International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct- 2026, PP 01-06

Emotion Recognition Using Artificial Intelligence: A Review Of Facial, Speech And Multimodal Affective Computing

Kashid Aditi Eknath; Prof. M. C. Kolse; Prof. S. K. Said
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India;

Abstract—Emotion recognition enables computers to infer human emotional states such as happiness, sadness, anger, fear, surprise and neutrality from facial expressions, voice and other behavioural signals. It underpins affective computing applications in healthcare, education, customer service, driver monitoring and human–computer interaction. This paper reviews AI-based emotion recognition. It introduces categorical and dimensional models of emotion, then surveys facial expression recognition from hand-crafted features to deep convolutional networks trained on laboratory and in-the-wild datasets, speech emotion recognition from acoustic features and deep models, and multimodal fusion that combines face, voice and text. Benchmark datasets including CK+, FER-2013, AffectNet and RAVDESS are discussed. Scientific and ethical debates are examined, particularly evidence that facial movements do not map reliably onto inner emotions across people and contexts, along with privacy and consent concerns. A five-layer architecture comprising input capture, preprocessing, feature learning, fusion and classification, and application with feedback is proposed, together with research gaps, objectives and a comparison of modalities. The review concludes that emotion recognition can make interaction more natural when used transparently, with consent and with awareness of its limits.

Index Terms—Emotion recognition, affective computing, facial expression recognition, speech emotion recognition, deep learning, multimodal fusion, ethics

I. Introduction

People constantly read each other's emotions from faces, voices and words, and adjust how they respond. Computers, by contrast, have traditionally ignored how users feel. Picard [1] coined the term affective computing for systems that recognize, interpret and respond to human emotion.

AI-based emotion recognition has advanced quickly with deep learning and large datasets. It is used to detect frustration in online learners, monitor patient mood, measure customer reactions and detect driver distress. At the same time, scientists have questioned how accurately emotions can be inferred from external signals, and privacy concerns are significant.

This paper reviews emotion recognition methods, proposes a system architecture and discusses technical and ethical challenges.

II. Theoretical Background

2.1 Models of Emotion

Categorical models use discrete emotions; Ekman and Friesen [2] proposed basic emotions with facial expressions recognized across cultures. Dimensional models describe emotion by valence (pleasantness) and arousal (activation). Fig. 1 summarizes emotion models and input modalities.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct- 2026, PP 01-06

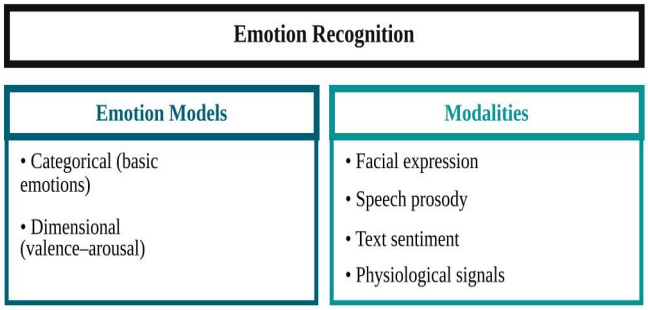


Fig. 1. Emotion models and recognition modalities.

B. Input Modalities

Facial images capture expressions; speech carries emotion in pitch, energy and rhythm; text conveys sentiment; and physiological signals such as heart rate reflect arousal.

III. Survey of Existing Approaches

Facial expression. Lucey *et al.* [3] released the extended Cohn–Kanade (CK+) dataset of posed expressions with action-unit labels. Goodfellow *et al.* [4] introduced FER-2013, a large set of low-resolution in-the-wild faces, through a representation-learning contest. Mollahosseini *et al.* [5] created AffectNet, with about a million web images annotated for categories, valence and arousal. Li and Deng [6] surveyed deep facial expression recognition and identified challenges of occlusion, pose and dataset bias.

Speech. Schuller [7] reviewed two decades of speech emotion recognition, from prosodic features with classical classifiers to deep learning. Livingstone and Russo [8] released RAVDESS, an audio-visual dataset of emotional speech and song by professional actors.

Multimodal. Poria *et al.* [9] reviewed affective computing from unimodal analysis to multimodal fusion and found that combining modalities generally improves accuracy.

Validity. Barrett *et al.* [10] reviewed scientific evidence and concluded that facial movements do not reliably and specifically indicate emotional states across people and contexts, cautioning against strong claims for facial emotion recognition. Table I summarizes these works.

TABLE I Summary of Literature Survey

Sr. No.	Ref.	Year	Methodology	Findings	Research Gap
1	[2]	1971	Cross-cultural facial expression study	Basic emotion categories	Posed photographs
2	[3]	2010	CK+ dataset	Action-unit and emotion labels	Posed, laboratory
3	[4]	2013	FER-2013 contest dataset	Large in-the-wild benchmark	Low resolution, noisy
4	[9]	2017	Multimodal affective computing review	Fusion improves accuracy	Synchronization complexity

5	[7]	2018	Speech emotion recognition review	Two decades of progress	Cross-corpus gaps
6	[5]	2019	AffectNet	Categories + valence–arousal	Annotation subjectivity
7	[10]	2019	Review of facial-emotion evidence	Limited reliability of facial inference	Challenges common assumptions
8	[6]	2022	Survey of deep FER	Mapped methods and challenges	Survey only

IV. Comparative Study

Table II compares modalities. Faces give rich cues but are affected by lighting and pose, speech works without a camera, text is easy to collect, and multimodal fusion is most robust.

TABLE II Comparison of Emotion Recognition Modalities

Modality	Accuracy	Robustness	Intrusiveness
Facial expression	High (posed)	Lighting / pose sensitive	Camera needed
Speech	Medium	Noise sensitive	Microphone needed
Text	Medium	Language dependent	Low
Physiological	Medium–High	Good	Wearable needed
Multimodal fusion	Highest	Most robust	Highest

V. Research Gaps

- Models trained on posed expressions perform poorly on spontaneous, real-world emotions.
- Datasets under-represent many ethnicities, ages and cultural expression styles.
- Inferring internal feelings from external signals has limited scientific validity.
- Multimodal systems are complex and require synchronized data.
- Privacy, consent and misuse in surveillance or hiring are serious concerns.

VI. Aims and Objectives

- To review emotion models and recognition modalities.
- To examine deep learning methods for facial and speech emotion recognition.
- To propose a multimodal emotion recognition architecture.
- To compare modalities in accuracy, robustness and intrusiveness.
- To discuss ethical, privacy and validity challenges.

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct- 2026, PP 01-06

VII. Proposed Architecture

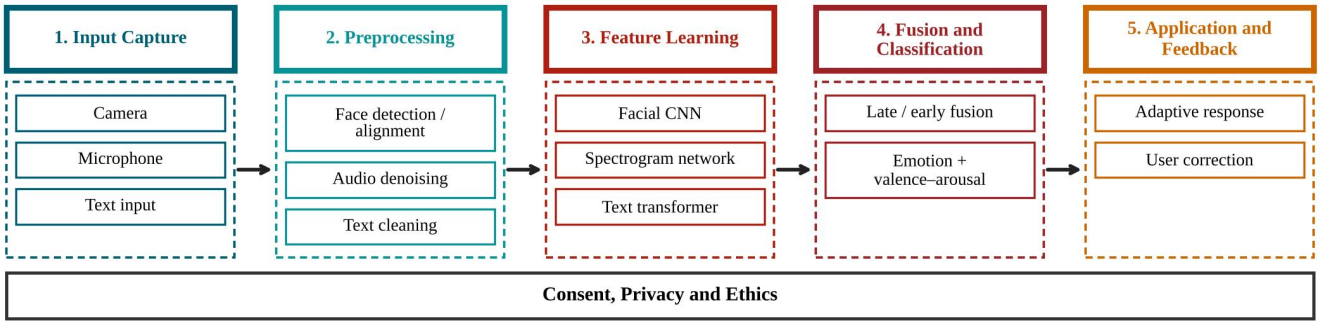


Fig. 2 shows the proposed architecture.

Fig. 2. Proposed multimodal emotion recognition architecture.

A. Input Capture Layer

A camera captures facial video, a microphone captures speech and, optionally, typed text is collected, all with user consent.

7.2 Preprocessing Layer

Faces are detected and aligned; audio is denoised and segmented; text is cleaned.

C. Feature Learning Layer

A CNN extracts facial expression features, a spectrogram-based network extracts vocal features and a transformer encodes text.

7.3 Fusion and Classification Layer

Features are fused and classified into emotion categories and valence–arousal estimates, with a confidence score.

7.4 Application and Feedback Layer

Applications adapt their behaviour, for example by offering help to a frustrated learner, and users can correct the system's estimates.

VIII. Process Flow and Evaluation

Fig. 3 shows the processing pipeline. A classifier outputs probabilities over K emotion classes using softmax

$$P(k \mid x) = e^{(z_k)} / \Sigma_j e^{(z_j)} \quad (1)$$

and, in late fusion, the final probability combines modality outputs with weights,

$$P = w_f \cdot P_face + w_s \cdot P_speech + w_t \cdot P_text \quad (2)$$

International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct- 2026, PP 01-06

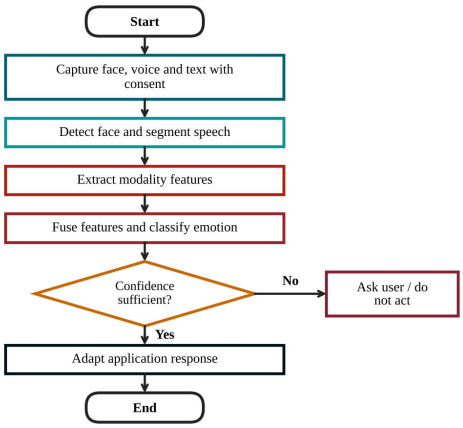


Fig. 3. Emotion recognition pipeline.

Performance is reported with accuracy, weighted F1-score and, for continuous dimensions, concordance correlation.

IX. Limitations and Future Scope

Table III summarizes challenges. Future directions include context-aware models, personalized calibration, privacy-preserving on-device processing, culturally diverse datasets and clear regulation restricting high-risk uses.

Challenge	Mitigation
Posed vs. spontaneous	In-the-wild training data
Dataset bias	Diverse, balanced datasets
Scientific validity	Context use; avoid strong claims
Privacy	Consent; on-device processing
Misuse	Restrict high-risk applications

TABLE III Challenges and Mitigation

X. Conclusion

AI-based emotion recognition can make human–computer interaction more natural and supportive. The reviewed literature shows strong progress in facial, speech and multimodal recognition driven by deep learning and large datasets, but also important limits in scientific validity, bias and privacy. Emotion recognition is best used transparently, with consent, as one input among many rather than as a definitive judgement of how a person feels.

10.1 Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

References

1. R. W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.

2. P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion,” J. Pers. Soc. Psychol., vol. 17, no. 2, pp. 124–129, 1971.
3. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,” in Proc. IEEE CVPR Workshops, 2010, pp. 94–101.
4. I. J. Goodfellow et al., “Challenges in representation learning: A report on three machine learning contests,” in Proc. Int. Conf. Neural Inf. Process. (ICONIP), 2013, pp. 117–124.
5. A. Mollahosseini, B. Hasani, and M. H. Mahoor, “AffectNet: A database for facial expression, valence, and arousal computing in the wild,” IEEE Trans. Affect. Comput., vol. 10, no. 1, pp. 18–31, 2019.
6. S. Li and W. Deng, “Deep facial expression recognition: A survey,” IEEE Trans. Affect. Comput., vol. 13, no. 3, pp. 1195–1215, 2022.
7. B. W. Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” Commun. ACM, vol. 61, no. 5, pp. 90–99, 2018.
8. S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS),” PLoS ONE, vol. 13, no. 5, Art. no. e0196391, 2018.
9. S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” Inf. Fusion, vol. 37, pp. 98–125, 2017.
10. 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,” Psychol. Sci. Public Interest, vol. 20, no. 1, pp. 1–68, 2019.