International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 3, Issue 3 July-Sept 2026 ISSN: 3048-975X
https://ijrasht.com/

Multidisciplinary Deep Learning Approaches for Emotion Recognition in Human–Computer Interaction Systems: A Comprehensive Review

Dhrub Kumar, Research Scholar, Department of Liberal Arts and Humanities, Swami Vivekanand Subharti University, Meerut, UP, India Dhrubkramc@gmail.com

Dr. Mohini Mittal, Assistant Professor (Psychology), Department of Liberal Arts and Humanities, Swami Vivekanand Subharti University, Meerut, UP, India agarwalmohini20@gmail.com

Abstract: Emotion recognition has emerged as a pivotal research frontier at the intersection of affective computing, signal processing, and human–computer interaction (HCI), driven by the growing demand for systems capable of perceiving, interpreting, and responding to human affective states in real time. This review synthesises the state of the art in deep learning-based emotion recognition, examining the multidisciplinary convergence of computer vision, speech processing, physiological signal analysis, and natural language understanding that underlies contemporary affect-sensing pipelines. The paper traces the evolution from handcrafted feature engineering to end-to-end representation learning, critically analysing convolutional neural networks, recurrent architectures, attention mechanisms, transformer-based models, and multimodal fusion strategies that have redefined benchmark performance on datasets such as IEMOCAP, CREMA-D, AffectNet, DEAP, and MELD. A structured taxonomy is proposed to classify existing approaches along the axes of modality (unimodal versus multimodal), learning paradigm (supervised, self-supervised, and few-shot), and application context (embodied agents, adaptive tutoring, mental health monitoring, and automotive safety). Architectural blueprints, mathematical formulations of loss functions and fusion operators, and comparative performance tables are presented to ground the discussion in empirical evidence. The review further examines real-world deployment through a case study of affect-aware conversational agents, surveys the software and hardware ecosystem supporting reproducible research, and evaluates recognition accuracy, latency, and robustness across cross-corpus and cross-cultural conditions. Persistent challenges—including data scarcity, annotation subjectivity, domain shift, and the fundamental ambiguity of emotional expression—are discussed alongside ethical considerations surrounding consent, bias, and affective privacy. The review concludes by outlining promising directions, including foundation-model-driven affective reasoning, neuro-symbolic integration, and personalised continual learning, offering a roadmap for researchers seeking to advance emotionally intelligent human–computer interaction systems.

Keywords: affective computing, emotion recognition, deep learning, multimodal fusion, human–computer interaction, facial expression analysis, speech emotion recognition, physiological signals, transformers, attention mechanisms

1. Introduction

The ambition to endow machines with the capacity to perceive and respond to human emotion has moved from speculative fiction to an active and rapidly maturing engineering discipline. As computing systems become deeply embedded in domains that demand sustained, nuanced interaction with people—virtual assistants, educational software, clinical monitoring tools, and autonomous vehicles—the ability to infer a user's affective state has shifted from a peripheral enhancement to a functional necessity. Human communication is inherently multimodal and emotionally coloured; a spoken sentence carries meaning not only through its lexical content but through prosody, facial expression, gesture, and physiological arousal. Systems that ignore this affective layer risk misinterpreting user intent, delivering interactions that feel mechanical, or, in safety-critical contexts, failing to detect states such as fatigue, frustration, or distress that carry direct consequences for user wellbeing. Emotion recognition, as a computational problem, therefore sits at the confluence of psychology, linguistics, signal processing, and machine learning, and its recent acceleration owes much to the broader deep learning revolution that has reshaped pattern recognition across virtually every perceptual modality.

Historically, affect-sensing systems relied on handcrafted features—facial action units derived from the Facial Action Coding System, prosodic descriptors such as pitch contour and energy, and statistical summaries of physiological time series—combined with classical classifiers including support vector machines and hidden Markov models. While these pipelines established the feasibility of automatic emotion recognition, they were constrained by the limited expressiveness of manually engineered representations and by their brittleness under variation in speaker identity, illumination, recording device, or cultural expression norms. The advent of deep convolutional and recurrent architectures, and more recently self-attention-based transformer models, has fundamentally altered this landscape by enabling systems to learn hierarchical, task-relevant representations directly from raw or lightly processed sensory data. This shift has produced substantial gains in recognition accuracy across benchmark corpora, but it has also introduced new methodological challenges: deep models are data-hungry, prone to overfitting on emotion corpora that remain comparatively small relative to those available for tasks such as object recognition or machine translation, and susceptible to encoding spurious correlations that do not generalize across populations.

This review is motivated by the observation that emotion recognition research, despite its shared reliance on deep learning, remains fragmented across disciplinary silos—computer vision researchers optimize facial expression classifiers largely in isolation from speech processing specialists working on prosodic emotion cues, while physiological signal researchers pursue electrodermal activity and electroencephalography-based affect models using methodologies that rarely intersect with either. A multidisciplinary synthesis is therefore both timely and necessary: it allows the field to identify common architectural patterns, shared evaluation weaknesses, and complementary strengths that individual modality-specific surveys tend to overlook. This paper adopts a deliberately cross-modal perspective, treating facial, vocal, physiological, textual, and multimodal fusion approaches as parts of a unified computational problem rather than as separate literatures.

The remainder of this review is organized to first establish the conceptual and historical background that motivates current research directions, before introducing a structured taxonomy that organizes the diverse methodological landscape. Subsequent sections detail core methodologies and representative architectures, present comparative empirical analysis across benchmark datasets, and examine real-world applications through a dedicated case study. The review then surveys the tools and technologies that constitute the practical implementation ecosystem, evaluates performance and robustness considerations, and critically discusses the challenges, ethical implications, and future research directions that will shape the next generation of emotionally aware human–computer interaction systems. Throughout, the review privileges depth of technical explanation over exhaustive enumeration, aiming to equip readers with both a conceptual map of the field and sufficient methodological detail to situate their own research contributions within it.

2. Background and Motivation

Affective computing was formally established as a research discipline through Picard's foundational articulation of the premise that computers capable of recognizing and expressing emotion could interact with humans more naturally and effectively [1]. This early framing anticipated many of the application domains that now animate contemporary research, including affect-aware tutoring systems, emotionally responsive assistive technologies, and physiological stress monitoring. The theoretical grounding for emotion recognition research draws principally from two competing psychological models of emotion. The categorical model, most closely associated with Ekman's cross-cultural studies of facial expression, posits a small set of discrete basic emotions—typically anger, disgust, fear, happiness, sadness, and surprise—that are held to be universally recognizable across cultures [2]. The dimensional model, by contrast, represents emotional states as continuous points within a low-dimensional space, most commonly the valence–arousal plane in which valence captures the pleasantness of an affective state and arousal captures its intensity, with a third dominance dimension sometimes appended to form the PAD (pleasure–arousal–dominance) space [3]. The choice between categorical and dimensional annotation schemes has direct downstream consequences for system design: categorical labels naturally support classification formulations with cross-entropy objectives, whereas dimensional labels are more naturally treated as regression targets and permit finer-grained characterization of affective states that resist discrete labeling, such as mixed or ambivalent emotions.

The motivation for pursuing deep learning approaches specifically, rather than continuing to refine classical pipelines, stems from several converging factors. First, the availability of large-scale annotated corpora—though still modest compared to other machine learning domains—has grown substantially over the past decade, with datasets such as AffectNet providing over a million facial images with both categorical and dimensional annotations, and IEMOCAP and MSP-IMPROV offering richly annotated multimodal conversational data [4], [5]. Second, advances in general-purpose deep learning architectures, including deep convolutional networks pretrained on large image corpora and, more recently, transformer models pretrained through self-supervised objectives on vast unlabeled speech and text corpora, have provided transferable representations that substantially reduce the annotated data required to achieve strong emotion recognition performance through fine-tuning or transfer learning. Third, the proliferation of low-cost sensing hardware—high-resolution cameras, microphone arrays, wearable physiological sensors, and consumer-grade electroencephalography headsets—has expanded the practical feasibility of multimodal data collection outside laboratory settings, motivating research into models robust to the noise and variability characteristic of in-the-wild deployment.

The practical stakes of this research extend across numerous application domains that collectively justify sustained investment. In education, affect-aware intelligent tutoring systems that detect confusion, boredom, or frustration can adapt instructional pacing and content delivery to sustain learner engagement, addressing a long-recognized limitation of static computer-based instruction [6]. In healthcare, automatic detection of depressive or anxious

affect from speech and facial cues offers a scalable complement to clinical assessment, particularly relevant given global shortages of mental health practitioners. In automotive systems, driver monitoring pipelines that fuse facial expression, gaze, and physiological signals to detect fatigue or emotional distraction directly address road safety, a domain where the cost of misclassification is measured in human lives rather than user experience alone. In customer service and conversational artificial intelligence, sentiment- and emotion-aware dialogue systems enable more contextually appropriate response generation, improving user satisfaction and enabling early detection of escalating frustration in support interactions.

Despite this clear motivation, the field confronts a persistent tension between laboratory performance and real-world reliability. Benchmark datasets are frequently collected under controlled conditions—posed or elicited expressions, scripted utterances, constrained lighting—that do not reflect the spontaneity, subtlety, and contextual entanglement of naturalistic emotional expression. This gap, often termed the "in-the-wild" problem, motivates much of the recent methodological innovation surveyed in this review, including domain adaptation techniques, self-supervised pretraining on unlabelled naturalistic data, and multimodal fusion strategies designed to compensate for the unreliability of any single modality under real-world noise conditions. Understanding this background is essential for situating the taxonomy and methodological review that follow, since the design choices underlying contemporary architectures are frequently direct responses to the specific limitations articulated here.

3. Taxonomy / Classification

A coherent taxonomy is indispensable for navigating a literature that spans modalities, learning paradigms, and application contexts with substantially different methodological conventions. This review organizes existing approaches along four principal axes: input modality, emotion representation model, learning paradigm, and fusion strategy. Table 1 summarizes this taxonomy, and the subsequent discussion elaborates the rationale and representative methods associated with each category.

The modality axis distinguishes unimodal systems, which infer affect from a single channel such as facial video, speech audio, physiological signals, or text, from multimodal systems that integrate two or more channels. Unimodal facial expression recognition remains the most extensively studied subfield, benefiting from mature convolutional and, increasingly, vision-transformer architectures. Speech emotion recognition constitutes the second major unimodal branch, exploiting prosodic, spectral, and increasingly self-supervised acoustic representations. Physiological emotion recognition, drawing on electroencephalography, electrodermal activity, heart rate variability, and electromyography, offers the advantage of resistance to voluntary suppression—since autonomic responses are considerably harder to consciously mask than facial or vocal expression—but suffers from higher instrumentation burden and lower ecological validity outside controlled settings. Text-based emotion recognition, operating on transcribed speech or written communication, has been transformed by pretrained language models and is increasingly integrated into multimodal conversational emotion recognition pipelines. Multimodal systems, the fastest-growing category in recent literature, seek to exploit the complementary and partially redundant information across channels to achieve robustness that no single modality can provide in isolation.

The emotion representation axis distinguishes categorical models, which frame the task as multi-class or multi-label classification over discrete emotion categories, from dimensional models, which frame the task as regression over continuous valence, arousal, and sometimes dominance axes. A smaller but growing body of work adopts appraisal-theoretic or componential models that attempt to capture the cognitive antecedents of emotional response, though these remain less common in deep learning implementations due to the difficulty of obtaining corresponding annotations at scale.

The learning paradigm axis reflects the maturation of the field's methodological toolkit. Fully supervised learning, requiring densely labeled training data, remains dominant but is increasingly supplemented by self-supervised pretraining, in which large volumes of unlabeled audio, video, or text are used to learn general-purpose representations later fine-tuned on smaller labeled emotion corpora. Few-shot and meta-learning approaches address the practical reality that many application-specific emotion recognition tasks—for instance, recognizing a particular individual's idiosyncratic expressive style—must operate with minimal labeled examples. Domain adaptation and transfer learning techniques explicitly target the cross-corpus generalization gap, while continual learning approaches, still relatively nascent in this domain, aim to allow models to incorporate new emotional expression patterns over time without catastrophic forgetting of previously learned representations.

The fusion strategy axis, applicable specifically to multimodal systems, distinguishes early fusion, in which raw or low-level features from different modalities are concatenated prior to joint processing; late fusion, in which independent unimodal models produce modality-specific predictions that are subsequently combined through weighted averaging, voting, or a shallow meta-classifier; and hybrid or intermediate fusion, in which modality-specific representations are learned partially independently before being merged at one or more intermediate network layers, often through attention-based or tensor-based fusion operators. Table 2 provides a comparative summary of these fusion strategies alongside their principal advantages and limitations, which is elaborated further in Section 5.

Table 1. Taxonomy of Deep Learning–Based Emotion Recognition Approaches

Taxonomy Axis	Categories	Representative Methods/Models
Input Modality	Facial (vision)	CNNs, Vision Transformers, 3D-CNNs for video
	Speech/Audio	CNN-LSTM, wav2vec 2.0, HuBERT
	Physiological	EEG-based CNN/GNN, EDA/HRV feature networks
	Text	BERT, RoBERTa, emotion-tuned LLMs
	Multimodal	Cross-modal transformers, tensor fusion networks
Emotion Model	Categorical	Ekman's six/seven basic emotions
	Dimensional	Valence-Arousal(-Dominance) regression
	Componential/Appraisal	Hybrid symbolic-neural appraisal models
Learning Paradigm	Supervised	End-to-end CNN/RNN classifiers
	Self-supervised pretraining	Contrastive and masked-prediction pretraining
	Few-shot/Meta-learning	Prototypical networks, MAML-based adaptation
	Domain adaptation	Adversarial domain-invariant learning
Fusion Strategy	Early fusion	Feature-level concatenation
	Late fusion	Decision-level combination
	Hybrid/Intermediate fusion	Cross-attention, tensor fusion

This taxonomy is not intended as a strict partition, since contemporary state-of-the-art systems frequently combine elements from multiple categories—for example, a multimodal system may employ self-supervised speech representations, supervised facial expression features, and hybrid attention-based fusion within a single architecture. Rather, the taxonomy provides an analytical vocabulary that structures the methodological review presented in the following section and clarifies the design space within which specific architectural choices should be evaluated.

4. Core Concepts and Methodologies

The methodological core of deep learning–based emotion recognition rests on the general principle of learning a mapping from raw or lightly pre-processed sensory input to an emotion label or continuous affective score, optimized end-to-end through gradient-based learning. This section elaborates the principal methodological families that instantiate this general principle across modalities, beginning with facial expression recognition, proceeding through speech and physiological emotion recognition, and concluding with the multimodal fusion methodologies that increasingly define the state of the art.

Facial expression recognition pipelines typically begin with face detection and alignment, commonly performed using multitask cascaded convolutional networks or similar landmark-based detectors, followed by feature extraction using deep convolutional architectures. Early deep learning approaches adapted architectures originally developed for general image classification, including VGG-style and residual networks, fine-tuned on facial expression corpora. More recent approaches exploit vision transformers, which partition an image into patches processed through self-attention layers, offering the capacity to model long-range spatial dependencies between facial regions—for instance, the coordinated movement of eyebrows and mouth corners characteristic of specific expressions—that convolutional receptive fields capture less directly. For video-based facial expression recognition, temporal modeling is essential to capture the dynamic unfolding of an expression from onset through

apex to offset; this is commonly achieved through 3D convolutional networks that jointly convolve over spatial and temporal dimensions, or through hybrid architectures that apply frame-level convolutional feature extraction followed by recurrent or transformer-based temporal aggregation. The loss function for categorical facial expression recognition is typically the standard cross-entropy objective, defined for a predicted class distribution

$\hat{y}$ and ground-truth one-hot label y over C classes as

$$\mathcal{L}_{CE} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$$

while dimensional prediction tasks commonly employ mean squared error or concordance correlation coefficient loss, the latter favored because it explicitly penalizes systematic bias and scale mismatch between predicted and ground-truth valence-arousal trajectories, a property standard regression losses do not capture.
Speech emotion recognition methodologies have undergone a particularly pronounced transformation with the introduction of self-supervised speech representation models. Traditional pipelines extracted handcrafted acoustic feature sets, most notably the extended Geneva Minimalistic Acoustic Parameter Set, comprising pitch, jitter, shimmer, formant frequencies, and spectral energy descriptors, which were subsequently fed to recurrent networks or support vector machines [7]. Contemporary approaches instead leverage self-supervised models such as wav2vec 2.0 and HuBERT, pretrained on thousands of hours of unlabeled speech through contrastive or masked-prediction objectives, whose intermediate layer representations have been shown to encode paralinguistic information relevant to emotion recognition even though the pretraining objective targets phonetic content rather than affect [8]. Fine-tuning these pretrained encoders on comparatively small labeled emotion corpora, often with a lightweight classification head, has produced substantial performance gains over models trained from randomly initialized weights, illustrating the broader trend toward transfer learning as a mitigation strategy for the data scarcity that has historically constrained emotion recognition research.

Physiological emotion recognition methodologies must contend with signal properties markedly different from vision or audio: electroencephalography data is high-dimensional, spatially structured across scalp electrodes, and characterized by substantial inter-subject variability arising from differences in skull morphology and baseline neural activity. Graph neural networks have emerged as a particularly well-suited architectural choice for electroencephalography-based emotion recognition, since they naturally accommodate the non-Euclidean spatial structure of electrode montages by representing electrodes as graph nodes and modeling functional connectivity as edge weights, allowing the network to learn which inter-electrode relationships are diagnostic of affective state [9]. Electrodermal activity and heart rate variability signals, by contrast, are typically processed through one-dimensional convolutional or recurrent networks operating on the raw or lightly filtered time series, with domain-informed preprocessing steps such as skin conductance response decomposition retained even within otherwise end-to-end pipelines, reflecting the persistent value of physiological domain knowledge even in the deep learning era.

Text-based emotion recognition has been reshaped by the widespread adoption of pretrained transformer language models, including BERT and its derivatives, which are fine-tuned on emotion-labeled text corpora such as GoEmotions or task-specific conversational datasets [10]. A distinguishing characteristic of text-based emotion recognition within conversational contexts is the importance of discourse-level context: the emotional interpretation of an utterance frequently depends on preceding conversational turns, motivating architectures that explicitly model conversational history through hierarchical recurrent or graph-based context encoders layered atop utterance-level transformer representations.

Multimodal fusion methodologies, discussed in greater architectural detail in Section 5, build upon these unimodal foundations by learning to combine complementary information across channels. A central methodological challenge in multimodal fusion is the differing temporal resolution and semantic granularity of different modalities—facial video is sampled at frame rate, speech audio at a much higher sampling rate, and text at the discrete level of tokens or words—necessitating alignment mechanisms, commonly implemented through cross-modal attention, that learn correspondences between temporally or semantically heterogeneous modality streams without requiring explicit forced alignment. Collectively, these methodological developments illustrate a coherent trajectory across modalities: from handcrafted features and shallow classifiers, toward deep supervised architectures, and most recently toward self-supervised pretraining and cross-modal attention mechanisms that increasingly define the methodological frontier of the field, setting the stage for the specific architectural implementations examined in the following section.

5. System Architecture / Algorithms

Translating the methodological principles outlined above into deployable systems requires attention to concrete architectural design. This section presents representative architectural blueprints for unimodal and multimodal emotion recognition systems, together with the mathematical formulation of key components, particularly attention-based fusion mechanisms that have become central to contemporary designs.

A representative unimodal facial expression recognition architecture proceeds through four stages: face detection and alignment, backbone feature extraction, temporal aggregation for video input, and classification. The backbone is typically a convolutional network such as a ResNet variant or a vision transformer pretrained on a large face recognition corpus and subsequently fine-tuned on expression-labeled data, exploiting the observation that identity-discriminative and expression-discriminative features share substantial low- and mid-level structure. Temporal aggregation over a sequence of frame-level embeddings ${e_1, e_2, \ldots, e_T}$ is commonly performed using a bidirectional long short-term memory network or a temporal transformer encoder, producing a sequence-level representation subsequently passed to a fully connected classification head with softmax output for categorical prediction or a linear output layer for dimensional regression.

For speech emotion recognition, a representative architecture consists of a self-supervised acoustic encoder, most commonly a wav2vec 2.0 or HuBERT model, whose output is a sequence of frame-level contextualized representations. These are aggregated through attentive statistics pooling, which computes a weighted average of frame-level embeddings using learned attention weights α_t, defined as

$$\alpha_t = \frac{\exp(w^T \tanh(W e_t + b))}{\sum_{t'=1}^{T} \exp(w^T \tanh(W e_{t'} + b))},

\qquad z = \sum_{t=1}^{T} \alpha_t e_t$$

where e_t denotes the frame-level embedding at time t, and W, w, and b are learned parameters. This pooled representation z is passed to a classification or regression head analogous to the facial expression pipeline. Attentive pooling is preferred over simple average pooling because emotional information is frequently concentrated in specific temporal segments—for instance, a vocal outburst or pitch spike—that uniform averaging would dilute.

Multimodal fusion architectures extend these unimodal backbones through one of the three fusion strategies introduced in Section 3. Early fusion architectures concatenate modality-specific feature vectors prior to a shared classification head, formulated simply as $z = [z_v; z_a; z_t]$ for visual, acoustic, and textual embeddings respectively, followed by a multilayer perceptron. While computationally efficient, early fusion assumes that a single shared representation space can adequately capture cross-modal interactions, an assumption that becomes increasingly strained as the number and heterogeneity of modalities grows. Late fusion architectures instead train independent unimodal classifiers and combine their output probability distributions, typically through weighted averaging $\hat{y} = \sum_{m} \beta_m \hat{y}_m$, where β_m denotes a learned or fixed modality weight, offering robustness to missing modalities at inference time but forgoing the opportunity to model fine-grained cross-modal dependencies during representation learning.

Hybrid fusion architectures, which represent the current state of the art for multimodal emotion recognition, employ cross-modal attention mechanisms that allow one modality's representation to be refined through attention over another modality's representation sequence. The cross-modal attention operation, adapted from the transformer attention mechanism, computes queries from one modality and keys and values from another, formulated as

$$\text{CrossAttn}(Q_v, K_a, V_a) = \text{softmax}\left(\frac{Q_v K_a^T}{\sqrt{d_k}}\right) V_a$$

where Q_v is derived from the visual modality and K_a, V_a from the acoustic modality, with d_k denoting the key dimensionality used for scaling. This formulation, central to the Multimodal Transformer architecture proposed by Tsai et al., allows the model to learn, for each visual time step, which acoustic time steps are most relevant to refining that visual representation, and vice versa when the roles of the modalities are reversed [11]. Stacking multiple such cross-modal attention blocks across all pairwise modality combinations, followed by self-attention within each refined modality stream and a final fusion layer, constitutes the dominant architectural pattern in recent state-of-the-art multimodal emotion recognition systems, including extensions incorporating

tensor fusion networks that explicitly model higher-order multiplicative interactions between modalities through outer-product operations, at the cost of increased parameter count scaling exponentially with the number of modalities [12].

Table 2. Comparison of Multimodal Fusion Strategies

Fusion Strategy	Architectural Principle	Advantages	Limitations
Early Fusion	Feature concatenation prior to joint modeling	Simple, computationally efficient, captures some joint structure	Sensitive to modality misalignment; degrades with missing modalities
Late Fusion	Independent unimodal models, decision-level combination	Robust to missing modalities; modular and interpretable	Cannot model fine-grained cross-modal interaction
Hybrid/Cross-Attention Fusion	Attention-based intermediate representation refinement	Captures asynchronous cross-modal dependencies; state-of-the-art accuracy	Higher computational cost; harder to interpret; more prone to overfitting on small corpora
Tensor Fusion	Outer-product modeling of joint modality interactions	Explicitly models higher-order interactions	Parameter count grows exponentially with modality count

Beyond the fusion mechanism itself, contemporary architectures increasingly incorporate graph-based context modeling for conversational emotion recognition, in which each utterance in a dialogue is represented as a graph node connected to preceding and following utterances, with edge weights learned to reflect contextual relevance, enabling graph convolutional or graph attention networks to propagate contextual information across the conversation while respecting its sequential and speaker-dependent structure, as instantiated in architectures such as DialogueGCN [13]. This architectural overview illustrates that, despite surface-level diversity across modalities and applications, contemporary emotion recognition systems converge on a shared set of architectural primitives—convolutional or transformer backbones for representation extraction, attention-based mechanisms for temporal and cross-modal aggregation—that recur with modality-specific adaptation throughout the literature, providing a foundation for the comparative empirical analysis presented in the following section.

6. Comparative Analysis

Empirical comparison across the architectural families introduced above is essential for grounding the preceding conceptual discussion in measurable evidence and for clarifying the practical trade-offs facing system designers. This section synthesizes reported performance across representative benchmark datasets and architectural approaches, while explicitly noting the methodological caveats that complicate direct cross-study comparison.

Facial expression recognition benchmarks report weighted or unweighted accuracy on datasets including FER2013, RAF-DB, and AffectNet, with recent vision-transformer-based approaches achieving unweighted accuracy in the range of 65–67 percent on the challenging eight-class AffectNet benchmark, an improvement of several percentage points over earlier convolutional baselines, though performance remains substantially constrained by label noise and inherent class imbalance favoring neutral and happy expressions over categories such as disgust and fear [4]. Speech emotion recognition performance is most commonly reported on IEMOCAP using four- or five-class categorical schemes, where self-supervised pretrained encoders such as wav2vec 2.0, fine-tuned end-to-end, achieve weighted accuracy exceeding 70 percent, representing a marked improvement over the 55–60 percent range typical of earlier handcrafted-feature pipelines combined with recurrent classifiers [8]. Multimodal systems evaluated on IEMOCAP or CMU-MOSEI consistently outperform their best constituent unimodal counterparts, with cross-attention-based multimodal transformers reporting accuracy gains of 3–6 percentage points over the strongest single-modality baseline, empirically substantiating the central premise that complementary modalities carry non-redundant affective information [11].

Table 3 consolidates representative reported performance figures drawn from widely cited studies, presented alongside dataset and evaluation protocol details to contextualise the comparison. Readers should interpret these figures as indicative rather than strictly comparable, since studies differ in class definitions, train-test partitioning

protocols (particularly speaker-dependent versus speaker-independent splits), and preprocessing pipelines, all of which materially affect reported accuracy.

Table 3. Representative Comparative Performance Across Modalities and Architectures

Modality	Dataset	Representative Architecture	Reported Metric	Approx. Performance
Facial (image)	AffectNet (8-class)	Vision Transformer fine-tuned	Unweighted accuracy	~65–67%
Facial (video)	RAF-DB	ResNet + temporal attention	Overall accuracy	~88–90%
Speech	IEMOCAP (4-class)	wav2vec 2.0 fine-tuned	Weighted accuracy	~70–72%
Speech	IEMOCAP (4-class)	CNN-BiLSTM (handcrafted features)	Weighted accuracy	~55–60%
EEG	DEAP (valence, binary)	Graph neural network	Accuracy	~90–93% (subject-dependent)
Text	GoEmotions	Fine-tuned BERT	Macro F1	~46–50% (27-class, multi-label)
Multimodal	CMU-MOSEI	Multimodal Transformer (cross-attention)	Accuracy (7-class sentiment)	~54–56%
Multimodal	IEMOCAP	Cross-attention fusion	Weighted accuracy	~75–78%

A recurring pattern in this comparative landscape is the substantial gap between subject-dependent and subject-independent (also termed cross-subject or leave-one-speaker-out) evaluation protocols, particularly pronounced for physiological modalities: electroencephalography-based valence classification accuracy on DEAP frequently exceeds 90 percent under subject-dependent evaluation, where training and test data derive from the same individuals, but drops markedly, often into the 55–65 percent range, under strict cross-subject evaluation, reflecting the substantial inter-individual variability in physiological affective response and raising important questions about the practical deployability of physiological emotion recognition systems intended for novel users without individual calibration [9]. A comparable, though generally less severe, degradation is observed for facial and speech modalities when models trained on one corpus are evaluated on a distinct corpus recorded under different elicitation protocols, camera or microphone characteristics, and population demographics, a phenomenon commonly termed the cross-corpus generalization gap and a subject of dedicated research using domain adaptation and adversarial feature alignment techniques [14].

Comparative analysis of learning paradigms further reveals that self-supervised pretraining confers particularly large benefits under conditions of limited labeled data: fine-tuning pretrained wav2vec 2.0 or vision transformer backbones on small emotion corpora consistently outperforms training equivalent architectures from randomly initialized weights, with the performance gap widening as the size of the labeled fine-tuning corpus decreases, underscoring the practical value of transfer learning strategies for a field structurally characterized by comparatively small labeled datasets relative to other computer vision or natural language processing tasks. Taken together, this comparative evidence supports three principal conclusions that recur throughout the remainder of this review: multimodal fusion reliably outperforms unimodal approaches when the constituent modalities are of reasonable individual quality; self-supervised pretraining substantially mitigates data scarcity; and cross-subject and cross-corpus generalization remain the field's most significant unresolved empirical challenge, a theme revisited in greater depth in Sections 10 and 11.

7. Applications

The methodological and architectural developments surveyed above find concrete expression across a diverse set of application domains, each imposing distinct constraints on system design and each illustrating different facets of the multidisciplinary character of emotion recognition research. This section surveys five principal application domains, examining how domain-specific requirements shape the adaptation of the general methodological toolkit described in Sections 4 and 5.

In education, affect-aware intelligent tutoring systems integrate facial expression and, increasingly, physiological or interaction-log-based affect signals to detect learner states such as confusion, boredom, engagement, and frustration, with the goal of dynamically adapting instructional content, pacing, or hint provision to sustain productive learning. Empirical studies have demonstrated that tutoring systems incorporating affect-sensitive intervention strategies produce measurable improvements in learning gains and engagement persistence relative to affect-agnostic baselines, particularly for learners exhibiting sustained confusion or boredom states that predict disengagement [6]. The educational application domain is methodologically distinctive in its reliance on classroom-deployed, often low-resolution and variably illuminated camera feeds, motivating research into lightweight facial expression models suitable for edge deployment on classroom hardware without reliance on continuous cloud connectivity, a constraint that has motivated recent work on model compression and knowledge distillation specifically targeting affect recognition pipelines.

In mental health and clinical applications, automatic emotion and affect recognition from speech, facial expression, and, in select research contexts, physiological signals has been explored as a scalable screening and monitoring tool for conditions including depression, anxiety, and post-traumatic stress disorder. The DAIC-WOZ corpus, comprising clinical interview recordings annotated with depression severity scores, has become a standard benchmark for multimodal depression detection research, with recent multimodal deep learning approaches combining facial action unit dynamics, prosodic speech features, and linguistic sentiment achieving substantially improved detection performance relative to unimodal baselines [15]. This application domain carries particularly acute ethical stakes, discussed further in Section 12, given the sensitivity of mental health data and the risk of both false positives, which may cause unwarranted distress or stigmatization, and false negatives, which may delay needed clinical intervention.

In automotive human-machine interaction, driver monitoring systems fuse facial expression, eye-gaze, head-pose, and increasingly physiological signals obtained from steering-wheel-embedded or seat-embedded sensors to detect fatigue, drowsiness, and emotional distraction states associated with elevated collision risk. These systems operate under stringent real-time latency constraints and must maintain robust performance across highly variable illumination conditions, occlusion from sunglasses or masks, and diverse driver demographics, motivating specialized architectural adaptations including infrared imaging to maintain performance under low-light conditions and lightweight convolutional architectures optimized for automotive-grade embedded processors with strict power and thermal budgets.

In conversational artificial intelligence and customer service, emotion- and sentiment-aware dialogue systems leverage the text- and speech-based methodologies described in Section 4 to detect user frustration, satisfaction, or escalating distress during interactions with virtual assistants or automated customer support systems, enabling dynamic response adaptation, appropriate escalation to human agents when frustration signals exceed a threshold, and post-hoc quality assurance analysis of customer interactions at scale. This domain has been a primary driver of research into conversational emotion recognition architectures that explicitly model dialogue-level context, as introduced in Section 5, since the emotional trajectory across a multi-turn interaction, rather than any single utterance in isolation, typically carries the most actionable signal for intervention.

In embodied and social robotics, emotion recognition constitutes a core perceptual component enabling socially assistive robots deployed in eldercare, autism intervention, and companionship applications to modulate their behavior in response to detected user affect, an application domain that additionally requires the integration of emotion recognition with emotion expression and generation capabilities to support genuinely bidirectional affective interaction, an area of active research extending beyond the perceptual focus of this review but directly dependent upon its outputs. Across all five domains, a recurring design tension emerges between the accuracy achievable under controlled, high-quality sensing conditions and the robustness required for sustained real-world deployment, a tension that motivates much of the ongoing methodological research into domain adaptation, multimodal robustness, and lightweight model design discussed in subsequent sections.

8. Case Study: Affect-Aware Conversational Agents in Mental Wellbeing Support

To ground the preceding survey in a concrete implementation context, this section presents a case study examining the design and evaluation of affect-aware conversational agents deployed for mental wellbeing support, a domain that has attracted substantial recent research attention and illustrates the practical integration of multiple methodological threads discussed throughout this review. Such systems typically combine automatic speech

recognition, text-based sentiment and emotion classification, and, in more sophisticated implementations, facial expression analysis obtained through the user's device camera during video-based interaction, fusing these signals to inform both the immediate conversational response generated by the system and longer-term mood trend monitoring intended to surface concerning patterns to the user or, with consent, to a supervising clinician.

The architectural pattern representative of this application class begins with a real-time speech-to-text pipeline whose transcription output is passed to a fine-tuned transformer-based emotion classifier operating over a taxonomy adapted from clinical mood assessment instruments rather than the basic emotion categories common in laboratory research, reflecting the domain-specific adaptation of general-purpose methodology to application context. In parallel, prosodic features extracted directly from the raw audio signal, bypassing the lossy intermediate step of text transcription, are processed through a self-supervised acoustic encoder analogous to the architecture described in Section 5, capturing paralinguistic affective cues such as vocal flatness or tremor that are known clinical markers of depressive and anxious states but that are not recoverable from text transcription alone. These two modality-specific predictions are combined through a late or hybrid fusion layer, with system designers frequently favoring late fusion in this domain specifically because it preserves interpretability and allows individual modality contributions to be audited, a property of particular importance given the clinical adjacency of the application and the corresponding requirement for explainability discussed further in Section 12. Evaluation of such systems proceeds along two complementary axes: recognition accuracy against clinician-annotated ground truth, typically reported using concordance correlation coefficients against continuous mood severity scales rather than simple classification accuracy, reflecting the dimensional rather than categorical framing appropriate to clinical mood assessment; and downstream engagement and outcome metrics, including sustained user engagement over multiweek deployment periods and, where ethically and logistically feasible, standardized clinical outcome measures such as the PHQ-9 depression severity instrument administered before and after a period of system use. Reported evaluations of representative systems in this space indicate that multimodal fusion of speech-derived acoustic and linguistic features improves concordance with clinician mood ratings relative to either modality in isolation, consistent with the broader multimodal fusion findings summarized in Section 6, while also revealing that system performance degrades measurably for users from linguistic or cultural backgrounds underrepresented in the training data, a finding that directly motivates the domain adaptation and dataset diversity discussion in Sections 11 and 12.

This case study illustrates several cross-cutting themes developed throughout this review in concrete form. First, it demonstrates the practical necessity of multimodal integration even in applications where a single modality, text, might appear at first sight sufficient, since acoustic and linguistic channels are shown to carry complementary clinically relevant information. Second, it illustrates the tension between the pursuit of maximal recognition accuracy and the practical requirement for interpretability and auditability in high-stakes application domains, a tension that has motivated increased research attention to explainable emotion recognition architectures that expose attention weights or modality contribution scores as interpretable byproducts of the fusion process rather than treating the fusion mechanism purely as an accuracy-optimizing black box. Third, the observed performance degradation across demographic subgroups within this case study directly anticipates the ethical and generalization challenges discussed in greater depth later in this review, illustrating that these are not abstract theoretical concerns but empirically documented limitations of systems already approaching real-world deployment.

9. Tools, Technologies, and Implementation

The practical realization of the architectures and methodologies surveyed above depends on a substantial ecosystem of software frameworks, pretrained models, and hardware platforms that has matured considerably over the past several years, substantially lowering the barrier to entry for emotion recognition research and enabling more consistent reproducibility across studies. This section surveys the principal components of this implementation ecosystem, organized by modality and functional role.

For deep learning implementation broadly, PyTorch has emerged as the dominant framework within the emotion recognition research community, owing to its dynamic computation graph facilitating the variable-length sequence processing common to speech and physiological signal modeling, and to the extensive availability of pretrained model checkpoints distributed through the Hugging Face Transformers and Hugging Face Hub ecosystems, which host pretrained wav2vec 2.0, HuBERT, BERT, and vision transformer checkpoints readily amenable to fine-tuning on emotion-labeled corpora with minimal boilerplate implementation. TensorFlow and its high-level Keras interface remain in use, particularly within legacy codebases and industrial deployment

pipelines that benefit from TensorFlow's more mature mobile and embedded deployment tooling through TensorFlow Lite, a consideration of particular relevance to the automotive and edge-deployment application contexts discussed in Section 7.

For facial expression recognition specifically, the OpenCV library and the dlib toolkit provide mature face detection and landmark localization functionality that continues to serve as a preprocessing front end even within otherwise deep-learning-centric pipelines, while the py-feat and OpenFace toolkits provide facial action unit extraction and expression classification pipelines widely adopted in psychology and affective computing research for their interpretability and grounding in the Facial Action Coding System. For speech emotion recognition, the openSMILE toolkit remains the standard implementation for extracting the extended Geneva Minimalistic Acoustic Parameter Set and related handcrafted acoustic feature sets, frequently used as a baseline or complementary feature source alongside self-supervised acoustic embeddings, while the SpeechBrain and NeMo toolkits provide end-to-end implementations of self-supervised speech representation learning and downstream fine-tuning pipelines directly applicable to speech emotion recognition tasks.

For physiological signal processing, the MNE-Python library provides comprehensive electroencephalography preprocessing functionality, including artifact removal through independent component analysis and standard filtering operations, while specialized libraries such as NeuroKit2 support electrodermal activity, heart rate variability, and electromyography feature extraction with implementations of the domain-informed preprocessing steps referenced in Section 4. Graph neural network implementations for electroencephalography-based emotion recognition are commonly built atop the PyTorch Geometric library, which provides efficient implementations of graph convolutional and graph attention layers directly applicable to the electrode-graph formulations discussed in Section 4.

Dataset accessibility constitutes a further critical component of the implementation ecosystem. Table 4 summarizes representative benchmark datasets referenced throughout this review alongside their modality coverage, annotation scheme, and approximate scale, providing implementation-relevant context for researchers selecting appropriate training and evaluation resources.

Table 4. Representative Benchmark Datasets for Emotion Recognition Research

Dataset	Primary Modality	Annotation Scheme	Approx. Scale	Notable Characteristic
AffectNet	Facial image	Categorical (8 classes) + valence-arousal	~1M images	Largest in-the-wild facial affect dataset
RAF-DB	Facial image	Categorical (basic + compound)	~30,000 images	Crowd-sourced real-world annotations
FER2013	Facial image	Categorical (7 classes)	~35,000 images	Low-resolution, widely used legacy benchmark
IEMOCAP	Audio-visual-text (dyadic)	Categorical + dimensional	~12 hours, 10 actors	Scripted and improvised dyadic interactions
CREMA-D	Audio-visual	Categorical (6 classes)	~7,400 clips	Crowd-sourced multi-rater intensity labels
MELD	Audio-visual-text (multiparty)	Categorical (7 classes)	~13,000 utterances	Multiparty conversational context (TV series)
CMU-MOSEI	Audio-visual-text	Sentiment + emotion	~23,500 utterances	Large-scale in-the-wild YouTube monologues
DEAP	EEG + peripheral physiology	Dimensional (valence, arousal, dominance)	32 participants, 40 trials each	Music-video elicitation protocol
DAIC-WOZ	Audio-visual-text	Clinical depression severity	~189 clinical interviews	Wizard-of-Oz virtual interviewer
GoEmotions	Text	Categorical (27 emotions + neutral, multi-label)	~58,000 Reddit comments	Fine-grained emotion taxonomy

The maturation of this tooling ecosystem has substantially reduced implementation barriers relative to a decade ago, when researchers routinely reimplemented handcrafted feature extraction pipelines from published descriptions with limited standardization. Nonetheless, meaningful implementation challenges persist: pretrained model checkpoints are predominantly trained on English-language and Western-demographic-skewed data, motivating ongoing efforts, still incomplete, to develop and release comparably resourced pretrained models and benchmark corpora for underrepresented languages and populations, a limitation directly relevant to the generalization and fairness considerations discussed in Sections 11 and 12. Furthermore, while individual modality-specific toolkits are individually mature, standardized end-to-end multimodal fusion frameworks remain comparatively immature relative to unimodal tooling, with most published multimodal architectures implemented as custom research code rather than through shared, actively maintained libraries, presenting an identifiable opportunity for tooling consolidation that would meaningfully improve reproducibility across the multimodal emotion recognition literature.

10. Performance Evaluation

Rigorous performance evaluation in emotion recognition research requires careful attention to metric selection, evaluation protocol, and the statistical characteristics of the underlying label distributions, each of which materially affects the interpretability and comparability of reported results. This section examines these evaluation considerations in detail, extending the comparative performance data presented in Section 6 with a critical discussion of methodological best practice and its frequent violation in published research.

Metric selection must be matched to the underlying task formulation. For categorical emotion classification, weighted accuracy, which accounts for class imbalance by weighting per-class accuracy by class frequency, and unweighted accuracy, equivalent to the macro-averaged recall across classes, are both commonly reported, with unweighted accuracy generally regarded as the more informative metric under the pronounced class imbalance characteristic of naturalistic emotion corpora, where neutral and happy expressions are substantially overrepresented relative to categories such as disgust or fear. Macro-averaged F1 score, which balances precision and recall on a per-class basis before averaging, is similarly favored over raw accuracy for imbalanced multi-label emotion classification tasks such as those posed by the GoEmotions dataset.

For dimensional emotion regression, the concordance correlation coefficient, defined for predicted and ground-truth sequences

$\hat{y}$ and y with means $\mu_{\hat{y}}, \mu_y$, variances $\sigma_{\hat{y}}^2, \sigma_y^2$, and covariance $\sigma_{\hat{y}y}$ as

$$\rho_c = \frac{2\sigma_{\hat{y}y}}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2}$$ is preferred over Pearson correlation or mean squared error alone because it simultaneously penalizes both poor correlation and systematic bias or scale mismatch, properties that a system optimizing purely for correlation could otherwise exploit by producing predictions correctly ordered but poorly calibrated in absolute magnitude.

Evaluation protocol design constitutes an equally consequential methodological consideration, particularly the choice between subject-dependent, subject-independent (leave-one-subject-out or leave-one-speaker-out), and cross-corpus evaluation, each answering a materially different research question. Subject-dependent evaluation, in which the same individuals appear in both training and test partitions, is appropriate for applications involving sustained interaction with a known, consistent user base, such as personalized wellbeing monitoring applications where a period of individual calibration is acceptable, but substantially overestimates performance for applications requiring immediate generalization to novel users, as the comparative degradation reported in Section 6 for electroencephalography-based recognition starkly illustrates. Subject-independent evaluation, though more methodologically demanding to implement correctly and more representative of realistic deployment conditions, remains inconsistently adopted across published research, with a non-trivial fraction of studies, particularly in the physiological signal recognition literature, continuing to report subject-dependent results without adequate discussion of the corresponding generalization limitations, a methodological weakness that this review explicitly flags as warranting greater community attention and more consistent reviewer scrutiny during the publication process.

Cross-corpus evaluation, in which a model trained on one dataset is evaluated on an entirely distinct dataset recorded under different elicitation protocols and population characteristics, provides the most stringent test of

generalization and is correspondingly the least commonly reported evaluation condition, owing to the practical challenge of reconciling differing emotion taxonomies and annotation protocols across corpora. Studies that do report cross-corpus evaluation consistently find substantial performance degradation relative to within-corpus evaluation, with speech emotion recognition accuracy frequently dropping by 15 to 25 percentage points when models trained on acted or elicited emotion corpora such as IEMOCAP are evaluated on more naturalistic in-the-wild corpora, underscoring that within-corpus benchmark performance, the dominant reporting convention in the field, substantially overstates the practical readiness of current systems for unconstrained real-world deployment [14].

Beyond accuracy-oriented metrics, computational efficiency evaluation, encompassing inference latency, model parameter count, and energy consumption, has received increasing attention as emotion recognition systems move toward edge and embedded deployment contexts such as the automotive driver monitoring application discussed in Section 7. Comparative studies of model compression techniques, including knowledge distillation and structured pruning applied to facial expression recognition backbones, report that compressed models retaining 90 to 95 percent of full-model accuracy while reducing parameter count by an order of magnitude are achievable, though such compression consistently incurs disproportionate accuracy loss on minority emotion classes, a finding with direct fairness implications given that minority classes such as fear and disgust are frequently also those of greatest clinical or safety relevance in applications such as distress detection. This pattern, wherein aggregate performance metrics obscure systematic degradation concentrated in specific subpopulations of either individuals or emotion categories, recurs throughout the performance evaluation literature and directly motivates the more disaggregated, fairness-aware evaluation practices advocated in Section 12 and the broader methodological reforms discussed as future research directions in Section 13.

11. Challenges and Limitations

Despite the substantial empirical progress documented throughout this review, deep learning–based emotion recognition confronts a set of persistent, structurally rooted challenges that distinguish it from many other deep learning application domains and that collectively constrain the field's readiness for unconstrained real-world deployment. This section critically examines these challenges, organized around data-related, methodological, and conceptual limitations.

Data scarcity and annotation subjectivity constitute the most immediately practical constraint on the field. Emotion-labeled corpora, even the largest currently available, remain modest in scale relative to the corpora underlying progress in adjacent domains such as general image classification or large language model pretraining, a gap attributable both to the substantially higher cost of emotion annotation, which typically requires multiple human raters per sample to achieve acceptable inter-rater reliability, and to legitimate privacy and consent concerns surrounding the collection of facial video, speech recordings, and physiological data, particularly from vulnerable populations. Annotation subjectivity compounds this scarcity: emotional expression interpretation varies measurably across annotators due to differences in cultural background, individual empathic disposition, and even momentary annotator mood state, producing label noise that establishes a ceiling on achievable model accuracy independent of architectural sophistication, since a portion of the discrepancy between model prediction and ground-truth label reflects genuine annotator disagreement rather than model error. Reported inter-annotator agreement on categorical emotion labeling, commonly quantified using Cohen's or Fleiss' kappa, frequently falls in the moderate range for naturalistic, spontaneous expression data, in contrast to the higher agreement typically achieved on posed or acted expression corpora, directly illustrating the tension between annotation reliability and ecological validity that pervades dataset construction decisions across the field.

Domain shift and cross-corpus generalization, empirically documented in Sections 6 and 10, represent a closely related but architecturally distinct challenge. Models trained on a specific corpus implicitly learn not only emotion-discriminative features but also corpus-specific nuisance correlations, including camera and microphone characteristics, recording environment acoustics, and demographic composition of the corpus population, that do not transfer to novel deployment contexts. While domain adaptation techniques, including adversarial feature alignment and self-supervised pretraining on unlabeled target-domain data, have demonstrated measurable mitigation of this gap, no currently published technique eliminates it, and the magnitude of residual domain shift remains substantial enough to caution strongly against deploying models validated exclusively on laboratory-collected corpora into naturalistic, in-the-wild application contexts without additional target-domain validation.

A more fundamental conceptual limitation concerns the adequacy of the categorical and dimensional emotion models themselves as targets for machine learning. Both frameworks, though methodologically convenient, represent substantial simplifications of the underlying psychological phenomenon: emotional experience is frequently ambivalent, context-dependent, and resistant to reduction into a small number of discrete categories or a low-dimensional continuous space, and there remains active, unresolved debate within affective science itself regarding the universality of basic emotion categories across cultures, with substantial evidence indicating that facial expression interpretation, in particular, is considerably more culturally variable than early cross-cultural studies suggested [16]. This conceptual tension implies that even a hypothetical model achieving perfect agreement with its training annotations would not necessarily constitute a valid general-purpose emotion recognizer, since the annotation scheme itself may inadequately represent the phenomenon under study, a limitation that architectural or data-scale improvements alone cannot resolve and that instead calls for closer, more sustained collaboration between machine learning researchers and affective science researchers in dataset and taxonomy design.

Robustness to naturalistic noise conditions constitutes a further practical limitation with direct implications for the application domains surveyed in Section 7. Facial expression recognition performance degrades measurably under partial occlusion, common in real-world settings due to face coverings, hands, or hair, and under extreme pose or illumination variation atypical of the relatively controlled conditions under which most training corpora are collected. Speech emotion recognition similarly degrades under background noise and channel distortion characteristic of telephony or far-field microphone deployment, conditions substantially different from the close-microphone, low-noise recording conditions of most benchmark corpora. Multimodal fusion offers partial mitigation of these single-modality robustness limitations, since a degraded or missing modality can in principle be compensated for by the remaining modalities, but this robustness benefit is realized only when the fusion architecture is explicitly designed and trained to handle missing or degraded modality conditions, a design consideration frequently absent from architectures optimized purely for accuracy under the clean, complete-modality conditions typical of benchmark evaluation, representing an identifiable gap between benchmark-optimized research practice and the robustness requirements of real deployment.

Finally, the computational and energy cost of state-of-the-art self-supervised pretrained backbones, while yielding substantial accuracy benefits as documented in Section 6, imposes practical deployment constraints particularly relevant to edge and embedded application contexts, creating a persistent tension between the pursuit of maximal recognition accuracy through increasingly large pretrained models and the latency, memory, and power constraints of real-world deployment platforms, a tension that motivates the model compression research discussed in Section 10 but that remains, at present, only partially resolved. Collectively, these challenges indicate that continued progress in emotion recognition will require advances not only in architectural sophistication but in dataset construction methodology, cross-disciplinary theoretical grounding, and robustness-oriented evaluation practice, themes that directly inform the future research directions articulated in Section 13.

12. Ethical and Societal Considerations

The deployment of emotion recognition technology raises ethical and societal considerations of a character and magnitude that distinguish it from many other machine learning application domains, arising from the technology's direct engagement with an intimate and involuntarily expressed dimension of human experience. This section critically examines these considerations across the dimensions of consent and privacy, bias and fairness, potential for misuse, and the contested scientific validity underlying commercial deployment claims.

Consent and privacy considerations are particularly acute given that emotional expression, unlike many other forms of personal data, is frequently produced involuntarily and continuously during ordinary social interaction, raising the possibility that emotion recognition systems could be deployed to infer affective states from individuals who have not meaningfully consented to such analysis, particularly in public or semi-public settings such as retail environments, workplaces, or public transportation where cameras and microphones may be present for ostensibly unrelated purposes such as security monitoring. This concern has motivated explicit regulatory attention: the European Union's Artificial Intelligence Act designates emotion recognition systems deployed in workplace and educational settings as carrying prohibited or high-risk status in most circumstances, reflecting a policy judgment that the potential for harm, including discriminatory employment decisions informed by inferred emotional states, outweighs the technology's purported benefits in these specific contexts [17]. Researchers and system designers bear a corresponding responsibility to implement affirmative, informed consent mechanisms wherever emotion recognition is deployed, to provide clear disclosure of data retention and downstream use practices, and to

critically evaluate whether a given application genuinely requires continuous affective monitoring or whether less privacy-invasive alternatives could achieve comparable functional outcomes.

Bias and fairness considerations manifest concretely in the demographic performance disparities documented empirically throughout this review, including the case study presented in Section 8, where system performance was shown to degrade for users from underrepresented linguistic and cultural backgrounds. These disparities trace directly to the demographic composition of training corpora, which, as noted in Section 9, remain predominantly skewed toward Western, English-speaking populations, and to the cultural variability in emotional expression norms discussed in Section 11, which implies that a model trained predominantly on one cultural population's expression patterns may systematically misclassify the expression patterns of other populations, a risk with direct consequences in high-stakes application contexts such as the clinical mental health and automotive safety domains surveyed in Section 7, where systematic misclassification for specific demographic subgroups could translate into inequitable access to beneficial system functionality or, worse, inequitable exposure to false-positive interventions such as unwarranted flagging for clinical concern. Mitigating these disparities requires deliberate, resource-intensive investment in demographically diverse and culturally representative dataset construction, together with disaggregated, subgroup-level performance evaluation as a standard reporting practice rather than the aggregate accuracy reporting that currently predominates in the literature, a methodological reform this review explicitly endorses as a necessary complement to the accuracy-oriented evaluation practices discussed in Section 10.

The potential for misuse of emotion recognition technology constitutes a further significant societal concern, encompassing applications such as covert emotional manipulation in advertising and political messaging, workplace emotional surveillance that could enable discriminatory management practices or create a chilling effect on authentic employee expression, and the deployment of emotion recognition in law enforcement or border security contexts, where the technology's contested scientific validity, discussed below, is particularly concerning given the severity of potential consequences from misclassification. Several jurisdictions and professional bodies have consequently called for moratoria or outright prohibition on emotion recognition deployment in specific high-stakes contexts, reflecting a precautionary policy stance that this review regards as substantively justified given the current state of the technology's demonstrated reliability limitations documented in Sections 6, 10, and 11.

Finally, the scientific validity of commercially deployed emotion recognition systems warrants explicit critical scrutiny, given a documented gap between the confident performance claims of some commercial emotion recognition vendors and the more circumscribed, caveat-laden findings of the peer-reviewed research literature surveyed throughout this review. A widely cited meta-analytic review of the psychological evidence underlying the assumption that discrete emotional categories can be reliably inferred from facial expression alone found the evidence for this assumption considerably weaker and more context-dependent than commonly presumed in commercial applications, concluding that facial configurations previously assumed to be universal and reliable indicators of specific emotions vary substantially across studies, contexts, and populations [16]. This finding directly undermines the scientific foundation of a substantial fraction of commercially deployed facial emotion recognition products and underscores the responsibility of researchers, reviewers, and technology deployers to maintain calibrated skepticism toward performance and validity claims that are not substantiated by rigorous, demographically diverse, and ecologically valid evaluation, a responsibility this review argues should extend to more conservative marketing and deployment practices industry-wide, particularly in consequential application domains such as those surveyed in Section 7.

13. Future Research Directions

Building on the challenges and ethical considerations articulated in the preceding two sections, this section outlines promising research directions that the authors regard as particularly likely to advance the field toward more accurate, robust, generalizable, and responsibly deployed emotion recognition systems.

Foundation-model-driven affective reasoning represents a natural extension of the self-supervised pretraining trend documented throughout this review, wherein large multimodal foundation models, pretrained on web-scale image, audio, video, and text corpora and subsequently instruction-tuned or fine-tuned on comparatively small emotion-labeled corpora, may offer substantially improved generalization relative to the modality-specific pretrained encoders currently dominant in the field, owing to the richer, more semantically grounded representations learned during large-scale multimodal pretraining. Preliminary studies evaluating large vision-language and multimodal large language models on emotion recognition tasks report competitive performance

relative to specialized architectures on several benchmarks, though with notable inconsistency across emotion categories and continued vulnerability to the demographic performance disparities discussed in Section 12, indicating that foundation-model adoption alone does not automatically resolve the field's generalization and fairness challenges and that continued targeted evaluation and fine-tuning remain necessary [18].

Neuro-symbolic and appraisal-theoretic integration constitutes a second promising direction, motivated directly by the conceptual limitations of purely data-driven categorical and dimensional emotion models discussed in Section 11. Hybrid architectures that incorporate structured, psychologically grounded appraisal-theoretic reasoning, for instance explicitly modeling the goal-relevance and coping-potential appraisal dimensions theorized to underlie emotional response generation, alongside conventional deep feature extraction, offer a potential pathway toward emotion recognition systems that better capture the context-dependent and situationally grounded character of genuine emotional experience rather than treating emotion recognition purely as pattern matching against surface-level expressive cues, though this direction remains comparatively nascent and would benefit substantially from closer collaboration between machine learning and affective science researchers, echoing the interdisciplinary imperative articulated in Section 11.

Personalized and continual learning approaches address the substantial inter-individual variability in emotional expression documented throughout this review, particularly acute for physiological modalities as shown in Section 6, by allowing models to incrementally adapt to an individual user's idiosyncratic expressive patterns over the course of sustained interaction while retaining generalizable population-level knowledge acquired during initial training, formulated as a continual learning problem requiring architectures resistant to catastrophic forgetting of prior knowledge upon exposure to new, individual-specific data. Few-shot personalization techniques, requiring only a small number of individual-specific calibration samples, represent a practically important special case of this direction with direct relevance to the subject-independent generalization challenges discussed in Section 10, since even modest individual calibration has been shown in preliminary studies to substantially narrow the gap between subject-dependent and subject-independent performance for physiological emotion recognition.

Robustness-oriented and uncertainty-aware architectures constitute a further priority direction, directly responsive to the noise robustness and domain shift challenges discussed in Section 11. This includes continued development of multimodal architectures explicitly trained under simulated missing-modality and degraded-modality conditions rather than exclusively on clean, complete-modality benchmark data, and the incorporation of calibrated uncertainty quantification, allowing deployed systems to abstain from confident prediction under conditions of high input ambiguity or out-of-distribution input, a capability of particular importance for the high-stakes clinical and automotive safety application domains surveyed in Section 7, where a confidently wrong prediction carries substantially greater cost than an appropriately flagged low-confidence prediction deferred to human judgment.

Finally, dataset and evaluation methodology reform represents perhaps the most immediately actionable future direction, requiring sustained community investment in demographically and culturally diverse dataset construction, standardized adoption of subject-independent and cross-corpus evaluation protocols as reporting requirements rather than optional supplementary analysis, and disaggregated subgroup-level performance reporting as a default rather than exceptional practice, extending the specific recommendations articulated in Sections 10 and 12 into a broader call for methodological standardization analogous to reporting guideline initiatives that have improved reproducibility and fairness accountability in adjacent fields such as clinical machine learning. Pursued collectively, these five directions—foundation-model integration, neuro-symbolic hybridization, personalized continual learning, robustness and uncertainty awareness, and methodological reform—chart a research agenda capable of addressing the field's most consequential open limitations while preserving the substantial empirical progress documented throughout this review.

14. Conclusion

This review has traced the multidisciplinary landscape of deep learning–based emotion recognition for human–computer interaction, synthesizing methodological developments across facial, vocal, physiological, textual, and multimodal modalities within a unified analytical framework. The evidence surveyed demonstrates substantial and continuing empirical progress: self-supervised pretraining has meaningfully mitigated the field's characteristic data scarcity, cross-modal attention mechanisms have enabled multimodal systems to reliably outperform their unimodal constituents, and the maturation of software tooling has lowered barriers to reproducible research. At the same time, this review has emphasized that benchmark-reported performance substantially overstates the

field's readiness for unconstrained real-world deployment, given the consistently documented degradation of recognition accuracy under cross-subject, cross-corpus, and naturalistic noise conditions, and given the conceptual limitations of the categorical and dimensional emotion models that underlie most current annotation and evaluation practice. The ethical considerations surrounding consent, demographic fairness, and the contested scientific validity of commercial deployment claims further underscore that continued technical progress must be accompanied by correspondingly rigorous ethical and regulatory attention. The future research directions articulated in this review—spanning foundation-model integration, neuro-symbolic hybridization, personalized continual learning, robustness-oriented architecture design, and methodological reform in dataset construction and evaluation practice—collectively chart a path toward emotion recognition systems that are not only more accurate but more genuinely generalizable, equitable, and responsibly deployed. Realizing this path will require sustained collaboration across the machine learning, affective science, human-computer interaction, and policy communities that this review has sought to bring into shared analytical view, in service of human-computer interaction systems capable of engaging with the emotional dimension of human experience with both technical sophistication and appropriate humility regarding the complexity of that which they seek to recognize.

References

[1] R. W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.
[2] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in Nebraska Symposium on Motivation, vol. 19, 1971, pp. 207–283.
[3] J. A. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178, 1980.
[4] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "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.
[5] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008.
[6] S. K. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," Learning and Instruction, vol. 22, no. 2, pp. 145–157, 2012.
[7] F. Eyben, K. R. Scherer, B. W. Schuller, et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190–202, 2016.
[8] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in Proc. Interspeech 2021, 2021, pp. 3400–3404.
[9] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," IEEE Transactions on Affective Computing, vol. 11, no. 3, pp. 532–541, 2020.
[10] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4040–4054.
[11] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019, pp. 6558–6569.
[12] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 1103–1114.
[13] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 154–164.
[14] J. Parry, D. Palaz, G. Clarke, et al., "Analysis of deep learning architectures for cross-corpus speech emotion recognition," in Proc. Interspeech 2019, 2019, pp. 1656–1660.
[15] J. Gratch et al., "The distress analysis interview corpus of human and computer interviews," in Proc. 9th International Conference on Language Resources and Evaluation (LREC), 2014, pp. 3123–3128.
[16] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," Psychological Science in the Public Interest, vol. 20, no. 1, pp. 1–68, 2019.
[17] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, 2024.
[18] Z. Lian, L. Sun, M. Xu, et al., "Explainable multimodal emotion reasoning," IEEE Transactions on Affective Computing, 2024.