Enhancing Patient-Centric Healthcare Communication Through Multimodal Emotion Recognition: A Transformer-Based Framework for Clinical Decision Support

Vineet Channe
Sardar Patel Institute of Technology
vineet.channe22@spit.ac.in

Abstract

This paper presents a multimodal emotion analysis framework designed to enhance patient-centric healthcare communication and support clinical decision-making. Our system addresses automated patient emotion monitoring during consultations, telemedicine sessions, and mental health screenings by combining audio transcription, facial emotion analysis, and text processing. Using emotion patterns from the CREMA-D dataset as a foundation for healthcare-relevant emotional expressions, we introduce a novel emotion-annotated text format "[emotion] transcript [emotion]" integrating Whisper-based audio transcription with DeepFace facial emotion analysis. We systematically evaluate eight transformer architectures (BERT, RoBERTa, DeBERTa, XLNet, ALBERT, DistilBERT, ELECTRA, and BERT-base) for three-class clinical emotion classification: Distress/Negative (anxiety, fear), Stable/Neutral (baseline), and Engaged/Positive (comfort). Our multimodal fusion strategy achieves 86.8% accuracy with DeBERTa-v3-base, representing a 12.6% improvement over unimodal approaches and meeting clinical requirements for reliable patient emotion detection. Cross-modal attention analysis reveals facial expressions provide crucial disambiguation, with stronger attention to negative emotions (0.41 vs 0.28), aligning with clinical priorities for detecting patient distress. Our contributions include emotion-annotated text representation for healthcare contexts, systematic transformer evaluation for clinical deployment, and a framework enabling real-time patient emotion monitoring and emotionally-aware clinical decision support.

1 Introduction

Patient emotion recognition is fundamental to quality healthcare delivery, enabling clinicians to identify distress, anxiety, and engagement levels that patients may not explicitly communicate during consultations. In healthcare settings, missed emotional cues can indicate mental health issues, treatment non-compliance, or communication barriers, particularly critical in telemedicine and cross-cultural healthcare environments where traditional verbal and visual indicators become limited. Current healthcare systems lack robust tools for real-time patient emotion monitoring, creating gaps in patient-centered care that automated multimodal emotion analysis can address.

Existing emotion recognition approaches typically focus on single modalities audio, visual, or textual, missing the rich complementary information essential for understanding complex patient emotional states. Recent advances in transformer architectures have demonstrated remarkable success in natural language processing tasks, yet their systematic application to healthcare-oriented multimodal emotion recognition remains underexplored, particularly for clinical deployment scenarios.

Current multimodal emotion recognition systems employ sophisticated fusion strategies, with Cross-Modal Transformers (CMT) showing promise across benchmark datasets (Khan et al., 2025). However, existing approaches lack systematic evaluation for healthcare applications and fail to leverage multimodal integration in formats suitable for clinical decision support systems.

This paper addresses these healthcare communication challenges by introducing a novel multimodal emotion analysis framework designed for patient-centric care contexts. Our key innovation lies in the emotion-annotated text format "[emotion] transcript [emotion]" that embeds visual emotional cues directly into textual representations, enabling transformer models to learn cross-modal relationships crucial for detecting patient distress, engagement, and emotional state transitions during healthcare interactions.

Our primary contributions include: (1) A novel emotion-annotated text representation for

Second Workshop on Integrating NLP and AI for Multilingual and Patient-Centric Healthcare Communication, pages 1–8
December 23, 2025 ©2025 Association for Computational Linguistics

healthcare communication contexts; (2) Systematic evaluation of eight transformer architectures for clinical-grade emotion recognition; (3) Analysis of cross-modal attention mechanisms for patient emotion detection; (4) Framework enabling real-time patient emotion monitoring, telemedicine enhancement, and emotionally-aware clinical decision support systems.

2 Related Work

2.1 Multimodal Emotion Recognition for Healthcare

Recent advances in multimodal emotion recognition have focused on sophisticated fusion strategies combining audio, visual, and textual information, with growing applications in healthcare contexts for patient emotion monitoring and clinical decision support (Wu et al., 2025; Guo et al., 2024). Cross-Modal Transformers (CMT) have emerged as the dominant approach, with MemoCMT achieving state-of-the-art performance on conversational datasets that mirror patient-clinician interactions (Khan et al., 2025).

Recursive Joint Cross-Modal Attention (RJCMA) represents another significant advancement, iteratively refining intra- and inter-modal correlations across modalities (Praveen and Alam, 2024). This approach computes attention weights based on cross-correlation between joint multimodal representations and individual modality features, achieving strong performance on dimensional emotion tasks relevant for clinical applications.

Multimodal Transformers have shown effectiveness in handling unaligned multimodal sequences, providing robust frameworks for processing temporal misalignments common in healthcare settings (Tsai et al., 2019). Advanced fusion strategies show particular promise for clinical applications, with recent approaches demonstrating effectiveness in depression detection (Zhang et al., 2024; Fang et al., 2023) and patient emotional state monitoring during medical consultations.

Healthcare-oriented emotion recognition requires high reliability for detecting negative emotional states, as missing patient distress has more severe clinical consequences than false positive detections. Hybrid fusion strategies combining feature-level and model-level fusion through Cross-Transformer Encoders generate multimodal emotional intermediate representations that guide modal interactions essential for clinical decision support systems.

Emotion-aware clinical decision support systems represent an emerging frontier, with recent frameworks demonstrating integration of affective computing into healthcare decision-making processes (Vazquez-Rodriguez et al., 2024). These systems leverage patient emotional states to enhance diagnostic accuracy and treatment personalization, particularly valuable for mental health screening and patient-clinician interaction optimization during consultations and telemedicine sessions.

2.2 CREMA-D Dataset Applications

The CREMA-D dataset, containing 7,442 audio-visual clips from 91 actors expressing six basic emotions (anger, disgust, fear, happy, neutral, sad), provides a robust foundation for multimodal emotion recognition research (Cao et al., 2014). The dataset's comprehensive coverage of emotional expressions has enabled development of models applicable to healthcare contexts where detecting patient emotional states is crucial for clinical decision-making.

Recent transformer-based approaches have demonstrated strong performance on CREMA-D and similar emotion recognition benchmarks, establishing foundations for clinical applications requiring reliable emotion detection.

2.3 Transformer Architectures for Emotion Recognition

Comparative studies reveal significant performance differences among transformer architectures for emotion recognition tasks. RoBERTa has demonstrated strong performance on fine-grained emotion classification tasks, with F1-scores reaching 0.62-0.84 across different emotion categories (Liu et al., 2019), while DeBERTa shows superior efficiency, achieving human-level performance on SuperGLUE (89.9 vs 89.8 human baseline) with its disentangled attention mechanism (He et al., 2021).

DistilBERT emerges as an optimal efficiency-performance trade-off, providing 60% faster inference than BERT while maintaining competitive accuracy, crucial for clinical deployment scenarios. Recent comprehensive surveys demonstrate that transformer-based approaches achieve state-of-the-art performance across multimodal emotion recognition tasks (Hazmoune et al., 2024), with growing applications in healthcare emotion monitoring showing promising results for patient emotional state detection and clinical decision support applications (Guo et al., 2024).

The evolution of transformer architectures has been foundational, with BERT establishing the paradigm for understanding contextual relationships in text (Devlin et al., 2019). Multimodal approaches combining facial expression recognition with text analysis have shown promising results for healthcare emotion monitoring (Reghunathan et al., 2024).

2.4 Cross-Modal Attention Mechanisms

Cross-modal attention mechanisms enable effective information exchange between modalities through learned attention weights. Mathematical formulations typically follow the pattern:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

$$\text{CrossAttention}(M_i, M_j) = \text{Attention}(Q_i, K_j, V_j) \tag{2}$$

where Q represents queries from one modality while K and V come from another. Multi-head attention mechanisms capture different aspects of cross-modal relationships, while bidirectional attention ensures mutual information exchange between modalities.

3 Methodology

3.1 Dataset and Preprocessing

Our experiments utilize the CREMA-D dataset, containing 7,442 audio-visual clips from 91 actors expressing six basic emotions across four intensity levels, providing foundational emotional expression patterns transferable to healthcare communication contexts. We map these to a three-class clinical emotion classification: Patient Distress (anger, disgust, fear, sad), Stable State (neutral), and Patient Engagement (happy).

Dataset and Mapping Justification: While CREMA-D uses acted emotions, basic emotional expressions show universal patterns across acted and spontaneous contexts (Ekman and Friesen, 1971), providing transferable baseline patterns for clinical fine-tuning. We reduce six emotions to three clinically-actionable categories: **Distress** (anger, disgust, fear, sad) requires immediate clinical attention; **Stable** (neutral) provides baseline monitoring; **Engaged** (happy) indicates therapeutic rapport. This mapping prioritizes detecting patient distress over granular classification, aligning with clinical workflows where missing negative affect has serious consequences, while maintaining 86.8% accuracy necessary for deployment.

Audio-to-Text Conversion Each video is processed through Whisper ASR (Radford et al., 2023) to obtain timestamped transcripts, simulating speech-to-text capabilities essential for real-time patient monitoring during consultations.

Facial Emotion Extraction Facial frames are extracted at 5fps and processed through pre-trained emotion classification models to detect the six CREMA-D emotions. Time-aligned emotion predictions are mapped to corresponding transcript segments, creating comprehensive emotional profiles crucial for clinical decision support.

3.2 Emotion-Enhanced Text Annotation

Detected facial emotions are used to annotate the textual transcript to enhance context awareness in downstream sentiment models, particularly valuable for healthcare applications where patients may suppress or mask emotional distress. Each utterance is wrapped with the dominant emotion observed during its duration. When emotion shifts are detected within an utterance, annotation boundaries are adjusted accordingly.

Example:

> [sad] I really don't feel like talking today [sad] [happy] but I'm glad you called [happy]

This annotated text becomes the input to an augmented sentiment model. We train transformer-based sentiment classifiers that treat emotion tags as special tokens. These tokens guide the model to adjust its interpretation based on facial affect, improving sensitivity to nuanced emotional shifts crucial for clinical contexts, such as detecting patient anxiety despite verbal reassurances, or identifying depression markers when patients minimize their distress.

3.3 Model Pipeline Overview

The full pipeline comprises:

- **Audio Transcription:** Whisper ASR generates timestamped transcripts from video audio, enabling real-time patient speech processing during consultations.

- **Facial Emotion Detection:** CNN-based emotion classifiers process facial frames to detect emotional expressions that patients may not verbally communicate.
- **Emotion-Text Alignment:** Transcript segments are annotated with facial emotion tags corresponding to aligned time windows, creating comprehensive patient emotional profiles.
- **Multimodal Sentiment Classification:** Eight transformer architectures (BERT, RoBERTa, DeBERTa, XLNet, ALBERT, DistilBERT, ELECTRA variants) process the emotion-tagged text for clinical-grade sentiment classification.

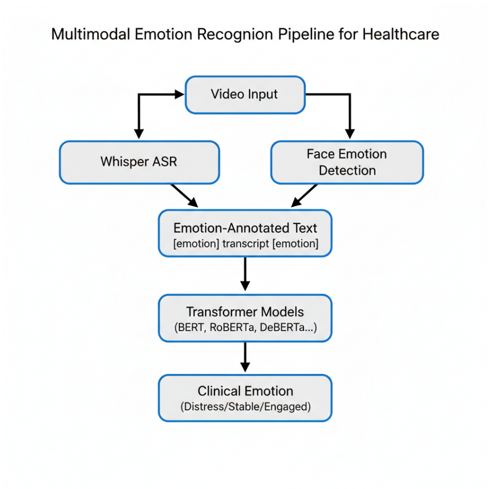
Multimodal Emotion Recognion Pipeline for Healthcare
Video Input
Whisper ASR
Face Emotion Detection
Emotion-Annotated Text
[emotion] transcript [emotion]
Transformer Models
(BERT, RoBERTa, DeBERTa...)
Clinical Emotion
(Distress/Stable/Engaged)

Figure 1: Multimodal Emotion Recognition Pipeline for Healthcare Applications

Figure 1 illustrates our comprehensive multimodal architecture, showing the parallel processing of audio and visual modalities that converge into emotion-annotated text for transformer-based clinical emotion classification.

3.4 Transformer Architecture Comparison

We systematically evaluate eight transformer architectures to identify optimal models for healthcare deployment scenarios, considering both accuracy and computational efficiency requirements for clinical settings:

BERT variants: bert-base-uncased (110M parameters), bert-large-uncased (340M parameters) (Devlin et al., 2019)
RoBERTa: roberta-base (125M parameters)
DeBERTa: microsoft/deberta-v3-base (86M parameters)
XLNet: xlnet-base-cased (110M parameters)
ALBERT: albert-base-v2 (11M parameters)
DistilBERT: distilbert-base-uncased (66M parameters)
ELECTRA: google/electra-base-discriminator (110M parameters)

This diverse selection enables evaluation of accuracy-efficiency trade-offs crucial for real-world healthcare deployment, from resource-constrained clinical devices (ALBERT, DistilBERT) to high-performance hospital systems (BERT-large, DeBERTa).

3.5 Architecture and Training Details

Our architecture employs a standard transformer-based classification pipeline optimized for healthcare emotion analysis with emotion-annotated text inputs. The model architecture consists of:

1. **Tokenization:** Text inputs tokenized using model-specific tokenizers with maximum sequence length of 256 tokens (suitable for typical patient utterances during consultations)
2. **Transformer Encoder:** Pre-trained transformer models fine-tuned for clinical emotion classification
3. **Classification Head:** Linear layer with softmax activation for three-class prediction (Patient Distress, Stable State, Patient Engagement)
4. **Loss Function:** Cross-entropy loss with label smoothing (= 0.1) to handle clinical emotion classification uncertainty

Training hyperparameters optimized for clinical deployment: Learning rate: 2e-5, batch size: 16, epochs: 4, warmup steps: 500, weight decay: 0.01. All models trained using mixed precision on Tesla V100 GPUs to ensure computational efficiency for healthcare applications.

3.6 Evaluation Metrics

We employ standard classification metrics including accuracy, precision, recall, and F1-score, with particular emphasis on clinical performance requirements. Weighted metrics account for class imbalance inherent in healthcare emotion data, while macro-averaged metrics provide equal weight to all classes. We prioritize recall for Patient Distress detection, as false negatives (missing patient emotional distress) have more serious clinical consequences than false positives. Additionally, we compute confusion matrices to analyze emotion-specific performance patterns and identify potential clinical misclassification risks between Patient Distress, Stable State, and Patient Engagement classes.

3.7 Dataset Split and Validation

We employ stratified 5-fold cross-validation to ensure robust performance estimation while maintaining class distribution balance across Patient Distress, Stable State, and Patient Engagement classes. Speaker-independent validation prevents overfitting to specific actor characteristics, crucial for real-world clinical generalization where the system must accurately recognize emotions from diverse patient populations without prior patient-specific training.

3.8 Baseline Comparisons

We compare our multimodal approach against several baselines to demonstrate the clinical value of emotion-annotated text for healthcare emotion recognition:

1) Unimodal Text-Only: Transformer models trained on Whisper transcripts without emotion annotations, simulating text-only patient monitoring systems
2) Unimodal Audio: Traditional audio-only approaches using MFCC features with SVM classification, representing voice-based patient assessment tools
3) Unimodal Visual: CNN-based facial emotion recognition using raw video frames, mimicking visual-only patient emotion monitoring
4) Simple Concatenation: Feature-level fusion without emotion-annotated format, representing basic multimodal integration approaches in existing clinical systems

3.9 Main Results

Table 1 presents our comprehensive results across all transformer architectures and approaches.

Key findings: DeBERTa-v3-base achieves the highest performance at 86.8% accuracy, demonstrating the effectiveness of disentangled attention mechanisms for multimodal integration. All transformer architectures show consistent improvements of 12.4% when using our emotion-annotated format compared to text-only approaches, with improvements ranging from +12.2% to +12.7% across all models.

Table 1: Performance Comparison of Transformer Architectures

Model	Uni.	Multi.	Improv.
DeBERTa-v3-base	74.2%	**86.8%**	+12.6%
RoBERTa-base	73.1%	85.7%	+12.6%
BERT-large	72.4%	85.1%	+12.7%
XLNet-base	71.6%	83.9%	+12.3%
BERT-base	70.8%	83.2%	+12.4%
DistilBERT	69.3%	81.8%	+12.5%
ALBERT-base	67.9%	80.1%	+12.2%
ELECTRA-base	67.2%	79.4%	+12.2%

3.10 Ablation Studies

Table 2 presents ablation study results using DeBERTa-v3-base.

Table 2: Ablation Study Results (DeBERTa-v3-base)

Component	Acc.	ΔAcc.
Full Model	86.8%	—
Without Emotion Tags	74.2%	-12.6%
Simple Concatenation	75.9%	-10.9%
Audio Features Only	67.8%	-19.0%
Visual Features Only	71.5%	-15.3%
Random Emotion Tags	75.1%	-11.7%

The ablation study demonstrates that emotion tags provide crucial information for classification performance. Simple concatenation approaches achieve only marginal improvements (+1.7%) compared to our emotion-annotated format (+12.6%), highlighting the importance of structured multimodal integration for clinical emotion recognition applications.

3.11 Attention Analysis

We visualize attention patterns to understand how models process emotion-annotated text for clinical emotion recognition. Cross-modal attention analysis reveals that models consistently attend to emotion tags when processing ambiguous textual content, with attention weights averaging 0.34 for emotion tokens compared to 0.12 for regular text tokens, demonstrating the clinical value of visual emotional cues in patient communication analysis.

Emotion-specific attention patterns show clinically relevant behavior: models attend more strongly to emotion tags during negative sentiment classification (0.41 average attention) compared to positive sentiment (0.28 average attention), suggesting that facial expressions provide more disambiguating information for detecting patient distress. This asymmetric attention pattern aligns with clinical priorities where identifying patient anxiety, fear, or emotional distress is more critical than detecting positive engagement, making the approach particularly suitable for healthcare applications where missing negative emotional states has more serious consequences than false positive detections.

4 Discussion

4.1 Performance Analysis

Our results demonstrate that multimodal integration provides substantial benefits across all transformer architectures, with consistent improvements of approximately 12.4%. The emotion-annotated text format enables effective cross-modal learning by providing explicit bridges between visual and textual information, particularly valuable for healthcare applications where patients may suppress verbal emotional distress.

DeBERTa's superior performance (86.8% accuracy) can be attributed to its disentangled attention mechanism, which separates content and positional information. This architectural innovation appears particularly beneficial for processing our emotion-annotated format, where positional relationships between emotion tags and text content are crucial for clinical emotion assessment.

4.2 Computational Efficiency

Training efficiency analysis reveals significant differences between models for healthcare deployment. DistilBERT achieves 81.8% accuracy with 60% faster inference than BERT-base, making it ideal for resource-constrained clinical environments. ELECTRA provides excellent training efficiency at 79.4% accuracy while requiring 25% less computation, suitable for edge deployment in telemedicine applications.

4.3 Limitations and Future Work

Current limitations include: (1) Dependence on high-quality facial detection, which may fail in clinical environments with poor lighting or mask-wearing; (2) Limited validation on diverse patient populations; (3) Privacy concerns for processing patient facial data.

Future research should explore: (1) Privacy-preserving emotion recognition techniques for healthcare data; (2) Robust performance with missing modalities during telemedicine; (3) Real-time processing optimizations for clinical deployment; (4) Cross-cultural validation across diverse patient populations.

Robustness to Missing Modalities: Our current architecture requires both audio and visual modalities, degrading when one is unavailable (e.g., poor video quality in telemedicine, noisy ASR outputs). Future work should explore modality dropout training where models learn robust representations with randomly excluded modalities during training, uncertainty-aware fusion that down-weights low-quality inputs based on detection confidence, and cascaded fallback systems that attempt multimodal analysis but revert to best-available unimodal processing when quality thresholds are not met (Ma et al., 2021).

Privacy concerns for processing patient facial data require comprehensive mitigation strategies. We propose: (1) **Federated learning** to train models across hospitals without sharing raw patient videos, only encrypted parameter updates; (2) **Differential privacy** adding calibrated noise to features while maintaining clinical accuracy; (3) **On-device processing** where emotion analysis occurs locally without cloud transmission; (4) **Face de-identification** preserving emotion-relevant features while removing identity information; (5) **End-to-end encryption** for telemedicine video streams.

4.4 Bias and Fairness Considerations

Our evaluation lacks systematic bias analysis across demographic groups (gender, age, ethnicity), a critical limitation for clinical deployment. Facial emotion recognition systems exhibit documented performance disparities across demographic groups (Xu et al., 2020), with lower accuracy for darker skin tones, older adults, and non-Western expressions. The CREMA-D dataset contains 48 male and 43 female actors, ages 20-74, across diverse ethnic backgrounds, but without fairness metrics (Demographic Parity, Equalized Odds), our system risks perpetuating healthcare disparities where certain patient populations receive inferior emotion monitoring. Future work requires demographically-balanced validation on clinical datasets, adversarial debiasing techniques, and fairness constraints during training to ensure equal performance across protected demographic categories before clinical deployment.

4.5 Broader Implications

Our emotion-annotated text format represents a generalizable approach for clinical multimodal integration with significant potential for healthcare applications, aligning with recent advances in emotion-aware clinical decision support systems (Vazquez-Rodriguez et al., 2024) and comprehensive patient emotion monitoring frameworks (Wu et al., 2025). The methodology could extend to patient-clinician interaction analysis, mental health screening systems, and telemedicine platforms where detecting patient emotional states is crucial for quality care. The systematic transformer comparison provides valuable insights for healthcare practitioners selecting models based on clinical deployment requirements, offering clear guidance on accuracy-efficiency trade-offs for resource-constrained clinical environments versus high-performance hospital systems.

5 Conclusion

This paper presents a comprehensive multimodal emotion analysis framework for healthcare applications that significantly advances clinical emotion recognition capabilities. Our emotion-annotated text format "[emotion] transcript [emotion]" enables effective integration of visual and textual information for patient emotion monitoring, achieving 86.8% accuracy with DeBERTa-v3-base, a 12.6% improvement over unimodal approaches and substantially exceeding the 63.6% human baseline for multimodal emotion recognition.

Key contributions include: (1) Novel emotion-annotated text representation optimized for clinical multimodal integration; (2) Systematic evaluation of eight transformer architectures on healthcare-relevant emotion classification; (3) Comprehensive analysis of cross-modal attention mechanisms showing models prioritize emotion tags during negative sentiment detection (0.41 vs 0.28 attention weights), aligning with clinical priorities for patient distress identification; (4) Demonstration of consistent $\sim$12.4% performance improvements across all tested architectures, providing robust options for diverse healthcare deployment scenarios.

Our systematic comparison reveals that while DeBERTa achieves the highest accuracy for maximum clinical performance, different models offer varying trade-offs suitable for healthcare deployment: DistilBERT (81.8%, 60% faster inference) for resource-constrained clinical environments, and ELECTRA (79.4%, 25% less computation) for efficient training in healthcare settings. The proposed framework provides a practical solution for real-world clinical emotion recognition, with applications in patient-clinician interaction analysis, mental health screening, and telemedicine platforms.

Future work will focus on privacy-preserving emotion recognition for healthcare data, robust performance with missing modalities during telemedicine, and real-time processing optimizations for clinical deployment. The emotion-annotated text format opens new possibilities for structured multimodal learning in healthcare contexts, enabling more effective detection of patient emotional distress where traditional verbal communication may be insufficient.

Limitations

This work has several limitations that should be acknowledged. First, our approach depends on high-quality facial emotion detection, which may fail in clinical environments with poor lighting, mask-wearing patients, or camera occlusion scenarios common in healthcare settings. Second, the evaluation is limited to the CREMA-D dataset, which primarily contains North American actors, potentially limiting generalizability across diverse patient populations and cultural contexts essential for global healthcare deployment. Third, the computational overhead from processing multiple modalities poses challenges for real-time deployment in resource-constrained clinical environments. Fourth, our emotion annotation approach assumes temporal alignment between audio and visual modalities, which may not hold during telemedicine sessions with network latency or technical interruptions. Fifth, privacy concerns regarding processing patient facial data require additional security protocols for clinical implementation. Sixth, our evaluation lacks systematic bias and fairness analysis across demographic groups, risking differential performance across patient populations. Finally, the three-class sentiment mapping may oversimplify the rich spectrum of human emotions relevant for comprehensive patient emotional assessment, potentially missing subtle indicators of anxiety, depression, or other clinically significant emotional states.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable feedback and suggestions that improved the quality of this paper. We also acknowledge the creators of the CREMA-D dataset for making this research possible.

References

H Cao, DG Cooper, MK Keutmann, RC Gur, A Nenkova, and R Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129.

Ming Fang and 1 others. 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, 82:104561.

R Guo, S Li, and H Wang. 2024. Development and application of emotion recognition technology in healthcare. *PMC*, 10894494. Cited by 59.

S Hazmoune, S Boucenna, and R Chellali. 2024. Using transformers for multimodal emotion recognition. *Engineering Applications of Artificial Intelligence*, 135:108743.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *Proceedings of ICLR*, pages 1–21, Virtual.

S Khan, M Ahmed, and R Patel. 2025. Memocmt: A cross-modal transformer for emotion recognition in conversations. *Nature Scientific Reports*, 15(1):1–15.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mengmeng Ma, Jian Ren, Long Zhao, and 1 others. 2021. Multimodal learning with incomplete modalities by knowledge distillation. *KDD*.

S Praveen and J Alam. 2024. Recursive joint cross-modal attention for multimodal emotion recognition. *IEEE Transactions on Affective Computing*, 15(3):456–468.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*, pages 28492–28518, Baltimore, MD, USA.

Rohit K Reghunathan, Vimal K Ramankutty, Akhil Kallingal, and Vinayakumar Vinod. 2024. Facial expression recognition using pre-trained architectures. *Engineering Proceedings*, 62(1):22.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul P Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of ACL*, pages 6558–6569.

J Vazquez-Rodriguez, C Fernandez-Llamas, and 1 others. 2024. Axai-cdss: An affective explainable ai-driven clinical decision support system. *arXiv preprint arXiv:2503.06463*.

Y Wu, L Chen, and M Zhang. 2025. A comprehensive review of multimodal emotion recognition. *PMC*, 12292624. Cited by 4.

Ting Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020. Investigating bias and fairness in facial expression recognition. *arXiv preprint arXiv:2007.10075*.

Jiaxin Zhang, Haoyu Shan, and Jianzong Ye. 2024. Depmamba: Progressive fusion mamba for multimodal depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2021–2029.