Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

TraceMind: A Multimodal Emotion Evaluation System for Psychologists

Abhay P
Dept. of CSE
APJ Abdul Kalam Tech. Univ.
Thrissur, Kerala

Adarsh K S
Dept. of CSE
APJ Abdul Kalam Tech. Univ.
Thrissur, Kerala

Anaswara Narayanan
Dept. of CSE
APJ Abdul Kalam Tech. Univ.
Thrissur, Kerala

Sam Lazar
Dept. of CSE
APJ Abdul Kalam Tech. Univ.
Thrissur, Kerala

Divya Jose
Assistant Professor & HOD(Coordinator)
Dept. of CSE, APJ Abdul Kalam Tech.
Univ. Thrissur, Kerala

***Abstract*—Psychological assessment often relies on subjective observation and self-reporting, which can be unreliable when patients are unable to articulate their emotional state due to trauma or alexithymia. This paper presents TraceMind, a multimodal AI system that assists clinicians by analyzing emotional states from patient video recordings. The system combines two independent streams: a visual stream using a ResNet-50 model fine-tuned on RAF-DB for facial expression recognition, and an auditory stream using a 1D Semi-CNN trained on a composite dataset combining RAVDESS, IEMOCAP, TESS, CREMA-D, and SAVEE. Both streams are synchronized into 3-second temporal windows and combined using a Weighted Average Fusion strategy ($\alpha = 0.6$) to produce a final emotion classification. Our experiments show that fusing both modalities improves detection accuracy for harder emotions like Fear and Anger compared to either stream alone. The system outputs a structured clinical PDF report with a temporal emotion profile of the session.**

***Index Terms*—Multimodal Emotion Recognition, ResNet-50, 1D Semi-CNN, Affective Computing, Psychological Assessment.**

I. Introduction

Emotion recognition is a practical need in clinical psychology and forensic investigation. Clinicians working with trauma patients, PTSD cases, or forensic subjects regularly face a core problem: patients either cannot accurately describe what they feel, or they deliberately conceal it. Standard tools like the Beck Depression Inventory depend on self-reporting, which breaks down exactly when it is needed most.

Two specific failure modes motivated this work. First, patients with PTSD or severe anxiety often exhibit alexithymia—the inability to identify and verbalize their own emotional state—making self-reported data unreliable [1]. Second, in forensic settings, subjects frequently engage in affective suppression, maintaining a neutral expression to avoid revealing distress [2]. Neither failure mode is detectable by observation alone, and studies report inter-observer agreement rates among trained clinicians as low as 60–70% [3].

Ekman's Facial Action Coding System established that while people can voluntarily control their facial expressions to some extent, vocal prosody—pitch, rhythm, tone, and timbre—is much harder to suppress consciously [4]. This creates a meaningful signal in the gap between what a subject's face shows and what their voice reveals. Prior work on datasets like IEMOCAP and CREMA-D confirms that acoustic features alone yield 65–75% accuracy for emotion detection, but combining them with visual features pushes that to 85%+ [5].

Existing systems largely address only one of these channels. Tools like OpenFace and DeepFace rely entirely on facial landmarks and fail under occlusion, poor lighting, or deliberate masking—all common in clinical and forensic environments [6]. They also produce raw probability scores rather than clinically interpretable reports, which limits their practical value to a practicing psychologist.

TraceMind was built to address these gaps. It combines a fine-tuned ResNet-50 visual model with a 1D Semi-CNN audio model, fusing their outputs per 3-second window across the full session and generating a structured PDF report that a clinician can actually use.

II. Related Work

A. Facial Expression Analysis

Facial Expression Recognition (FER) has been studied extensively in affective computing. Most systems are grounded in Ekman's Facial Action Coding System (FACS) [1], which maps facial muscle movements to discrete emotional states. Early methods used handcrafted geometric features and landmark detection [3], which worked well in controlled lab settings but degraded significantly with real-world variation in lighting and head pose.

Deep learning shifted the field toward CNNs, which learn directly from pixel data without manual feature engineering [5]. Modern CNN-based FER systems achieve strong benchmark numbers, but remain brittle to occlusion and pose variation [6]. More importantly for clinical use, purely visual

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

systems cannot detect emotions that a subject is actively suppressing—they only see the face, not the voice.

B. Speech and Audio Emotion Recognition

Speech Emotion Recognition (SER) extracts emotion from the acoustic properties of speech rather than its content. The key features are prosodic: pitch (fundamental frequency), energy, speech rate, and timbre, typically captured through Mel-Frequency Cepstral Coefficients (MFCCs) [4]. Models range from SVMs on handcrafted features to RNNs and CNNs operating directly on MFCC sequences [5].

Text-based emotion analysis has also been explored using NLP and transformer models like BERT [7]. However, text analysis alone misses tonal information entirely. The phrase "I'm fine" can signal either contentment or distress depending completely on how it is said—a distinction that text cannot resolve but audio can [5].

C. Multimodal Fusion and Remaining Gaps

Multimodal Emotion Recognition (MER) combines multiple channels to improve robustness, particularly in cases where one channel is unreliable. Empirical results consistently show that fusing visual and audio features outperforms either alone, especially in noisy or real-world conditions [5], [6].

Despite this, several practical gaps remain. Most commercial systems are still unimodal and cannot detect affective incongruence—the clinically significant case where a subject's face and voice send contradictory emotional signals [8]. Systems that do output multimodal results typically return raw probability vectors that are difficult for a non-technical clinician to interpret, with no temporal structure showing how emotion changed across a session [8]. Many proposed systems are also computationally heavy or require specialized hardware, making them unsuitable for routine clinical use [6].

TraceMind was designed specifically around these three gaps: detecting affective incongruence through fusion, generating interpretable clinical reports, and remaining lightweight enough to run on standard hardware.

III. Methodology

The TraceMind pipeline has three stages: independent feature extraction from the visual and audio streams, temporal segmentation into 3-second windows, and decision-level fusion.

A. Visual Modality: Facial Expression Recognition

The visual stream processes each video frame to extract a facial emotion probability vector. Face detection uses a Haar Cascade Classifier for speed. Detected face regions are cropped, converted from BGR to RGB, and resized to 224×224 pixels before being passed to the classification model.

We use ResNet-50 [9], trained in two stages. Initial training combined FER-2013 [10] (35,887 grayscale images) with CK+ [11] (593 video sequences from 123 subjects) to establish a general feature base and calibrate Action Unit detection. The model was then fine-tuned on RAF-DB [12] (29,672 real-world facial images) to handle the lighting variation and head pose shifts typical in clinical recordings. The final model classifies each detected face into seven categories: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise.

B. Audio Modality: Speech Emotion Recognition

The audio stream analyzes the paralinguistic properties of speech. Audio is extracted from the video, denoised using spectral gating, and segmented into 3-second chunks aligned with the visual windows. For each chunk, 40 MFCC features are extracted at a sample rate of 22,050 Hz.

A custom 1D Semi-Convolutional Neural Network processes the MFCC sequences. The model was trained on a composite corpus we refer to as the Level 4 Corpus, containing over 20,000 samples from five sources: RAVDESS [13], IEMOCAP [14], TESS [15], CREMA-D [16], and SAVEE [17]. These were combined to address two structural weaknesses common in individual SER datasets: limited sample count and gender imbalance. TESS provides predominantly female-speaker samples and SAVEE provides male-speaker samples, helping to distribute the training set more evenly across both.

C. Text Modality: Transcript Emotion Analysis

Audio from the session is transcribed using Google Speech Recognition and passed to a pretrained transformer model for emotion classification. We use a RoBERTa-based model fine-tuned on emotion corpora [7], deployed as a zero-shot classifier without additional fine-tuning on our data. The model returns a probability distribution over the same seven emotion classes. Because the full transcript is processed as a single unit rather than per window, the text modality contributes a global session-level signal rather than a frame-level one. This is a known limitation discussed in Section V.

D. Multimodal Fusion

All three streams produce a 7-class probability vector. Visual and audio vectors are combined per 3-second window using a weighted average at the decision level:

$$P_{final} = \alpha P_v + (1 - \alpha) P_a \tag{1}$$

where P_v is the visual probability vector, P_a is the audio probability vector, and $\alpha = 0.6$. The visual stream is weighted slightly higher because facial expressions carry stronger valence information in controlled conditions, while the audio stream acts as a corrective signal in cases of affective incongruence. The text modality applies an additional dynamic weight adjustment: if the transcript-level emotion contradicts the visual output, the visual weight is reduced to 0.15 and the audio and text weights are increased accordingly.

The dominant emotion for each window is the class with the highest fused probability. The full sequence of window-level labels is aggregated into a session-level report with a temporal emotion timeline.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

IV. Experimental Results and Analysis

A. Datasets and Setup

1) Visual Datasets: The visual model was trained in two stages. Initial training used FER-2013 (35,887 grayscale images) and CK+ (593 video sequences from 123 subjects) to establish a general feature base and refine Action Unit detection. The model was then fine-tuned on RAF-DB (29,672 real-world facial images) to improve performance under the lighting and pose conditions typical in clinical recordings.

2) Audio Datasets: The audio model was trained on a composite corpus we refer to as the Level 4 Corpus. RAVDESS, IEMOCAP, and CREMA-D contribute diverse emotional intensities and naturalistic speech patterns. TESS and SAVEE were specifically included to correct the gender imbalance present in most individual SER datasets.

B. Visual Model Performance

After initial training on FER-2013 and CK+, the ResNet-50 baseline accuracy was approximately 65%. Fine-tuning on RAF-DB improved this to **79%**, with the most significant gains in Fear and Surprise, which had been consistently confused at the baseline stage.

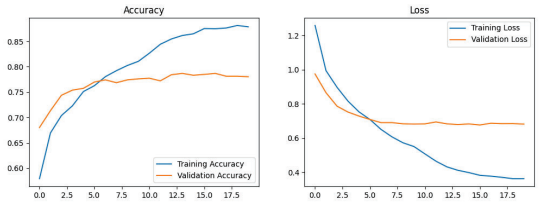

Fig. 1. Training and validation accuracy/loss for ResNet-50 (fine-tuned on RAF-DB).

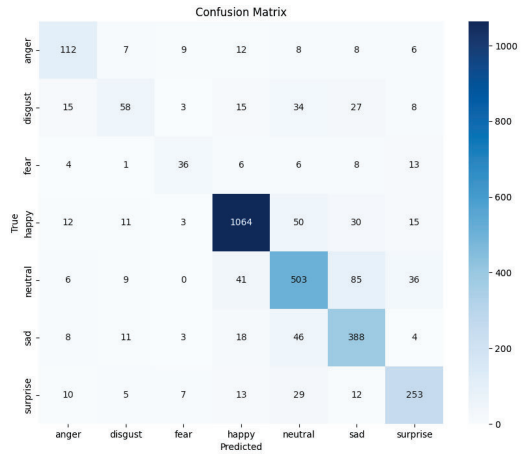

Fig. 2. Confusion matrix for the visual model evaluated on RAF-DB test set.

Table I shows the per-class performance. Happy achieves the highest F1-score (0.90), reflecting its large support in the dataset (1185 samples). Disgust is the weakest class (F1: 0.44), which we attribute to its visual similarity to Angry and Fear—a known challenge in FER literature that is partially corrected by the audio stream during fusion.

TABLE I
Performance Metrics: Visual Modality (ResNet-50)

Class	Precision	Recall	F1-Score	Support
Anger	0.67	0.69	0.68	162
Disgust	0.57	0.36	0.44	160
Fear	0.59	0.49	0.53	74
Happy	0.91	0.90	0.90	1185
Neutral	0.74	0.74	0.74	680
Sad	0.70	0.81	0.75	478
Surprise	0.76	0.77	0.76	329
Accuracy			**0.79**	**3068**
Macro Avg	0.70	0.68	0.69	3068
Weighted Avg	0.79	0.79	0.78	3068

C. Audio Model Performance

The 1D Semi-CNN achieved **80%** overall accuracy on the held-out test set. Class weighting during training helped address the imbalance in Fear and Disgust samples. Fear in particular performs well at the audio level—precision 0.94, F1-score 0.90—which is noteworthy given how poorly it performs in the visual stream (F1: 0.53). This asymmetry directly motivates the fusion approach: Fear is better detected through voice than face.

Fig. 3. Training and validation accuracy/loss for the 1D Semi-CNN.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

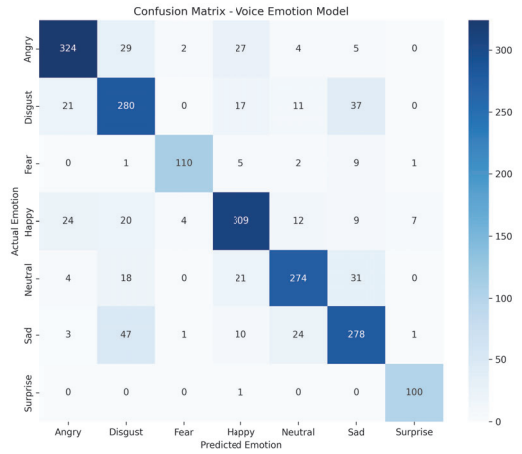

Fig. 4. Confusion matrix for the audio model on the Level 4 Corpus test set.

TABLE II
PERFORMANCE METRICS: AUDIO MODALITY (1D SEMI-CNN)

Class	Precision	Recall	F1-Score	Support
Angry	0.86	0.83	0.84	391
Disgust	0.71	0.77	0.74	366
Fear	0.94	0.86	0.90	128
Happy	0.79	0.80	0.80	385
Neutral	0.84	0.79	0.81	348
Sad	0.75	0.76	0.76	364
Surprise	0.92	0.99	0.95	101
Accuracy			**0.80**	**2083**
Macro Avg	0.83	0.83	0.83	2083
Weighted Avg	0.81	0.80	0.80	2083

D. Multimodal Fusion Analysis

The weighted average fusion ($\alpha = 0.6$) was evaluated qualitatively on video samples not used during training. The most informative test cases were those involving affective incongruence—where the face and voice disagreed.

In one representative case, a subject maintained a neutral facial expression throughout while their vocal tone indicated anger. The visual-only model classified the segment as Neutral (correct face read, wrong emotional conclusion). The fused model correctly returned Angry, demonstrating that the audio stream successfully overrode an incomplete visual signal.

This behavior is consistent with the per-class accuracy gap between modalities: Fear and Disgust are weakly detected visually but strongly detected by audio. The fusion weights were chosen to allow the audio stream enough influence to correct these cases without overriding the face when both channels agree.

E. System Interface and Clinical Deployment

TraceMind is deployed as a full-stack mobile application with a Flutter frontend and a Flask REST API backend hosted on Hugging Face Spaces. The interface is designed specifically for psychologists, not technical users—every screen reflects a clinical workflow rather than a developer tool. Figure 5 shows the patient management screen, where each patient is assigned a unique TraceMind ID (e.g., TM-2025-12-0001) alongside their diagnosed condition. This structured record system allows a clinician to track emotional profiles across multiple sessions over time.

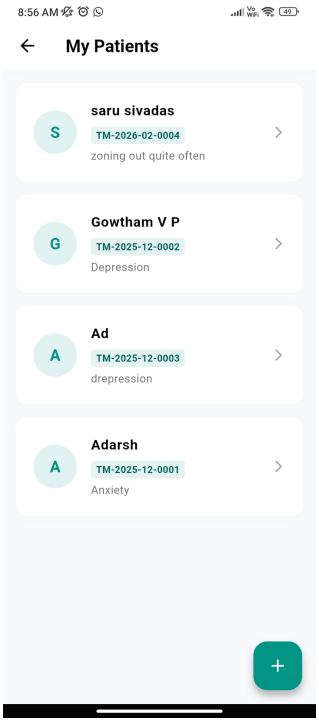

Fig. 5. Patient list screen showing TraceMind IDs and diagnosed conditions.

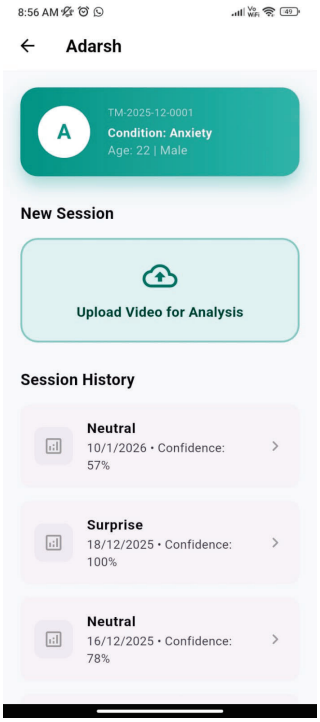

Fig. 6. Patient detail screen showing session history with dominant emotion and confidence per session.

Figure 6 shows the per-patient session history. Each completed analysis is recorded with a date, dominant emotion, and confidence score, allowing a clinician to observe how a patient's emotional state changes across sessions—a longitudinal view not available in single-point assessment tools.

The core clinical output is the session report, shown in Figures 7 and 8. The report has three components: an Executive Summary with the dominant emotion and a plain-language description, an Emotional Distribution bar chart showing the percentage breakdown across all seven emotion categories, and a Temporal Analysis table showing the per-window emotion and confidence score at 3-second intervals. This temporal breakdown is the primary differentiator from existing tools—it shows not just what the patient felt overall, but when during the session each emotional state occurred.

The dashboard (Figure 9) provides the clinician with a session-level overview: total patient count and reports ready, followed by a recent activity log showing all completed analyses with their dominant emotion and confidence. This gives a practicing psychologist an at-a-glance view of their caseload without navigating into individual patient records.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

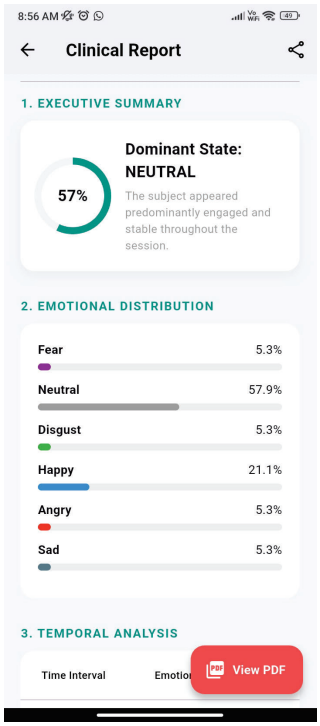

Fig. 7. Clinical report: Executive Summary and Emotional Distribution showing 57.9% Neutral dominant state.

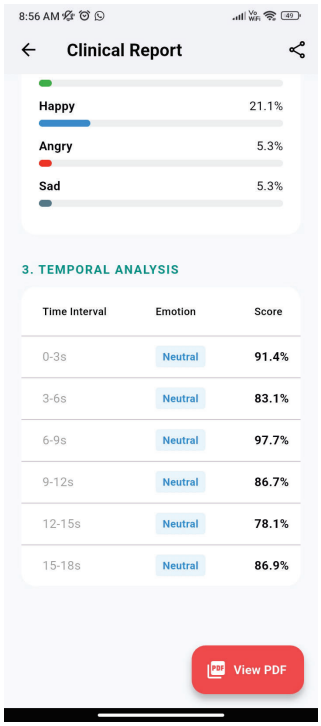

Fig. 8. Clinical report: Temporal Analysis showing per-window emotion labels and confidence scores at 3-second intervals.

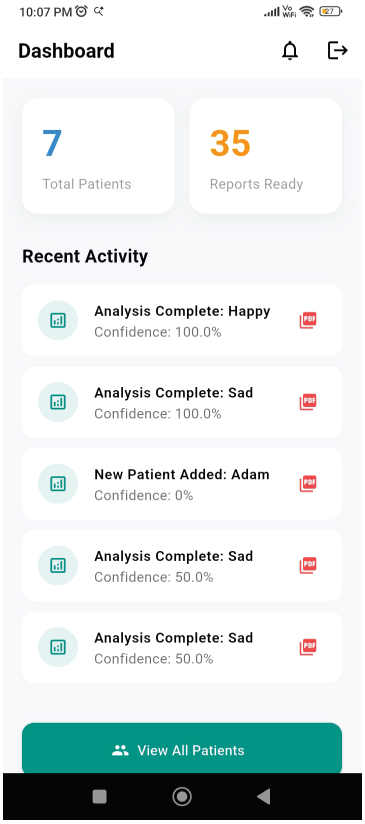

Fig. 9. Dashboard showing total patients, reports ready, and recent analysis activity with per-session dominant emotion.

V. Conclusion

TraceMind demonstrates that combining facial and audio emotion signals gives meaningfully better results than either alone, particularly for the emotion categories most relevant to clinical use—Fear and Anger. The visual ResNet-50 reached 79% accuracy after progressive fine-tuning on RAF-DB, and the audio 1D Semi-CNN reached 80% on the Level 4 Corpus. More importantly, the fusion step corrected cases that either model alone got wrong, including instances where a patient masked their facial expression while their vocal prosody remained revealing.

The clinical PDF output addresses a gap we observed in existing tools: most produce raw scores that are not interpretable without technical background. The temporal report generated by TraceMind gives a psychologist a readable session-level summary with per-window emotion labels.

There are clear limitations. The current text modality is not time-aligned to video windows—the full transcript is analyzed as a single unit. Incorporating word-level timestamps from a speech recognizer would allow text emotion to be fused per segment rather than globally. Real-time operation is also not yet supported; the current pipeline processes recordings offline. Both are targets for the next development phase.

Acknowledgment

The authors thank their faculty mentors and project guide for their feedback throughout the development of TraceMind. Thanks also to classmates who participated in early testing and code reviews. This work would not have been possible without the publicly available datasets (RAF-DB, RAVDESS, IEMOCAP, TESS, CREMA-D, SAVEE) and the open-source libraries on which TraceMind is built.

References

[1] R. Rahman, T. Islam, Md. H. Ahmed, "Detecting Emotion from Text and Emoticon," *International Journal of Computer Applications*, vol. 167, no. 9, 2017.
[2] B. R. Reddy, E. Mahender, "Speech to Text Conversion using Android Platform," *International Journal of Engineering Research and Applications*, vol. 3, no. 1, pp. 253–258, 2013.
[3] W. Wolf, "Key Frame Selection by Motion Analysis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 2411–2414, 1996.
[4] A. Trivedi, N. Pant, P. Shah, S. Sonik, S. Agrawal, "Speech to Text and Text to Speech Recognition Systems – A Review," *IOSR Journal of Computer Engineering*, vol. 20, no. 2, pp. 36–43, 2018.
[5] M. Rocamora, P. Herrera, "Comparing Audio Descriptors for Singing Voice Detection in Music Audio Files," in *Proc. Brazilian Symposium on Computer Music*, 2007.
[6] R. Olusegun, T. Oladunni, H. Audu, Y. Houkpati, S. Bengesi, "Text Mining and Emotion Classification on Monkeypox Twitter Dataset," *IEEE Access*, vol. 11, pp. 32–45, 2023.
[7] J. Devlin, M. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, pp. 4171–4186, 2019.
[8] S. Poria, E. Cambria, R. Bajpai, A. Hussain, "A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.
[9] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE CVPR*, pp. 770–778, 2016.
[10] I. J. Goodfellow et al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," in *Proc. ICONIP*, 2013.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

[11] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-Specified Expression," in *Proc. IEEE CVPR Workshops*, pp. 94–101, 2010.

[12] S. Li, W. Deng, J. Du, "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild," in *Proc. IEEE CVPR*, pp. 2584–2593, 2017.

[13] S. R. Livingstone, F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *PLOS ONE*, vol. 13, no. 5, p. e0196391, 2018.

[14] C. Busso et al., "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[15] K. Dupuis, M. K. Pichora-Fuller, "Toronto Emotional Speech Set (TESS)," *Scholars Portal Dataverse*, 2010.

[16] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, R. Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[17] S. Haq, P. J. B. Jackson, "Multimodal Emotion Recognition," in *Machine Audition: Principles, Algorithms and Systems*, W. Wang, Ed. IGI Global, pp. 398–423, 2010.