical reliability regardless of split ratio.

Subject-independent experiments were also performed to test generalization across individuals. Results confirmed that the proposed HCNN effectively transfers learned features to unseen subjects, maintaining satisfactory accuracy in bimodal recognition. Evaluating feature-level fusion at different HCNN layers further demonstrated effective integration of EEG and facial modalities. These outcomes highlight the model's potential for practical emotion recognition applications.

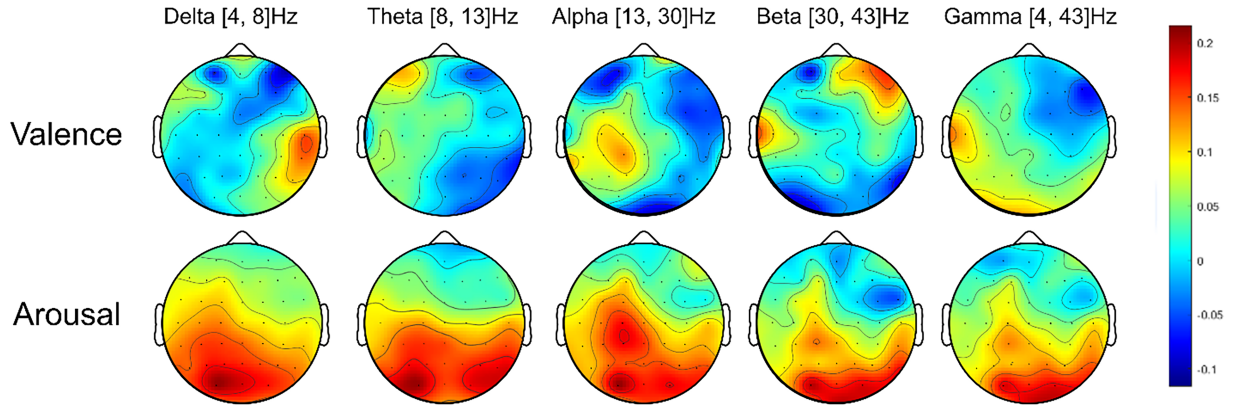


Figure 11: The correlation maps between PSD values, averaged overall sequences, using DEAP dataset and continuous valence for delta, theta, alpha, beta and gamma bands.

Table 7: Average accuracies and standard deviation (%) of varying train-test split ratios on MAHNOB-HCI dataset.

Train/Test Split	Emotion	Accuracy (%)	*p*-Value*
90%/10%	Valence	91.40 ± 2.13	0.012
	Arousal	90.72 ± 2.48	0.021
80%/20%	Valence	89.95 ± 3.43	0.019
	Arousal	88.86 ± 2.91	0.020
70%/30%	Valence	87.03 ± 4.87	0.022
	Arousal	86.65 ± 5.11	0.024

Note: **p*-values are computed from paired statistical tests on subject-level accuracies. Standard deviations are reported only for accuracy values.

Table 8: Average accuracies and standard deviation (%) of varying train-test split ratios on DEAP dataset.

Train/Test Split	Emotion	Accuracy (%)	*p*-Value*
90%/10%	Valence	87.84 ± 2.87	0.019
	Arousal	88.09 ± 3.19	0.015
80%/20%	Valence	86.93 ± 3.48	0.020
	Arousal	84.89 ± 3.92	0.021
70%/30%	Valence	83.78 ± 5.24	0.024
	Arousal	83.66 ± 5.16	0.019

Note: **p*-values are computed from paired statistical tests on subject-level accuracies. Standard deviations are reported only for accuracy values.

6.1 Overfitting Analysis

Given the limited size of the benchmark datasets, overfitting was a potential concern. To mitigate this, dropout layers and batch normalization were employed throughout the network, and training and validation performance were continuously monitored. The close alignment between training and validation accuracy indicates that the model generalizes well within the evaluated datasets. Early stopping was applied to prevent excessive fitting to the training data, contributing to stable convergence behavior.

6.2 Sensitivity Analysis

The robustness of the proposed model was examined under controlled input perturbations. Experimental results show that the HCNN maintains stable performance under low-level noise conditions, with only marginal accuracy degradation. However, higher noise levels ($\sigma > 0.1$ in EEG signals) resulted in more pronounced performance drops, suggesting sensitivity to severe signal corruption. These findings indicate that while the model is reasonably robust, additional adaptive noise filtering or more advanced augmentation strategies could further enhance resilience.

6.3 Limitations and Future Work

Despite the encouraging results, several limitations should be acknowledged. First, both DEAP and MAHNOB-HCI datasets were collected in controlled laboratory environments, which may not fully reflect real-world emotional variability. Second, the current framework focuses on binary valence and arousal classification, which may oversimplify complex affective states. Third, although subject-independent performance is promising, further validation on larger and more diverse populations is required.

Future work will focus on extending the framework to real-time emotion recognition in dynamic environments, incorporating additional affective dimensions such as dominance, and exploring discrete emotion categories. Integrating lightweight attention mechanisms and evaluating the model on larger-scale or in-the-wild datasets may further improve generalizability and applicability.

7 Conclusion

This study presented a hierarchical bimodal learning framework for emotion recognition that integrates EEG and facial expression data through feature-level fusion. The proposed CNN-based architecture was evaluated on the DEAP and MAHNOB-HCI benchmark datasets. Dataset-specific preprocessing procedures were applied to mitigate noise in EEG and facial signals, and class balancing strategies were employed within the training data to address class distribution imbalance. The HCNN framework was used to classify emotional states along the valence and arousal dimensions.

Experimental results indicate that the fusion of hierarchical deep features with complementary handcrafted descriptors yields improved and consistent performance across both datasets. By preserving multi-level representations, the proposed approach reduces information loss associated with premature feature aggregation and enhances the robustness of bimodal feature learning. The results further demonstrate that the framework generalizes reasonably well across subjects under subject-independent evaluation settings.

Overall, the proposed HCNN provides an effective multimodal learning strategy for affective computing and brain–computer interaction research. While the reported results are encouraging, further validation on larger-scale and real-world datasets is necessary to assess robustness under unconstrained conditions. Future work will focus on extending the framework toward real-time emotion recognition, incorporating additional affective dimensions such as dominance, and improving generalizability through adaptive fusion and lightweight model optimization.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the Department of Scientific Research at Shandong Xiehe University, Jinan, China.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Muhammad Shafiq, methodology, Muhammad Shafiq; software, Salman Afsar Awan; validation, Salman Afsar Awan; formal

analysis, Tahir Kamal; investigation, Tahir Kamal; resources, Salman Afsar Awan; data curation, Muhammad Shafiq; writing—original draft preparation, Muhammad Shafiq; writing—review and editing, Tahir Kamal, and Salman Afsar Awan; visualization, Tahir Kamal; supervision, Salman Afsar Awan; project administration, Muhammad Shafiq; funding acquisition, Muhammad Shafiq. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, Muhammad Shafiq, upon reasonable request.

Ethics Approval: This study used publicly available datasets (DEAP and MAHNOB-HCI). The datasets were collected and released for research purposes by the original authors with appropriate ethical approval. Therefore, additional ethical approval was not required for this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Nomenclature

HCNN	Hierarchical Convolutional Neural Network
HCI	Human–computer interaction
EEG	Electroencephalogram
FACS	Facial Action Coding System

References

1. Al-Saadawi HFT, Das B, Das R. A systematic review of trimodal affective computing approaches: text, audio, and visual integration in emotion recognition and sentiment analysis. Expert Syst Appl. 2024;255(3):124852. doi:10.1016/j.eswa.2024.124852.
2. Cai Y, Li X, Li J. Emotion recognition using different sensors, emotion models, methods and datasets: a comprehensive review. Sensors. 2023;23(5):2455. doi:10.3390/s23052455.
3. Ghaleb E, Popa M, Asteriadis S, editors. Multimodal and temporal perception of audio-visual cues for emotion recognition. In: Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII); 2019 Sep 3–6; Cambridge, UK. doi:10.1109/acii.2019.8925444.
4. Rehman AU, Ali Naqvi R, Rehman A, Paul A, Sadiq MT, Hussain D. A trustworthy SIoT aware mechanism as an enabler for citizen services in smart cities. Electronics. 2020;9(6):918. doi:10.3390/electronics9060918.
5. Shafiq M, Fan Q, Alghamedy FH, Obidallah WJ. DualEye-FeatureNet: a dual-stream feature transfer framework for multi-modal ophthalmic image classification. IEEE Access. 2024;12:143985–4008. doi:10.1109/access.2024.3469244.
6. Wu D, Zhang J, Zhao Q. Multimodal fused emotion recognition about expression-EEG interaction and collaboration using deep learning. IEEE Access. 2020;8:133180–9. doi:10.1109/access.2020.3010311.
7. Liu H, Lou T, Zhang Y, Wu Y, Xiao Y, Jensen CS, et al. EEG-based multimodal emotion recognition: a machine learning perspective. IEEE Trans Instrum Meas. 2024;73(8):1–29. doi:10.1109/tim.2024.3369130.
8. Almanza-Conejo O, Almanza-Ojeda DL, Contreras-Hernandez JL, Ibarra-Manzano MA. Emotion recognition in EEG signals using the continuous wavelet transform and CNNs. Neural Comput Appl. 2023;35(2):1409–22. doi:10.1007/s00521-022-07843-9.
9. Mutawa AM, Hassouneh A. Multimodal Real-Time patient emotion recognition system using facial expressions and brain EEG signals based on Machine learning and Log-Sync methods. Biomed Signal Process Control. 2024;91(6):105942. doi:10.1016/j.bspc.2023.105942.
10. Nakisa B, Rastgoo MN, Rakotonirainy A, Maire F, Chandran V. Automatic emotion recognition using temporal multimodal deep learning. IEEE Access. 2020;8:225463–74. doi:10.1109/access.2020.3027026.
11. Jirayucharoensak S, Pan-Ngum S, Israsena P. EEG-based emotion recognition using deep learning network with principal component based covariate shift adaptation. Sci World J. 2014;2014(2):627892. doi:10.1155/2014/627892.
12. Riyad M, Khalil M, Adib A, editors. Cross-subject EEG signal classification with deep neural networks applied to motor imagery. In: Mobile, secure, and programmable networking. Cham, Switzerland: Springer International Publishing; 2019. p. 124–39. doi:10.1007/978-3-030-22885-9_12.

13. Yang H, Han J, Min K. A multi-column CNN model for emotion recognition from EEG signals. Sensors. 2019;19(21):4736. doi:10.3390/s19214736.
14. Ghaleb E, Popa M, Asteriadis S. Metric learning based multimodal audio-visual emotion recognition. IEEE Multimed. 2019;27(1):37–48. doi:10.1109/mmul.2019.2960219.
15. Cho J, Hwang H. Spatio-temporal representation of an electoencephalogram for emotion recognition using a three-dimensional convolutional neural network. Sensors. 2020;20(12):3491. doi:10.3390/s20123491.
16. Tang H, Liu W, Zheng WL, Lu BL, editors. Multimodal emotion recognition using deep neural networks. In: Neural information processing. Cham, Switzerland: Springer International Publishing; 2017. p. 811–9. doi:10.1007/978-3-319-70093-9_86.
17. Bao LQ, Qiu JL, Tang H, Zheng WL, Lu BL, editors. Investigating sex differences in classification of five emotions from EEG and eye movement signals. In: Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2019 Jul 23–27; Berlin, Germany. doi:10.1109/embc.2019.8857476.
18. Bilal A, Sun G, Li Y, Mazhar S, Latif J. Lung nodules detection using grey wolf optimization by weighted filters and classification using CNN. J Chin Inst Eng. 2022;45(2):175–86. doi:10.1080/02533839.2021.2012525.
19. Zhou J, Wei X, Cheng C, Yang Q, Li Q. Multimodal emotion recognition method based on convolutional auto-encoder. Int J Comput Intell Syst. 2018;12(1):351–8. doi:10.2991/ijcis.2019.125905651.
20. Pei G, Li H, Lu Y, Wang Y, Hua S, Li T. Affective computing: recent advances, challenges, and future trends. Intell Comput. 2024;3(5):76. doi:10.34133/icomputing.0076.
21. Wu Y, Mi Q, Gao T. A comprehensive review of multimodal emotion recognition: techniques, challenges, and future directions. Biomimetics. 2025;10(7):418. doi:10.3390/biomimetics10070418.
22. Shafiq M, Ayub S, Muthevi AK, Prabhu MR. AI-driven Life Cycle Assessment for sustainable hybrid manufacturing and remanufacturing. Int J Adv Manuf Technol. 2024;184(3):1002. doi:10.1007/s00170-024-14930-9.
23. Sharma A, Sharma K, Kumar A. Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion. Neural Comput Appl. 2023;35(31):22935–48. doi:10.1007/s00521-022-06913-2.
24. Zhang Y, Cheng C, Zhang Y. Multimodal emotion recognition using a hierarchical fusion convolutional neural network. IEEE Access. 2021;9:7943–51. doi:10.1109/access.2021.3049516.
25. Scherer KR. What are emotions? And how can they be measured? Soc Sci Inf. 2005;44(4):695–729. doi:10.1177/0539018405058216.
26. Ekman P, Friesen WV, O'Sullivan M, Chan A, Diacoyanni-Tarlatzis I, Heider K, et al. Universals and cultural differences in the judgments of facial expressions of emotion. J Pers Soc Psychol. 1987;53(4):712–7. doi:10.1037/0022-3514.53.4.712.
27. Russell JA. A circumplex model of affect. J Pers Soc Psychol. 1980;39(6):1161–78. doi:10.1037/h0077714.
28. Cimtay Y, Ekmekcioglu E. Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset EEG emotion recognition. Sensors. 2020;20(7):2034. doi:10.3390/s20072034.
29. Ramzan M, Dawn S. Fused CNN-LSTM deep learning emotion recognition model using electroencephalography signals. Int J Neurosci. 2023;133(6):587–97. doi:10.1080/00207454.2021.1941947.
30. Shafiq M, Thakre K, Krishna KR, Robert NJ, Kuruppath A, Kumar D. Continuous quality control evaluation during manufacturing using supervised learning algorithm for Industry 4.0. Int J Adv Manuf Technol. 2023;117(3):408. doi:10.1007/s00170-023-10847-x.
31. Soleymani M, Lichtenauer J, Pun T, Pantic M. A multimodal database for affect recognition and implicit tagging. IEEE Trans Affect Comput. 2012;3(1):42–55. doi:10.1109/t-affc.2011.25.
32. Rosenberg EL, Ekman P. What the face reveals: basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford, UK: Oxford University Press; 2020.
33. Alghamedy FH, Shafiq M, Liu L, Yasin A, Ali Khan R, Mohammed HS. Machine learning-based multimodel computing for medical imaging for classification and detection of Alzheimer disease. Comput Intell Neurosci. 2022;2022(6):9211477. doi:10.1155/2022/9211477.
34. Yu H, Liu H. Regression-based facial expression optimization. IEEE Trans Human Mach Syst. 2014;44(3):386–94. doi:10.1109/thms.2014.2313912.

35. Araújo JDL, da Cruz LB, Diniz JOB, Ferreira JL, Silva AC, de Paiva AC, et al. Liver segmentation from computed tomography images using cascade deep learning. Comput Biol Med. 2022;140(8):105095. doi:10.1016/j.compbiomed.2021.105095.
36. Rajan S, Chenniappan P, Devaraj S, Madian N. Novel deep learning model for facial expression recognition based on maximum boosted CNN and LSTM. IET Image Process. 2020;14(7):1373–81. doi:10.1049/iet-ipr.2019.1188.
37. Huang H, Hu Z, Wang W, Wu M. Multimodal emotion recognition based on ensemble convolutional neural network. IEEE Access. 2019;8:3265–3271.
38. Joshi V. M, Ghongade R. B. Optimal number of electrode selection for EEG based emotion recognition using linear formulation of differential entropy. Biomed Pharmacol J. 2020;13(2):645–53.
39. Zhao Y, Cao X, Lin J, Yu D, Cao X. Multimodal emotion recognition model using physiological signals. arXiv:1911.12918. 2019.
40. Xu D, Qin X, Dong X, Cui X. Emotion recognition of EEG signals based on variational mode decomposition and weighted cascade forest. Math Biosci Eng. 2022;20(2):2566–87. doi:10.3934/mbe.2023120.
41. Wang T, Huang X, Xiao Z, Cai W, Tai Y. EEG emotion recognition based on differential entropy feature matrix through 2D-CNN-LSTM network. EURASIP J Adv Signal Process. 2024;2024(1):49. doi:10.1186/s13634-024-01146-y.
42. Alidoost Y, Mohammadzadeh Asl B. Entropy-based emotion recognition using EEG signals. IEEE Access. 2025;13(6):51242–54. doi:10.1109/access.2025.3553809.
43. Liu R, Chao Y, Ma X, Sha X, Sun L, Li S, et al. ERTNet: an interpretable transformer-based framework for EEG emotion recognition. Front Neurosci. 2024;18:1320645. doi:10.3389/fnins.2024.1320645.
44. Choudhury N, Das D, Deka D, Ghosh R, Deb N, Ghaderpour E. NeuroFeat: an adaptive neurological EEG feature engineering approach for improved classification of major depressive disorder. Biomed Signal Process Control. 2026;113(3):109031. doi:10.1016/j.bspc.2025.109031.
45. Chen W, Liao Y, Dai R, Dong Y, Huang L. EEG-based emotion recognition using graph convolutional neural network with dual attention mechanism. Front Comput Neurosci. 2024;18: 1416494. doi:10.3389/fncom.2024.1416494.
46. Kumfor F, Landin-Romero R, Devenney E, Hutchings R, Grasso R, Hodges JR, et al. On the right side? A longitudinal study of left-versus right-lateralized semantic dementia. Brain. 2016;139(3):986–98. doi:10.1093/brain/awv387.
47. Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, et al. DEAP: a database for emotion analysis; using physiological signals. IEEE Trans Affect Comput. 2012;3(1):18–31. doi:10.1109/t-affc.2011.15.
48. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imag. 2018;9(4):611–29. doi:10.1007/s13244-018-0639-9.
49. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. doi:10.1109/cvpr.2016.90.
50. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90. doi:10.1145/3065386.
51. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. doi:10.1109/cvpr.2015.7298594.
52. Zhang C, Ma G, Zhang L, Shihada B. Graph neural networks empowered origin-destination learning for urban traffic prediction. CAAI Trans Intel Tech. 2025;10(4):1062–76. doi:10.1049/cit2.70021.
53. Hjorth B. EEG analysis based on time domain properties. Electroencephalogr Clin Neurophysiol. 1970;29(3):306–10. doi:10.1016/0013-4694(70)90143-4.
54. Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol. 2000;278(6):H2039–49. doi:10.1152/ajpheart.2000.278.6.h2039.
55. Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Machine Intell. 2002;24(7):971–87. doi:10.1109/tpami.2002.1017623.

56. Choi DY, Kim DH, Song BC. Multimodal attention network for continuous-time emotion recognition using video and EEG signals. IEEE Access. 2020;8:203814–26. doi:10.1109/access.2020.3036877.
57. Li M, Qiu M, Kong W, Zhu L, Ding Y. Fusion graph representation of EEG for emotion recognition. Sensors. 2023;23(3):1404. doi:10.3390/s23031404.
58. Cimtay Y, Ekmekcioglu E, Caglar-Ozhan S. Cross-subject multimodal emotion recognition based on hybrid fusion. IEEE Access. 2020;8:168865–78. doi:10.1109/access.2020.3023871.
59. Zhang H. Expression-EEG based collaborative multimodal emotion recognition using deep AutoEncoder. IEEE Access. 2020;8:164130–43. doi:10.1109/access.2020.3021994.
60. Huang Y, Yang J, Liu S, Pan J. Combining facial expressions and electroencephalography to enhance emotion recognition. Future Internet. 2019;11(5):105. doi:10.3390/fi11050105.
61. Hassan CAU, Ehatisham-ul-Haq M, Murtaza F, Ullah Yasin A, Sajid Ullah S. EmoTrans attention based emotion recognition using EEG signals and facial analysis with expert validation. Sci Rep. 2025;15(1):22004.