# A Comprehensive Review of Multimodal Emotion Recognition Techniques Using Machine Learning

*D. Ujwalla Gawande[1], Dr. Hemalatha[2]*
*[1]Global Postdoctoral & Researcher Programme , Lincoln University College Malaysia*
*[1]Professor, Department of IT, Yeshwantrao Chavan College of Engineering,*
*Nagpur, Maharashtra, India*

*[2]Professor*
*Department of computer science and business systems*
*Panimalar Engineering college*
*Poonamallee. Chennai. Tamilnadu. India*
*pithemalatha@gmail.com*

## Abstract

Emotion recognition allows intelligent systems to recognize and address human emotions, making human computer interaction better. The single-modality-based approaches, i.e., speech or facial expressions alone, usually failed to recognize the full spectrum of human emotions under realistic conditions. The advance in machine learning and deep learning has given rise to particularly important multimodal and multilingual emotion recognition methods. In such a system, facial expressions, voice, and text are the different sources used by integrating these to improve the detection of emotion in various languages. This study present recent advance on multimodal and multilingual emotion recognition, while giving special emphasis on convolutional, recurrent, and transformer models. It covers systems like Multilingual Speech Emotion Recognition (MSER) and shows how they apply to India and other low-resource languages. It identifies emotion-aware music, activity, and education recommendation systems for their potential to boost personalization and user engagement. It discusses specific challenges on dataset imbalance, cultural differences, subjective emotional boundaries, and privacy issues. Finally, this paper provides a direction for creating contextually aware, adaptive, and ethically sensitive emotion recognition systems.

**Keywords: Emotion recognition, Multimodal learning, Multilingual speech emotion recognition, Deep learning, Affective computing, Personalized recommendation**

## Introduction

Emotion recognition is one of the important areas in artificial intelligence because it is applied in human-computer interaction, behavioral analysis, and affective computing. Many work in the same area has been perform to create machines that will be able to identify and then respond to human emotions through facial expressions, intonation of voice, gestures, and language patterns [7], [8], [27]. Various ER technology is used in adaptive learning, mental health monitoring, and entertainment, among other more advanced personal assistants, where emotion understanding allows for better personalization and user experience [1], [3], [21].

Most of the Emotion recognition works generally used unimodal data, e.g., either facial expressions or speech, with handcrafted features like Mel-Frequency Cepstral Coefficients (MFCC) or Local Binary Patterns (LBP) [25], [30]. These methods utterly fail, however, under real-world circumstances because affect is multimodal and culturally biased. Single-modality may fail in occlusion, noise, or linguistic variation, which motivated the development of multimodal emotion recognition systems that make use of multiple channels of information [9], [32], [36].

Machine learning and deep learning have also made emotion recognition easier by the automatic extraction of features and learning of representations. The main use of CNN is for the visual and

audio analyses, while hybrid models such as CNN-LSTM networks describe both spatial and temporal structures of emotional signals [9], [27]. Recent works have been using attention-based and transformer architectures that yield state-of-the-art performance by modeling effectively long-range dependencies and cross-modal relationships effectively [5], [34], [35]. These models not only improve the accuracy of recognition but also assist in understanding the development of emotions across various modalities.

This attracts interest in the research of multilingual emotion detection with emotion-aware systems being rolled out across the world. Variations in phonetics, prosody, and linguistic context across different languages make it challenging to design generalized systems [15, 37]. Models trained on English corpora fail to capture emotions correctly across other languages, including Indian and low-resource ones. Cross-lingual adaptation, transfer learning, and pre-trained models like wav2vec 2.0 embeddings are effective remedies [35, 37].

Recent works also focus on emotion-aware recommendation systems with dynamic adaptation of content, music, or activities considering user emotions [1], [2], [4], [14]. By utilizing multi-modal cues, these applications provide context-sensitive personalized responses that further increase user engagement. For example, the inclusions of emotion recognition in e-learning systems allow educators to monitor student engagement and emotional states in real-time manners [6].

Even with their huge success, they all have some challenges. The number of labels in multimodal datasets is always limited, and the boundaries of emotions are not objective. More to these are issues related to fusing heterogeneous modalities leading to synchronization and interpretability problems [16], [22], [34]. Besides that, multilingual emotion recognition raises a lot of ethical and privacy concerns about cross-cultural data gathering and use.

This paper presented an overall review on multimodal and multilingual emotion recognition using machine learning and deep learning. Section 2 briefed on unimodal and multimodal techniques, Section 3 reviewed multilingual and cross-lingual systems, Section 4 discussed emotion-aware applications, and Section 5 mentioned challenges and future research directions in the area of affective computing.

## 2. Literature Review

Recent advances in machine learning, deep learning, and affective computing have fostered development in emotion recognition systems. The presented work has evolved from unimodal to multimodal and multilingual methods of analyzing facial, speech, and text cues together during the last decade. Thereby, these systems are closing the gap between human expression of emotion and computational understanding by learning data-driven representations of emotions that generalize across cultural and linguistic settings. Major advances in recognition of facial, speech, multimodal, and multilingual emotion recognition and their applications with mention of inherent challenges are discussed in the following subsections.

### 2.1 Facial Emotion Recognition

Facial Emotion Recognition is arguably the most investigated area in affective computing since facial expressions are amongst the richest forms of nonverbal communication. Earlier FER systems used handcrafted features such as LBP, Gabor filters, and Eigenfaces. With the advent of deep learning, all these aforementioned features became automated under the convolutional architectures that pushed the accuracy even further. Kaur and Kumar [7] summarized deep learning-based FER approaches and described that CNNs consistently outperform old-fashioned classifiers in both controlled and naturalistic

conditions. Meena et al. [27] proposed a deep CNN model trained on the FER-2013 dataset with over 91% accuracy for simple emotions. Lee et al. [28] proposed a knowledge-distillation model that reduced inference time without losses in terms of accuracy. Wang et al. [29] presented OAENet, a spatial-oriented attention ensemble network able to sense fine-grained spatial relationships and featuring an accuracy of 92%. Manalu and Rifai [9] combined CNN and RNN layers in order to represent temporal changes of facial sequences and received better emotion discrimination from dynamic video frames. Bakariya et al. [30] connected music recommendation with facial emotion detection, thus proving applicability for FER in adaptive user experience systems. To put it in a nutshell, all these experiments already position FER as a solid real-time scalable technology for emotion perception.

Table 2.1: Facial Emotion Recognition Summary

| **Reference** | **Dataset** | **Algorithm / Model** | **Performance / Accuracy** |
|---|---|---|---|
| Kaur & Kumar [7] | FER-2013 | CNN | 91% |
| Meena et al. [27] | FER-2013 | Deep CNN | 91% |
| Lee et al. [28] | CK+ | Knowledge Distillation CNN | 89% |
| Wang et al. [29] | RAF-DB | OAENet (Oriented Attention Ensemble) | 92% |
| Manalu & Rifai [9] | CK+ | CNN + RNN | 90% |
| Bakariya et al. [30] | Custom Dataset | CNN + Recommendation | 88% |

**Issues and Limitations in FER**

1. **Environmental sensitivity:** Most current FER models perform perfectly in controlled environments but degrade for changing illumination conditions, occlusion, or head poses.
2. **Cultural variation:** Emotional expressions vary across cultures, limiting model generalization.
3. **Temporal dynamics:** Static image-based models cannot be that sensitive to sequential facial movements.
4. **Dataset bias:** Most large annotated datasets lie in limited domains, which hurts robustness to real-world scenarios.

While FER has achieved much towards maturity, generalization and dynamic recognition remain areas where it always lags. Typically, the multimodal approaches try to circumvent such limitations by combining facial cues with speech or text cues.

## 2.2 Speech Emotion Recognition

By analysing the building cures, Speech Emotion Recognition investigates acoustic and prosodic features such as tone, rhythm, and pitch which carry critical information. Early SER employed features such as MFCC, ZCR, and pitch contours as the input to an SVM or HMM classifier. Deep learning replaced these hand-crafted techniques with data-driven representations that capture subtle emotional messages.
Alhussein et al. [8] developed a comprehensive review of SER and documented that repetitive models like LSTM and GRU were highly robust across varying acoustic conditions. Oh et al. [10] proposed an autoencoder-based multi-detection SER system for noisy mobility settings, with the enhancement in emotion classification performance. Pepino et al. [35] employed wav2vec 2.0 embeddings to extract high-level features independent of the language of the raw audio, while improved cross-linguistic emotion recognition is observed. Somvanshi et al. [37] applied these techniques to multilingual scenarios, especially Indian languages, where deep models obtained substantially higher accuracy relative to traditional models. All these works point out SER as a primary central modality for end-to-end affective analysis.

Table 2.2: Speech Emotion Recognition Summary

| Reference | Dataset | Algorithm / Model | Performance / Accuracy |
|---|---|---|---|
| Alhussein et al. [8] | IEMOCAP | LSTM / GRU | 86% |
| Oh et al. [10] | Emo-DB | Autoencoder + Classifier | 85% |
| Pepino et al. [35] | MSP-IMPROV | wav2vec 2.0 embeddings | 88% |
| Somvanshi et al. [37] | Indian Languages | Multilingual LSTM | 85% |

**Issues and Limitations in SER**

1. **Noise sensitivity:** Acoustic features are sensitive to environmental noise.
2. **Language dependency:** Classical models do not generalize across languages.
3. **Data requirements:** Deep models require big annotated datasets, which hardly exist for any language.
4. **Speaker variation:** Variation in vocal pitch, rate, and style can decrease accuracy.

While SER is back in the spotlight of affective computing, it needs robust, multilingual, and noise-tolerant architectures before real-world applications can be reliably conducted.

### 2.3 Multimodal Fusion Systems

While unimodal methods can capture partial emotion cues, human emotion is inherently multimodal. MER integrates visual, auditory, and textual inputs into a unified representation. The most central problem is how to synchronously harmonize and merge heterogeneous data.
Chaudhari et al. [31] proposed a hybrid CNN-BiLSTM model that combined the facial and speech modalities and outperformed the best single-source baseline models. Tang et al. [32] used feature-fusion to generate context-relevant visual and audio features; spatial and temporal, respectively. Tomar et al. [36] utilized deep fusion networks that included

learning inter-modal correlations simultaneously to achieve over 93% accuracy on standard benchmark datasets. Ahmed et al. [22] classified fusion methods into early, late, and hybrid; however, emphasized that only attention-based transformers provide the most effective fusion in a multimodal fusion. N. S. et al. [21] deployed an AI-Bot that merged facial and voice information for tracking emotional states in mental well-being applications.

Table 2.3: Multimodal Emotion Recognition Summary

| Reference | Dataset | Algorithm / Model | Performance / Accuracy |
|---|---|---|---|
| Chaudhari et al. [31] | CMU-MOSEI | CNN (Face) + BiLSTM (Speech) | 92% |
| Tang et al. [32] | RAVDESS | Feature Fusion (Visual + Audio) | 91% |
| Tomar et al. [36] | IEMOCAP | Deep Fusion Network | 93% |
| Ahmed et al. [22] | Multiple | Transformer-based Fusion | 94% |
| N. S. et al. [21] | Custom | Multimodal AI-Bot | 90% |

**Limitations and Issues in MER**

1. **Alignment:** Temporal misalignment between modalities will reduce performance.
2. **Computational cost:** Multimodal models are more computationally expensive than unimodal ones.
3. **Data scarcity:** Collecting synchronized, labeled multimodal datasets is complicated.
4. **Model complexity:** Fusion architectures are harder to train and less interpretable.

Despite these disadvantages, MER always does better than the unimodal methods of real-world and dynamic data.

## 2.4 Multilingual Emotion Recognition

Recognizing emotions across different languages comes with several challenges, including cultural differences, variations in speech sounds, and ambiguous meanings. Multilingual Emotion Recognition (MLER) aims to create models that can interpret emotional signals regardless of language. Cai et al. [15] reviewed methods for emotion recognition in multiple languages and pointed out the importance of transfer learning and multilingual embedding spaces. Pepino et al. [35] used self-supervised wav2vec 2.0 embeddings to capture acoustic features that are less dependent on language, which improved cross-lingual performance. Somvanshi et al. [37] studied deep learning models for Multilingual Speech Emotion Recognition (MSER) in Indian languages, using multilingual fine-tuning to deal with low-resource datasets.

Ahmed et al. [22] highlighted the need for consistent multilingual datasets and context-aware feature learning to reduce bias between language groups. Their work shows that multilingual emotion recognition must go beyond simple translation and consider how emotions are expressed in different cultures. The field is gradually moving toward universal models that combine linguistic and non-linguistic cues, aiming for more generalizable, global emotion-aware systems.

Table 2.3: Multimodal Emotion Recognition Summary

| Reference | Dataset | Algorithm / Model | Performance / Accuracy |
|---|---|---|---|
| Chaudhari et al. [31] | CMU-MOSEI | CNN (Face) + BiLSTM (Speech) | 92% |
| Tang et al. [32] | RAVDESS | Feature Fusion (Visual + Audio) | 91% |
| Tomar et al. [36] | IEMOCAP | Deep Fusion Network | 93% |
| Ahmed et al. [22] | Multiple | Transformer-based Fusion | 94% |
| N. S. et al. [21] | Custom | Multimodal AI-Bot | 90% |

## 2.5 Applications and Challenges

Emotion recognition is being used in many real-world systems. Some systems try to suggest music, activities, or learning content based on a user's mood, which can help keep them engaged [1], [2], [4], [5], [14], [30]. For example, Aly [6] found that tracking student emotions during online classes can improve learning outcomes. Sarala et al. [1] created a system that recommends tasks for each learner, and N. S. et al. [21] developed a multimodal bot that looks at both voice and facial expressions to assess mental health.

Even with these developments, there are still challenges. High-quality datasets are hard to get, and many models only work for certain languages. How emotions are expressed can vary a lot between cultures, which makes recognition tricky. Combining data from different sources is also not easy, especially if the timing or sampling rates do not match. Privacy and bias remain major concerns [16], [22], [34]. The environment, social context, and the speaker's intentions can also affect results. Future research will need to focus on fairness, explainability, protecting data, and making systems more adaptive so they can work in real time across different settings.

**Issues and Limitations in MLER**

1. **Low-resource languages:** Small annotated datasets restrict model accuracy.
2. **Cultural differences:** Emotional expressions vary in meaning and intensity across cultures.
3. **Feature harmonization:** Aligning features across languages and modalities remains challenging.
4. **Model adaptation:** Transfer learning requires careful tuning to prevent performance loss.

Research in multimodal and multilingual emotion recognition is now aiming at models that can work across cultures and languages.

## Conclusion

Multimodal and multilingual emotion recognition (MMER) has emerged as a pivotal area in affective computing, enabling intelligent systems to perceive and respond to human emotions in complex, real-world scenarios. This review highlights the progression from traditional unimodal approaches to sophisticated deep learning and transformer-based architectures capable of integrating facial, speech, textual, and physiological signals. Multimodal fusion and cross-lingual adaptation have demonstrated significant improvements in recognition accuracy, robustness, and context-awareness. Applications in education, healthcare, entertainment, human–robot interaction, and smart assistants underscore the transformative potential of these systems. Despite substantial advancements, challenges remain in modality fusion, dataset scarcity, cross-cultural generalization, real-time processing, and ethical considerations. Future research should focus on developing scalable, low-latency, and ethically responsible MMER frameworks, incorporating adaptive fusion strategies, universal affective representations, and culturally sensitive designs. Addressing these challenges will pave the way for next-generation emotion-aware systems that are robust, inclusive, and capable of enhancing human–computer interaction across diverse linguistic and cultural contexts.

## References

[1] P. Sarala, G. Neha, N. S. V. Tanmayi, and A. N. V. Reddy, “EMOREC: Personalized Emotion Recognition and task recommendation system for enhanced learning experiences,” 2022 International Conference on Inventive Computation Technologies (ICICT), Apr. 2024, doi: 10.1109/icict60155.2024.10545013.

[2] S. Khedkar, “Activity recommendation system based on emotion recognition,” INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, vol. 09, no. 06, pp. 1–9, Jun. 2025, doi: 10.55041/ijsrem49641.

[3] H. D. P. M. Hettiarachchi, M. N. Ekanayake, G. C. Kaveendhya, O. V. Koralage, P. Samarasekara, and D. Kasthurirathna, “Emotional State Monitoring with Personalized Activity Recommendations,” 2023 5th International Conference on Advancements in Computing (ICAC), Colombo, Sri Lanka, 2023, Pp. 436-441, pp. 436–441, Dec. 2023, doi: 10.1109/icac60630.2023.10417468.

[4] S. M. Florence and M. Uma, “Emotional Detection and Music Recommendation System based on User Facial Expression,” IOP Conference Series Materials Science and Engineering, vol. 912, no. 6, p. 062007, Aug. 2020, doi: 10.1088/1757-899x/912/6/062007.

[5] L. Zhao, G. Liu, S. Yan, and J. Zhang, “Emotion-driven music recommendation system based on fully convolutional recurrent attention networks and collaborative filtering,” Alexandria Engineering Journal, vol. 125, pp. 354–366, Apr. 2025, doi: 10.1016/j.aej.2025.03.114.

[6] M. Aly, “Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model,” Multimedia Tools and Applications, Jun. 2024, doi: 10.1007/s11042-024-19392-5.

[7] M. Kaur and M. Kumar, “Facial emotion recognition: A comprehensive review,” Expert Systems, vol. 41, no. 10, Jun. 2024, doi: 10.1111/exsy.13670.

[8] G. Alhussein, I. Ziogas, S. Saleem, and L. J. Hadjileontiadis, “Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis,” Artificial Intelligence Review, vol. 58, no. 7, Apr. 2025, doi: 10.1007/s10462-025-11197-8.

[9] H. V. Manalu and A. P. Rifai, “Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm,” Intelligent Systems With Applications, vol. 21, p. 200339, Feb. 2024, doi: 10.1016/j.iswa.2024.200339.

[10] J. M. Oh, J. K. Kim, and J. Y. Kim, “Multi-Detection-Based speech emotion recognition using Autoencoder in mobility service environment,” Electronics, vol. 14, no. 10, p. 1915, May 2025, doi: 10.3390/electronics14101915.

[11] Z. Leng et al., “Emotion Recognition on the Go: Utilizing Wearable IMUs for Personalized Emotion Recognition,” UbiComp ’24: Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing, vol. 1, pp. 537–544, Sep. 2024, doi: 10.1145/3675094.3678452.

[12] A. Salvi, P. Shevkar, and Prof. Harshil Kanakia, “Emotion based Music Recommendation using FaceRecognition,” journal-article, 2024. [Online]. Available: https://www.ijnrd.org/papers/IJNRD2410111.pdf

[13] J. Hou, “Deep Learning-Based Human Emotion Detection Framework using Facial Expressions,” Journal of Interconnection Networks, vol. 22, no. Supp01, Jan. 2022, doi: 10.1142/s0219265921410188.

[14] V. Vijayalakshmi, P. Shrivastav, and G. Thiyagarajan, “Facial expression-based AI system for personalized music recommendations,” in Advances in intelligent systems research/Advances in Intelligent Systems Research, 2025, pp. 273–285. doi: 10.2991/978-94-6463-738-0_23.

[15] Y. Cai, X. Li, and J. Li, “Emotion recognition using different sensors, emotion models, methods and datasets: A Comprehensive review,” Sensors, vol. 23, no. 5, p. 2455, Feb. 2023, doi: 10.3390/s23052455.

[16] C. Zhu, “Research on Emotion Recognition-Based Smart Assistant System: Emotional Intelligence and Personalized Services,” Journal of System and Management Sciences, vol. 13, no. 5, Sep. 2023, doi: 10.33168/jsms.2023.0515.

[17] H.-G. Kim, G. Y. Kim, and J. Y. Kim, “Music recommendation system using human activity recognition from accelerometer data,” IEEE Transactions on Consumer Electronics, vol. 65, no. 3, pp. 349–358, Jun. 2019, doi: 10.1109/tce.2019.2924177.

[18] R. Hou, “Music content personalized recommendation system based on a convolutional neural network,” Soft Computing, vol. 28, no. 2, pp. 1785–1802, Dec. 2023, doi: 10.1007/s00500-023-09457-2.

[19] C. Li, I. Ishak, H. Ibrahim, M. Zolkepli, F. Sidi, and C. Li, “Deep Learning-Based Recommendation System: Systematic Review and Classification,” IEEE Access, vol. 11, pp. 113790–113835, Jan. 2023, doi: 10.1109/access.2023.3323353.

[20] N. A. Albatayneh, K. I. Ghauth, and F.-F. Chua, “Discriminate2Rec: Negation-based dynamic discriminative interest-based preference learning for semantics-aware content-based recommendation,” Expert Systems With Applications, vol. 199, p. 116988, Mar. 2022, doi: 10.1016/j.eswa.2022.116988.

[21] S. N, M. K, B. Prasanth, J. C, V. S. K. B, and M. C. Prabhu, “AI-Bot Powered Emotion Detection and Mental Wellness System with Facial and Voice Analysis for Personalized Recommendations,” 2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE), pp. 1–6, May 2025, doi: 10.1109/iccrtee64519.2025.11053026.

[22] N. Ahmed, Z. A. Aghbari, and S. Girija, “A systematic survey on multimodal emotion recognition using learning algorithms,” Intelligent Systems With Applications, vol. 17, p. 200171, Jan. 2023, doi: 10.1016/j.iswa.2022.200171.

[23] P. Dandavate et al., “Emotion detection for adaptive experiences,” in Lecture notes in networks and systems, 2025, pp. 413–423. doi: 10.1007/978-981-96-7499-2_35.

[24] N. Holla, P. Prabal, N. H. U, N. Patil, and D. E. Nathaniel, “A personalized mental health chatbot with emotion detection and content suggestion,” Proceedings Volume 13660, International Conference on Advances in Photonics Science (ICAPS 2024); 136600J (2025), p. 31, Jun. 2025, doi: 10.1117/12.3070958.

[25] J. A. Ballesteros, G. M. Ramírez V., F. Moreira, A. Solano, and C. A. Pelaez, “Facial emotion recognition through artificial intelligence,” Frontiers in Computer Science, vol. 6, Jan. 2024, doi: 10.3389/fcomp.2024.1359471.

[26] A. Ezquerra, F. Agen, R. B. Toma, and I. Ezquerra-Romano, “Using facial emotion recognition to research emotional phases in an inquiry-based science activity,” Research in Science & Technological Education, pp. 1–24, Jul. 2023, doi: 10.1080/02635143.2023.2232995.

[27] G. Meena, K. K. Mohbey, A. Indian, M. Z. Khan, and S. Kumar, “Identifying emotions from facial expressions using a deep convolutional neural network-based approach,” Multimedia Tools and Applications, vol. 83, no. 6, pp. 15711–15732, Jul. 2023, doi: 10.1007/s11042-023-16174-3.

[28] K. Lee, S. Kim, and E. C. Lee, “Fast and accurate facial expression image classification and regression method based on knowledge distillation,” Applied Sciences, vol. 13, no. 11, p. 6409, May 2023, doi: 10.3390/app13116409.

[29] Z. Wang, F. Zeng, S. Liu, and B. Zeng, “OAENet: Oriented attention ensemble for accurate facial expression recognition,” Pattern Recognition, vol. 112, p. 107694, Oct. 2020, doi: 10.1016/j.patcog.2020.107694.

[30] B. Bakariya, A. Singh, H. Singh, P. Raju, R. Rajpoot, and K. K. Mohbey, “Facial emotion recognition and music recommendation system using CNN-based deep learning

techniques," Evolving Systems, vol. 15, no. 2, pp. 641–658, May 2023, doi: 10.1007/s12530-023-09506-z.

[31] A. Chaudhari, C. Bhatt, T. T. Nguyen, N. Patel, K. Chavda, and K. Sarda, "Emotion recognition system via facial expressions and speech using machine learning and deep learning techniques," SN Computer Science, vol. 4, no. 4, Apr. 2023, doi: 10.1007/s42979-022-01633-9.

[32] G. Tang, Y. Xie, K. Li, R. Liang, and L. Zhao, "Multimodal emotion recognition from facial expression and speech based on feature fusion," Multimedia Tools and Applications, vol. 82, no. 11, pp. 16359–16373, Nov. 2022, doi: 10.1007/s11042-022-14185-0.

[33] A. Vlachostergiou, G. Caridakis, and S. Kollias, "Investigating Context Awareness of Affective Computing Systems: A Critical Approach," Procedia Computer Science, vol. 39, pp. 91–98, Jan. 2014, doi: 10.1016/j.procs.2014.11.014.

[34] G. Pei, H. Li, Y. Lu, Y. Wang, S. Hua, and T. Li, "Affective Computing: recent advances, challenges, and future trends," Intelligent Computing, vol. 3, Dec. 2023, doi: 10.34133/icomputing.0076.

[35] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings," arXiv (Cornell University), Jan. 2021, doi: 10.48550/arxiv.2104.03502.

[36] P. S. Tomar, K. Mathur, and U. Suman, "Fusing facial and speech cues for enhanced multimodal emotion recognition," International Journal of Information Technology, vol. 16, no. 3, pp. 1397–1405, Jan. 2024, doi: 10.1007/s41870-023-01697-7.

[37] S. Somvanshi, A. Meher, P. Hagawane, M. Adhav, and J. Harne, "Multilingual speech emotion recognition using deep learning," International Journal of Scientific Research in Computer Science Engineering and Information Technology, pp. 183–189, Dec. 2024, doi: 10.32628/cseit2410772.