Emotion Recognition Using Machine Learning and Computer Vision: A Hybrid CNN–ViT Multimodal Framework

Deepa Chandrashekhar Rathod, Bhumika B K, Arpitha G A, Brunda U Jajur, Usha K

Department of Computer Science and Engineering, Jain Institute of Technology, Davangere

DOI: https://dx.doi.org/10.51244/IJRSI.2026.1304000249

Received: 22 April 2026; Accepted: 27 April 2026; Published: 19 May 2026

ABSTRACT

Automatic recognition of human emotions from facial expressions and multimodal signals constitutes a foundational challenge in affective computing and human–computer interaction, with broad applications spanning healthcare monitoring, autonomous vehicle safety, educational technology, and social robotics. Despite remarkable progress driven by deep learning, particularly convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Vision Transformers (ViT), achieving robust emotion recognition in unconstrained, real-world environments remains an open problem. This paper presents a comprehensive synthesis of over twenty-five state-of-the-art studies on facial and multimodal emotion recognition, encompassing CNN-based systems trained on FER2013, CK+, RAF-DB, and AffectNet; transformer-based hybrid architectures; and multimodal fusion systems integrating facial, speech, and electroencephalography (EEG) cues evaluated on RAVDESS, IEMOCAP, CMU-MOSEI, eNTERFACE'05, and MAHNOB-HCI. Building upon these insights, this work proposes a novel Hybrid CNN–ViT Multimodal Emotion Recognition (HCV-MER) framework comprising: (i) a squeeze-and-excitation ResNet combined with a Vision Transformer facial backbone incorporating region-specific attention over eyes and mouth; (ii) a lightweight temporal aggregation unit for video-level inference; and (iii) a cross-modal attention fusion module integrating facial and speech streams. Experimental evaluations target FER2013, RAF-DB, CK+, and RAVDESS using TensorFlow, PyTorch, and OpenCV. Expected improvements over baseline CNN architectures range from five to ten percentage points on challenging in-the-wild benchmarks. The paper further analyzes unresolved challenges including cross-domain generalization, demographic fairness, micro-expression recognition, and privacy-preserving deployment.

Keywords — Facial emotion recognition; Convolutional neural networks; Vision Transformer; Multimodal fusion; Affective computing; Human–computer interaction; Deep learning; FER2013; RAVDESS.

INTRODUCTION

Human emotions are complex psychological and physiological states that profoundly influence behavior, decision-making, and social interaction. The ability of machines to automatically perceive, interpret, and respond to human affective states—a capability broadly termed affective computing—has emerged as a critical frontier in the development of intelligent, human-centric systems [1]. Facial emotion recognition (FER), which involves the automated inference of emotional states from visual representations of the human face, occupies a particularly prominent position within this field due to the central communicative role of facial expressions and the scalability of visual data acquisition [6][16].

The practical significance of FER extends across numerous high-impact application domains. In healthcare, automated emotion monitoring facilitates the assessment of neurological and psychiatric conditions, assists in pain detection for non-verbal patients, and supports mental health interventions [1][18]. In automotive contexts, driver monitoring systems utilize real-time emotion and attention analysis to preemptively address fatigue or distraction, thereby enhancing road safety [15]. Educational technologies leverage student affect recognition to tailor instruction and detect disengagement, while social robotics and human–robot interaction (HRI) systems employ FER to enable more natural, empathetic agent behaviors [1][11][15]. Beyond these, applications in surveillance, security, and affective entertainment continue to motivate sustained research investment [8][13].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Early FER systems followed a structured three-stage pipeline consisting of face detection, handcrafted feature extraction, and classification. Prominent feature extraction methodologies included Local Binary Patterns (LBP), Gabor filter banks, Principal Component Analysis (PCA), and optical flow descriptors, which were subsequently fed into discriminative classifiers such as Support Vector Machines (SVMs), AdaBoost, or random forests [8][16]. While these approaches demonstrated satisfactory performance under controlled laboratory conditions, they exhibited fundamental limitations in terms of representational capacity and robustness to real-world variability, including uncontrolled illumination, occlusion, non-frontal head pose, and significant inter-subject diversity [6][8][16].

The emergence and rapid proliferation of deep learning have precipitated a paradigm shift in FER research. Convolutional neural networks, endowed with the ability to learn rich hierarchical feature representations directly from raw pixel data, have systematically supplanted handcrafted approaches on standard benchmarks [2][6][11]. Subsequent advances introduced CNN–RNN hybrids that explicitly model temporal dynamics in video sequences, recurrent architectures such as Long Short-Term Memory (LSTM) and Bidirectional LSTM networks capturing emotion-relevant motion patterns [2][16]. More recently, Vision Transformers have demonstrated compelling performance by modeling long-range dependencies across facial regions through self-attention mechanisms, offering a complementary perspective to the locally-biased inductive priors of CNNs [16][20].

Despite achieving accuracies exceeding 95% on controlled benchmark datasets such as CK+, state-of-the-art systems still exhibit notable performance degradation when confronted with the unpredictability of real-world conditions [15]. Furthermore, facial expressions alone may constitute an incomplete and potentially misleading index of emotional state, motivating the development of multimodal emotion recognition systems that integrate complementary cues from speech prosody, physiological signals such as EEG, or combinations thereof [5][10][12][19]. Such multimodal systems consistently outperform their unimodal counterparts, particularly when fusion strategies are designed to leverage the complementarity and contextual dependencies between modalities.

This paper makes the following primary contributions: (i) a structured synthesis of more than twenty-five recent works spanning CNN-based FER, transformer-based FER, and multimodal emotion recognition, organized to reveal methodological trends, benchmark performance trajectories, and persistent challenges; (ii) an explicit identification of research gaps derived from cross-study analysis; and (iii) the formulation of a novel Hybrid CNN–ViT Multimodal Emotion Recognition (HCV-MER) framework that integrates a region-attentive facial backbone, a lightweight temporal module, and a cross-modal attention fusion architecture, targeting improved generalization and practical deployability.

RELATED WORK

Traditional Computer Vision Approaches

Prior to the deep learning era, FER systems were architecturally constrained by the representational limits of handcrafted features. The standard pipeline comprised three sequential stages: face detection and localization, low-level or mid-level feature extraction, and classification [8][16]. Face detection commonly relied on Haar cascade classifiers or Viola-Jones detectors [9][13]. Feature extraction methodologies spanned a broad spectrum, including Local Binary Patterns for encoding local texture statistics, Gabor filter banks sensitive to orientation and frequency, PCA for dimensionality reduction, and optical flow for capturing motion across temporal sequences [8][16]. These features were subsequently processed by discriminative classifiers—most commonly SVMs, AdaBoost ensembles, or random forests.

Canedo and Neves [8] conducted a systematic review of these conventional FER pipelines and concluded that, while accuracy on controlled datasets could be acceptable, performance degraded substantially under uncontrolled conditions characterized by spontaneous expressions, illumination changes, and pose variation. This body of work established the fundamental necessity of more expressive representation learning, ultimately motivating the transition to end-to-end deep learning approaches [6][8].

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

CNN-Based Facial Emotion Recognition

Convolutional neural networks have become the predominant paradigm for FER, owing to their capacity for hierarchical, task-adaptive feature learning from raw image data without the need for manual feature engineering [2][6][11][16]. Large-scale systematic reviews, including those by Cirneanu et al. [6] covering 155 papers and Karnati et al. [2], confirm that CNNs and their architectural variants account for the overwhelming majority of high-performing FER systems.

Standard CNN architectures achieve approximately 64–75% accuracy on the challenging FER2013 dataset, which contains approximately 35,000 in-the-wild grayscale images annotated across seven basic emotion categories [9][15][17]. On the more controlled CK+ dataset comprising laboratory-captured facial sequences, accuracies routinely exceed 95% [15][16]. Intermediate performance is observed on RAF-DB, a real-world dataset with diverse subjects and imaging conditions, where Huang et al. [11] achieved 83.37% validation accuracy using a squeeze-and-excitation ResNet (SE-ResNet) pre-trained on AffectNet and fine-tuned via transfer learning. This study additionally identified the nose and mouth regions as the most diagnostically critical areas for discriminating among emotion categories [11].

Transfer learning and systematic hyperparameter optimization have proven particularly effective strategies for overcoming dataset size limitations. Agung et al. [3] demonstrated this approach on the Emognition dataset, which contains ten fine-grained emotion categories including complex affective states such as amusement and awe. Using fine-tuned Inception-V3 and MobileNet-V2 models with Taguchi-optimized hyperparameters, they achieved 96% accuracy and F1=0.95, illustrating that even highly nuanced emotion categories can be distinguished with appropriate model selection and optimization [3]. Similarly, Raj and Demirkol [15] employed CNN architectures augmented with data augmentation and transfer learning, reaching up to 95% on CK+ while highlighting the persistent accuracy gap when the same models are evaluated on FER2013, underscoring the fundamental challenge of cross-dataset generalization.

Real-time FER applications represent a practically significant segment of the CNN literature. Systems integrating CNNs with OpenCV-based face detection—typically using Haar cascades—for live video processing have demonstrated around 80–85% accuracy on FER2013 while maintaining sufficiently low latency for deployment on edge devices [4][7][9][13]. Juniawan et al. [9] further incorporated Grad-CAM visualization to provide interpretable insight into network decision-making, identifying which facial regions most strongly activate the network for each predicted emotion class.

Recurrent and Hybrid CNN–RNN Models

Static frame-level FER ignores the temporal dynamics of facial expression evolution, which carry meaningful information about the nature and intensity of emotional experiences. To address this limitation, hybrid architectures combining CNNs for spatial feature extraction with RNNs—particularly LSTMs and Bidirectional LSTMs—for temporal sequence modeling have been extensively explored [2][16]. These CNN–LSTM architectures process sequences of per-frame CNN feature vectors through recurrent units that maintain and update a hidden state across time, enabling the capture of motion-related cues and expression dynamics.

Reviews by Karnati et al. [2] and Ko [16] confirm that CNN–LSTM hybrids generally outperform static CNNs when sufficient temporally-labeled video data is available. However, the associated computational cost and the relative scarcity of densely annotated video FER datasets present practical barriers to widespread adoption. These limitations have motivated the search for more computationally efficient alternatives, including temporal attention pooling and lightweight temporal aggregation strategies [2][16][18].

Vision Transformers for FER

The application of Transformer architectures to computer vision—originally introduced as ViT by Dosovitskiy et al.—has opened compelling new directions for FER [16][20]. Unlike CNNs, which are inherently biased toward local spatial patterns through their convolutional inductive priors, Vision Transformers employ multi-

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

head self-attention to model global contextual relationships across the entire image or feature map, enabling flexible, long-range dependency capture between spatially distant facial regions.

Chaudhari et al. [20] proposed ViTFER, a hybrid architecture combining a ResNet-18 backbone for local feature extraction with a Vision Transformer encoder for global relational reasoning. To address the class imbalance prevalent in individual FER datasets, they constructed AVFER, a composite dataset merging FER2013, AffectNet, and CK+, and demonstrated that their hybrid model achieved superior classification performance over prior CNN-only baselines on this unified benchmark. Systematic reviews [16] confirm that ViT and hybrid CNN–ViT architectures represent an emerging frontier in FER, though they characteristically require large-scale training data and substantial computational resources, presenting accessibility challenges for resource-constrained research settings.

Multimodal Emotion Recognition

Recognizing that the face provides only one window into emotional experience, and that vocal prosody, linguistic content, physiological signals, and bodily posture each carry complementary affective information, multimodal emotion recognition has attracted growing attention. Multimodal systems aim to exploit the complementarity and mutual reinforcement among modalities to achieve more accurate, robust, and generalizable emotion inference than is possible from any single modality in isolation.

Jiménez et al. [5] developed an audio-visual multimodal system for emotion recognition on the RAVDESS dataset, combining a pre-trained Spatial Transformer Network (STN) coupled with a Bidirectional LSTM for facial video processing with a transfer-learned PANNs CNN-14 model for speech spectrogram analysis. Late fusion of the two modality streams yielded 80.08% accuracy over eight emotion categories, representing a substantial improvement over unimodal baselines and demonstrating the practical utility of transfer learning in data-constrained multimodal settings. Mamieva et al. [10] further advanced multimodal fusion through an attention-based architecture that learned to dynamically weight the contributions of facial and speech feature encoders. Applied to IEMOCAP and CMU-MOSEI, this approach achieved weighted accuracies of 74.6% and 80.7%, respectively, with substantial F1-score improvements attributable to the attention mechanism's ability to suppress modality-specific noise.

The question of optimal fusion strategy—whether to combine modalities at the feature level or at the decision level—has been investigated by Tomar et al. [12], who evaluated both paradigms on the eNTERFACE'05 dataset. Their findings conclusively favored feature-level fusion, which enables the model to jointly reason over complementary representations from both modalities rather than independently committing to modality-specific decisions before combination. Pan et al. [19] introduced Deep-Emotion, a comprehensive three-branch architecture integrating an improved GhostNet for efficient facial processing, a lightweight fully convolutional network (LFCNN) for speech spectrogram encoding, and a tree-structured LSTM (tLSTM) for EEG sequence modeling. Decision-level fusion of all three branches on CK+, EMO-DB, and MAHNOB-HCI demonstrated state-of-the-art performance, establishing the feasibility and advantage of higher-order multimodal integration.

Surveys and Identified Open Issues

Multiple comprehensive surveys and systematic literature reviews provide meta-level perspectives on the FER and multimodal emotion recognition landscape [1][2][6][8][16][18]. Collectively, these works identify a consistent set of open challenges. The scarcity of large, balanced, and demographically diverse datasets—particularly for non-basic emotions and micro-expressions—constitutes a fundamental bottleneck for supervised learning approaches [1][2][6][16][20]. The performance gap between controlled laboratory datasets such as CK+ and in-the-wild benchmarks such as FER2013 and AffectNet remains substantial, pointing to unresolved issues of domain shift and distribution mismatch [6][8][11][15]. Technical challenges including occlusion, significant head pose variation, illumination inhomogeneity, and individual identity bias compound the difficulty of real-world recognition [2][6][8][11][16]. Finally, the deployment of emotion recognition systems in high-stakes, sensitive domains raises pressing concerns regarding demographic bias, algorithmic fairness, data privacy, and informed consent, which remain inadequately addressed in the literature [8][18].

Research Gap

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Despite the substantial body of research surveyed, a critical analysis reveals several persistent and under addressed gaps. First, while CNN-based FER has reached impressive performance on individual benchmarks, cross-dataset generalization remains poor. Models trained on controlled datasets such as CK+ frequently fail to maintain accuracy when evaluated on in-the-wild datasets like FER2013, with performance drops of more than 30 percentage points being commonly reported [6][8][11][15]. This domain gap indicates that current architectures are prone to overfitting to dataset-specific biases in lighting, subject demographics, expression elicitation conditions, and annotation protocols.

Second, while hybrid CNN–ViT architectures have shown promise for FER by capturing both local texture features and global structural relationships, the systematic integration of region-specific attention—explicitly guided toward diagnostically informative facial regions such as the eyes and mouth—remains underexplored [11][20]. Existing attention mechanisms in the FER literature are largely learned without explicit spatial priors, potentially resulting in suboptimal allocation of model capacity.

Third, multimodal emotion recognition systems, though consistently superior to unimodal approaches, have yet to achieve the full potential of cross-modal complementarity. Most existing systems employ either simple concatenation or late decision-level fusion, which do not permit rich interactions between complementary feature spaces [5][10][12][19]. Cross-modal attention mechanisms that allow each modality stream to condition its representations on contextual information from other streams remain underutilized in this domain.

Fourth, many state-of-the-art systems prioritize benchmark accuracy at the expense of computational efficiency, limiting real-world deployability on resource-constrained edge platforms [4][9][19]. The simultaneous achievement of high accuracy, robustness, and real-time efficiency represents an insufficiently addressed engineering challenge. Finally, fairness, demographic generalization, and privacy-preserving deployment are acknowledged as critical concerns in surveys [8][18] but receive minimal empirical treatment in primary research, leaving a significant gap between system capabilities and responsible deployment requirements.

Proposed Method: HCV-MER Framework

To address the identified research gaps, this paper proposes the Hybrid CNN–ViT Multimodal Emotion Recognition (HCV-MER) framework, a unified deep learning architecture comprising three principal components: (i) a region-attentive CNN–ViT facial recognition stream; (ii) a lightweight temporal aggregation module for video-level inference; and (iii) a cross-modal attention fusion module for integrating facial and speech representations.

CNN–ViT Facial Stream

The facial recognition stream processes input images of size 224×224×3 through a four-stage pipeline. In Stage 1, a squeeze-and-excitation ResNet-18 (SE-ResNet-18) backbone extracts rich hierarchical spatial features, producing a feature map $F \in R^{h\times w\times d}$. The squeeze-and-excitation mechanism explicitly recalibrates channel-wise feature responses, enabling the backbone to emphasize expression-relevant channels while suppressing uninformative ones [11]. In Stage 2, the resulting feature map is partitioned into non-overlapping patches of fixed size, which are then flattened and linearly projected into a sequence of patch embeddings $Z_0 \in R^{P\times D}$, following the ViT patch embedding paradigm [20].

Stage 3 applies a multi-layer Vision Transformer encoder with multi-head self-attention to the patch embedding sequence. This encoder captures long-range contextual dependencies among facial regions, allowing the model to relate the state of the eye region to that of the mouth or brow, even across large spatial distances in the image. A learnable class token appended to the sequence aggregates global emotion-discriminative information. Stage 4 incorporates a Region Attention Regularization mechanism: spatial attention maps produced by the transformer encoder are subject to a regularization loss that penalizes insufficient attention to the eye and mouth regions, which have been empirically identified as the most diagnostically informative facial areas [11][20]. This spatial prior serves to focus the model's representational capacity on affectively relevant anatomy.

Temporal Aggregation Module

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

For video-level emotion recognition, per-frame CNN–ViT embeddings are aggregated across time using a lightweight temporal attention mechanism. Rather than processing the full temporal sequence through a recurrent unit, which incurs substantial computational overhead, the temporal module learns a set of attention weights over the frame sequence, producing a single temporally-weighted facial embedding that emphasizes the most emotionally salient frames while down-weighting transitional or neutral frames [2][5][16]. This design balances temporal expressiveness with computational efficiency, supporting real-time deployment without sacrificing meaningful temporal modeling.

Speech Stream

The speech recognition stream accepts log-Mel spectrogram representations of aligned speech segments as input. A pre-trained PANNs CNN-14 backbone [5]—fine-tuned on emotion labels—extracts a sequence of high-level speech feature vectors encoding prosodic, rhythmic, and spectral characteristics associated with emotional expression. Attention-based temporal pooling over the speech feature sequence produces a fixed-dimensional speech embedding $s \in R^{d_s}$ that emphasizes temporally salient acoustic events [10][19].

Cross-Modal Attention Fusion

The facial embedding $v \in R^{d_v}$ and speech embedding $s \in R^{d_s}$ are independently projected to a common embedding space of dimension d via learned linear transformations. The projected embeddings are concatenated and passed through a multi-head cross-modal attention module, which learns to dynamically weight and recombine features from both modalities based on their mutual contextual relevance. Formally:

$$z = \mathrm{Attn}([W_v \cdot v ; W_s \cdot s])$$

where W_v and W_s are projection matrices, $[\cdot;\cdot]$ denotes concatenation, and Attn is a multi-head attention operation [10][12]. The fused representation z is subsequently processed by a fully connected classification head with softmax activation to produce emotion posterior probabilities. The entire network is trained end-to-end using a class-balanced cross-entropy loss to mitigate the impact of label imbalance prevalent across FER datasets [3][20].

METHODOLOGY

Datasets

Five benchmark datasets are employed in this study. FER2013 consists of approximately 35,000 48×48 grayscale facial images distributed across seven basic emotion categories (angry, disgust, fear, happy, neutral, sad, surprise), collected in unconstrained conditions and exhibiting substantial inter-class overlap, mislabeling, and resolution limitations [9][15][17]. The CK+ (Extended Cohn-Kanade) dataset provides 593 laboratory-elicited image sequences from 123 subjects, with the peak expression frame of each sequence carrying a validated categorical label; this dataset is widely used for benchmarking due to its high labeling quality, yielding consistently high reported accuracies [15][16][19].

RAF-DB (Real-world Affective Faces Database) contains 15,339 real-world facial images with compound and basic emotion labels, annotated by multiple independent raters to ensure label reliability [11][15]. AffectNet is a large-scale in-the-wild database comprising over one million facial images with both categorical and dimensional (valence-arousal) labels, commonly used for large-scale pre-training [2][6][11][20]. RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) provides 7,356 audio-visual recordings of 24 professional actors portraying eight emotion categories, serving as the primary benchmark for multimodal audio-visual evaluation [5].

Data Preprocessing

All input facial images undergo a standardized preprocessing pipeline. Face detection and localization are performed using a Multi-task Cascaded Convolutional Network (MTCNN) detector for images and OpenCV DNN-based detection for real-time video streams [7][9][13]. Detected face regions are cropped, aligned based

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

on detected facial landmarks (eye centers and nose tip), and resized to 224×224 pixels for the CNN–ViT backbone. Images are normalized to zero mean and unit variance per channel using ImageNet statistics to facilitate pre-trained backbone fine-tuning.

Data augmentation is applied during training to improve robustness to real-world variability. Augmentation transformations include random horizontal flipping, random rotation within ±15 degrees, random brightness and contrast perturbation, and random affine scaling [3][9][11][15]. For speech inputs, log-Mel spectrograms are computed with 128 Mel filter banks over 25ms windows with 10ms hop length. Spectrogram augmentation employs SpecAugment—specifically frequency masking and time masking—to simulate acoustic variability [5][10][19].

Feature Extraction

The SE-ResNet-18 backbone extracts spatial features at multiple levels of abstraction through its convolutional stack. Squeeze-and-excitation blocks inserted after each residual stage explicitly recalibrate channel-wise feature responses through a global average pooling squeeze operation followed by a two-layer fully connected excitation network, producing channel attention weights that multiplicatively gate the residual features [11]. This mechanism has been empirically shown to enhance discriminability for subtle expression differences by amplifying emotion-relevant channel activations.

For speech processing, the PANNs CNN-14 backbone—a deep 2D CNN pre-trained on AudioSet—is fine-tuned on emotion-annotated speech data. This backbone processes log-Mel spectrograms through fourteen convolutional layers with batch normalization and ReLU activations, extracting rich acoustic feature representations encoding spectral shape, temporal rhythm, and energy distribution—all of which correlate with emotional prosody [5][19].

Model Architecture

The complete HCV-MER architecture integrates the four components described in Section IV. The CNN facial backbone (SE-ResNet-18) produces a spatial feature map that is divided into 7×7=49 patches, each of dimension d=512, yielding a patch embedding sequence of length 50 (including the class token). The ViT encoder comprises six transformer layers, each with eight attention heads and a feed-forward dimension of 2048. The temporal attention module is a single-layer attention mechanism with learnable query vector operating over the sequence of T=16 frame-level class token embeddings. The speech encoder (PANNs CNN-14) produces frame-level acoustic features of dimension 2048, which are pooled to a fixed 512-dimensional embedding through attention-weighted temporal pooling. Cross-modal fusion employs four attention heads operating over the 512-dimensional projected facial and speech embeddings. The classification head consists of two fully connected layers (512→256→K), where K is the number of target emotion classes.

Tools and Implementation

Model implementation is carried out in Python 3.10 using the PyTorch 2.0 and TensorFlow 2.13 frameworks [3][4][7][9][11][20]. Face detection and real-time video processing pipelines are implemented with OpenCV 4.8, utilizing Haar cascades for rapid detection and MTCNN for high-accuracy offline processing [7][9][13]. The PANNs speech backbone is accessed through the publicly available pretrained model repository. Model training is conducted on NVIDIA A100 GPUs with 40GB memory. Grad-CAM visualizations are generated using the pytorch-grad-cam library to provide interpretable attribution maps for model predictions [9].

Experimental Setup

The proposed HCV-MER framework is evaluated under two experimental configurations. In the unimodal FER configuration, only the CNN–ViT facial stream is active, and evaluations are performed on FER2013, CK+, and RAF-DB using the standard train/validation/test splits established in the FER literature. In the multimodal configuration, both facial and speech streams are active, and evaluations are performed on RAVDESS using a stratified five-fold cross-validation protocol.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

All models are optimized using the AdamW optimizer with an initial learning rate of 3×10^{-4}, a cosine annealing learning rate schedule, and a weight decay of 1×10^{-4}. Training employs a mini-batch size of 64 images. Early stopping is applied based on validation accuracy with a patience of 15 epochs. For cross-dataset generalization experiments, models pre-trained on AffectNet are fine-tuned on RAF-DB and FER2013 using layer-wise learning rate decay, with lower layers frozen for the first ten fine-tuning epochs.

To address class imbalance in FER2013 and RAF-DB, a class-balanced sampling strategy is employed during training, ensuring each class is represented with approximately equal frequency within each mini-batch [3][20]. The region attention regularization loss weight is set to $\lambda=0.1$, balancing classification accuracy against spatial attention guidance. Performance is evaluated using overall accuracy, weighted accuracy (WA), macro-averaged F1 score, per-class precision and recall, and confusion matrices, enabling a comprehensive multi-faceted comparison with baseline and prior-art systems [5][10][15][19][20].

RESULTS AND ANALYSIS

Table I presents a comparative summary of representative prior works alongside expected HCV-MER results, establishing the performance context within the FER and multimodal emotion recognition literature.

TABLE I. Comparative Overview of Existing and Proposed Emotion Recognition Systems

Study / Year	Modality	Model	Dataset(s)	Result	Key Contribution
Pereira et al. [1], 2024	Face, Pose	CNN, Faster R-CNN, ViT	77-paper SLR	Review	PRISMA taxonomy; data scarcity & hardware limits highlighted
Karnati et al. [2], 2023	Face	Deep CNNs, CNN-LSTM	Multiple FER	Review	Occlusion, pose, illumination, identity bias analysis
Agung et al. [3], 2024	Face (Video)	Inception-V3, MobileNet-V2	Emognition (10 classes)	96% Acc, F1=0.95	Transfer learning + Taguchi hyperparameter optimization
Raj & Demirkol [15], 2025	Face	CNN + CV Pipeline	FER2013, RAF-DB, CK+	95% (CK+), 64% (FER2013)	Real-time HRI; exposes lab-to-wild domain gap
Huang et al. [11], 2023	Face	SE-ResNet CNN	AffectNet, RAF-DB	83.37% (RAF-DB)	Nose & mouth identified as critical regions
Chaudhari et al. [20], 2022	Face	ResNet-18 + ViT (ViTFER)	AVFER (composite)	SOTA on AVFER	Hybrid ResNet-ViT; AVFER dataset for class imbalance
Cirneanu et al. [6], 2023	Face	CNNs, RNNs, GANs	155-paper SLR	Review	Endorses CNNs; covers applications and open issues
Ko [16], 2018	Face	CNN, CNN-LSTM	FER benchmarks	Review	Hybrid CNN-LSTM FER; micro-

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

					expression difficulty noted
Jiménez et al. [5], 2021	Face + Speech	STN + Bi-LSTM; PANNs CNN-14	RAVDESS	80.08% (8 emotions)	Multimodal TL with late fusion on audio-visual data
Mamieva et al. [10], 2023	Face + Speech	Attention-based fusion	IEMOCAP, CMU-MOSEI	WA 74.6% / 80.7%	Attention fusion of heterogeneous multimodal features
Tomar et al. [12], 2024	Face + Speech	Feature + Decision Fusion	eNTERFACE'05	Improved over unimodal	Feature-level fusion outperforms decision-level fusion
Pan et al. [19], 2023	Face + Speech + EEG	GhostNet, LFCNN, tLSTM	CK+, EMO-DB, MAHNOB-HCI	SOTA on multiple datasets	Three-branch multimodal with decision-level fusion
Juniawan et al. [9], 2024	Face (Real-time)	CNN + OpenCV	FER2013 + custom	85%	Real-time FER with Haar cascades and Grad-CAM
Jha et al. [17], 2025	Face + Identity	CNN + SVM/KNN	FER2013, JAFFE	74.78% / 98.65%	Integrated face recognition + FER; HCI interface design
Hassan et al. [18], 2025	Multi (Survey)	CNN, RNN, Transformers	Multiple	Review	Ethics, bias, privacy in emotion recognition reviewed

On FER2013, the proposed CNN–ViT facial stream is expected to achieve approximately 71–74% accuracy, representing an improvement of approximately 7–10 percentage points over standalone CNN baselines operating on the same data [9][15][17]. This improvement is attributed to the ViT encoder's capacity to integrate global facial context, complementing the local feature sensitivity of the SE-ResNet-18 backbone. The region attention regularization is expected to particularly benefit the recognition of subtle expressions such as fear and disgust, which are systematically confused by attention-agnostic models.

On CK+, the model is expected to achieve 96–98% accuracy, consistent with the consistently high performance of deep models on this controlled dataset and attributable primarily to the high quality and consistency of CK+ imagery [15][16]. On RAF-DB, the transfer learning protocol (AffectNet pre-training followed by RAF-DB fine-tuning) is projected to yield approximately 86–88% accuracy, surpassing the SE-ResNet baseline of 83.37% reported by Huang et al. [11] through the additional representational capacity of the ViT encoder and region attention mechanism.

For the multimodal configuration evaluated on RAVDESS, the HCV-MER framework is anticipated to achieve approximately 84–87% weighted accuracy over eight emotion categories. This projection is grounded in the 80.08% baseline established by Jiménez et al. [5] using a comparable multimodal late-fusion architecture, with expected gains attributable to the feature-level cross-modal attention fusion, which enables richer inter-modal interaction compared to decision-level combination. The attention module is expected to be especially effective for ambiguous cases where the facial expression and speech prosody carry complementary rather than redundant emotional cues.

Analysis of predicted confusion matrices indicates that the most frequent error patterns involve the fear–surprise and sad–neutral pairs, consistent with findings across the FER literature [3][8][9][15]. These confusions likely

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

reflect genuine perceptual ambiguity in the source images rather than model deficiencies per se. The region attention mechanism is expected to reduce confusion between disgust and anger by directing representational focus toward the mouth region, whose morphological changes are differentially diagnostic for these two categories [11].

DISCUSSION

Strengths of the Proposed Approach

The HCV-MER framework consolidates several independently validated strengths from the literature into a unified architecture. The combination of CNN spatial sensitivity and ViT global relational reasoning has been established as a productive hybrid strategy by ViTFER [20] and corroborated by systematic reviews [16]. The explicit spatial guidance through region attention regularization addresses a limitation identified by Huang et al. [11], who showed empirically that nose and mouth regions carry disproportionate discriminative value. Lightweight temporal attention aggregation strikes a pragmatic balance between the temporal modeling capabilities of full recurrent architectures and the computational constraints of real-time deployment [2][5][16].

The cross-modal attention fusion module represents the most significant architectural innovation of HCV-MER, departing from the simple late-fusion or concatenation strategies used in the majority of multimodal FER systems [5][12]. By enabling each modality's representation to attend to and selectively integrate information from the other, the fusion module can exploit subtle inter-modal complementarities—for instance, a neutral facial expression paired with a distressed vocal prosody—that static fusion strategies would fail to utilize effectively [10][12].

Limitations and Open Challenges

Several limitations of the current framework merit acknowledgment. The ViT encoder and cross-modal attention fusion collectively require substantial computational resources during training and inference, which may limit deployability on low-power edge devices without model compression, knowledge distillation, or quantization [4][9][19]. The framework's reliance on large, well-labeled training corpora—particularly AffectNet for pre-training—creates a dependency on resources that may be unavailable in low-resource languages or domains [1][2][16][18].

Cross-cultural generalization remains an underaddressed challenge. The facial expression conventions encoded in benchmark datasets such as FER2013 and AffectNet reflect predominantly Western, WEIRD (Western, Educated, Industrialized, Rich, Democratic) populations and expression elicitation paradigms, potentially limiting model applicability across demographically diverse user populations [6][8][18]. Preliminary analyses of FER model performance stratified by gender, age, and ethnicity consistently reveal performance disparities, underscoring the need for explicit fairness-aware training objectives and evaluation protocols [8][18].

The explainability of model decisions beyond Grad-CAM attribution maps—which are useful but provide only post-hoc, input-gradient-based explanations—remains limited. For deployment in sensitive contexts such as healthcare or judicial proceedings, more principled and certifiable forms of decision explainability are required [8][18].

Future Work

Several directions present compelling opportunities for extending the proposed framework. Micro-expression recognition—the detection of brief, involuntary, low-intensity facial movements lasting 40–500 milliseconds—represents one of the most challenging and consequential frontiers in FER [1][2][16]. Addressing this challenge will likely require high-frame-rate video capture, optical flow-based temporal descriptors sensitive to subtle motion, and self-supervised pre-training strategies that leverage the abundance of unlabeled video data [16][18].

Self-supervised and few-shot learning paradigms offer promising avenues for reducing the dependence of multimodal emotion recognition systems on large quantities of labeled training data. Contrastive learning

objectives—such as learning audio-visual correspondences from naturally paired speech and facial video—could substantially reduce the labeled data requirements for multimodal systems while potentially improving cross-modal representational alignment [16][18][20].

Cross-cultural and demographically fair FER is an urgent research priority. Future work should invest in constructing balanced, diverse datasets spanning multiple cultural backgrounds, age groups, and demographic identities, and should integrate fairness constraints—such as equalized odds or demographic parity objectives—directly into the training loss to mitigate discriminatory model behavior [6][8][18]. Privacy-preserving deployment through federated learning, differential privacy, and on-device inference deserves concerted engineering attention, particularly as emotion recognition systems are increasingly considered for deployment in regulated domains such as healthcare [18].

Finally, comprehensive user-centered evaluations in realistic deployment contexts—assessing not only model accuracy but also user comfort, perceived appropriateness, and practical utility in healthcare, education, and HRI applications—are essential for bridging the gap between benchmark performance and real-world impact [1][8][11][15][19].

CONCLUSION

This paper has presented a comprehensive review of the emotion recognition landscape, synthesizing methodological advances and benchmark results from over twenty-five recent studies spanning CNN-based FER, Vision Transformer approaches, and multimodal fusion systems. The analysis reveals a clear trajectory toward increasingly expressive architectures that combine the local sensitivity of convolutional networks with the global contextual reasoning of transformers, and toward multimodal systems that transcend the limitations of single-channel emotion inference.

Building upon identified research gaps—including cross-dataset generalization deficits, insufficient exploitation of region-specific facial anatomy, limitations of static fusion strategies, and the underaddressed intersection of performance and responsible deployment—this work proposed the Hybrid CNN–ViT Multimodal Emotion Recognition (HCV-MER) framework. The framework integrates a region-attentive SE-ResNet–ViT facial backbone, a computationally efficient temporal attention aggregation module, and a cross-modal attention fusion architecture combining facial and speech streams. Experimental projections suggest substantial improvements over CNN baselines on in-the-wild benchmarks and competitive multimodal performance on RAVDESS.

The broader challenge of robust, fair, and trustworthy emotion recognition in unconstrained real-world conditions remains open. Addressing it will require not only architectural innovation but also richer and more diverse datasets, principled fairness and privacy interventions, and sustained engagement between the research community, affected communities, and regulatory stakeholders. The authors hope that this synthesis and the accompanying framework contribute meaningfully to these ongoing efforts.

REFERENCES

1. R. Pereira et al., "Systematic Review of Emotion Detection with Computer Vision and Deep Learning," Sensors, vol. 24, 2024.
2. M. Karnati et al., "Understanding Deep Learning Techniques for Recognition of Human Emotions Using Facial Expressions: A Comprehensive Survey," IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1–28, 2023.
3. E. S. Agung et al., "Image-based facial emotion recognition using convolutional neural network on Emognition dataset," Scientific Reports, vol. 14, 2024.
4. S. Taru et al., "Emotion Recognition System from Facial Expressions Using Machine Learning," Journal of Artificial Intelligence and Capsule Networks, vol. 7, 2025.
5. C. L. Jiménez et al., "Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning," Sensors, vol. 21, no. 22, 2021.
6. Cirneanu et al., "New Trends in Emotion Recognition Using Image Analysis by Neural Networks, A Systematic Review," Sensors, vol. 23, 2023.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

7. M. Ranjani et al., "Emotion Recognition Using CNN," in Proc. ICOEI, 2025.
8. D. Canedo and A. J. R. Neves, "Facial Expression Recognition Using Computer Vision: A Systematic Review," Applied Sciences, vol. 9, no. 21, p. 4678, 2019.
9. G. Juniawan et al., "Real-Time Facial Emotion Detection Application with Image Processing Based on CNN," International Journal of Electrical Engineering, Mathematics and Computer Science, vol. 3, no. 1, 2024.
10. D. Mamieva et al., "Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features," Sensors, vol. 23, 2023.
11. Z. Huang et al., "A study on computer vision for facial emotion recognition," Scientific Reports, vol. 13, 2023.
12. P. S. Tomar et al., "Fusing facial and speech cues for enhanced multimodal emotion recognition," International Journal of Information Technology, vol. 16, 2024.
13. D. Shukla et al., "Human Face Detection and Emotion Recognition Using OpenCV through AI," in Proc. IEMECON, 2024.
14. Yushchenko et al., "Evaluating CNN, RNN, and Vision Transformer for Emotion Recognition: Strengths and Weaknesses," in Proc. IEEE eStream, 2025.
15. R. Raj and I. Demirkol, "An improved facial emotion recognition system using CNN for optimization of human robot interaction," Scientific Reports, vol. 15, 2025.
16. B. C. Ko, "A Brief Review of Facial Emotion Recognition Based on Visual Information," Sensors, vol. 18, no. 2, p. 401, 2018.
17. B. K. Jha et al., "Human Emotion Detection and Face Recognition System," International Journal on Engineering Technology, vol. 6, 2025.
18. S. E. A. Hassan et al., "Survey on Emotion Recognition Using Deep Learning," in Proc. ICEEM, 2025.
19. J.-Y. Pan et al., "Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG," IEEE Open Journal of Engineering in Medicine and Biology, vol. 4, 2023.
20. Chaudhari et al., "ViTFER: Facial Emotion Recognition with Vision Transformers," Applied System Innovation, vol. 5, no. 6, p. 117, 2022.
21. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021.
22. J. Hu et al., "Squeeze-and-Excitation Networks," in Proc. IEEE CVPR, 2018.
23. K. He et al., "Deep Residual Learning for Image Recognition," in Proc. IEEE CVPR, 2016.
24. G. Qiuqiang et al., "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880–2894, 2020.
25. R. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLOS ONE, vol. 13, no. 5, 2018.
26. P. Lucey et al., "The Extended Cohn-Kanade Dataset (CK+): A complete expression dataset for action unit and emotion-specified expression," in Proc. IEEE CVPRW, 2010.
27. S. Li et al., "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," in Proc. IEEE CVPR, 2017.
28. Mollahosseini et al., "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, 2019.