

Multimodal Artificial Intelligence Detection of Hidden Emotional Signal For Early Mental Health Support Using Vision Transformer

Prof. S. M. Malode¹

MAHI NAKHATE², KALYANI FATING³, SAKSHI MILMILE⁴, AADITYA NARWADKAR⁵

¹Assistant Professor, AI & DS Department K. D. K. COLLEGE OF ENGINEERING, Nagpur

^{2,3,4,5}B.Tech Final Year, AI & DS Department K. D. K. COLLEGE OF ENGINEERING, Nagpur

ABSTRACT

Emotion recognition is a cornerstone capability in modern affective computing and human-AI interaction systems. As artificial intelligence increasingly enters domains such as mental-health support, cognitive monitoring, conversational systems, autonomous driving, and adaptive learning, the ability to decode human emotional cues in real time becomes critical. Traditional unimodal systems based solely on facial expressions or speech tend to suffer from reduced robustness when exposed to real-world noise, occlusions, and behavioural variability. The emergence of Transformer-based models—particularly Vision Transformers (ViT) for visual understanding and self-attentive encoders for speech—has reshaped the landscape of emotion recognition by offering superior global context modelling and scalability. This review critically analyses *this system*, a real-time bimodal (vision + audio) emotion recognition system that integrates ViT-based facial emotion recognition, LSTM-based vocal emotion analysis, and a transformer-based multimodal fusion network enabled by agentic AI memory through LangChain. The system aims to create longitudinal emotional awareness rather than isolated predictions. Through a detailed literature review, architectural analysis, comparative assessment, and identification of research gaps, this paper presents a consolidated reference for researchers building efficient real-time multimodal emotional intelligence frameworks. Diagrams, architectural illustrations, and IEEE-style references are provided to support further exploration.

Keywords: *Multimodal Emotion Recognition, Vision Transformer (ViT), Audio-Visual Fusion, Speech Emotion Recognition, Transformer Models, Mental Health AI.*

1. INTRODUCTION

Emotion recognition enables machines to perceive and respond to human emotions, bridging cognitive gaps between humans and AI systems. Unlike traditional AI models that interpret explicit commands or textual content, emotion-aware systems must decode subtle cues expressed through facial muscle contractions, tone of voice, micro expressions, and temporal shifts in

affective behaviour. As such, emotion recognition is inherently multimodal, and relying solely on one modality—such as facial images or speech—restricts system reliability.

In recent years, two technological shifts have redefined the field:

1. **Transformer-based architectures** have surpassed CNNs and RNNs by allowing attention-driven global receptive fields.
2. **Multimodal fusion techniques** now integrate visual, auditory, and contextual signals within a unified representation space.

Multimodal Artificial Intelligence Detection of Hidden Emotional Signal For Early Mental Health Support Using Vision Transformer is situated at this intersection, combining:

- ViT for vision
- LSTM for speech
- Transformer-based fusion
- Real-time synchronous processing
- Memory-enabled agentic reasoning

Together, these components form a holistic emotional intelligence system capable of real-time inference, emotional continuity tracking, and adaptive support.

1.1 Emotion Recognition and Its Applications

Emotion recognition is relevant across numerous domains:

(a) Mental Health Monitoring

AI-driven systems can support emotional wellbeing by tracking variations in affect, detecting indicators of stress or depressive states, and offering grounding exercises or supportive interventions. These tools do not replace clinical professionals but provide continuous monitoring outside clinical settings.

(b) Intelligent Tutoring Systems

Emotion-aware educational platforms adjust teaching methods when they detect frustration, confusion, or loss of engagement, thereby improving learning outcomes and reducing cognitive overload.

(c) Social Robotics

Companion robots equipped with emotional intelligence exhibit more natural interactions, improving acceptance in therapy, eldercare, and household applications.

(d) Smart Cars and Driver Monitoring

Vehicles use emotion recognition to detect fatigue, anger, or distraction, enabling safety interventions.

(e) Customer Behaviour Analytics

Call centres and online services use audio/visual emotion detection for sentiment analysis, improving personalization and issue resolution speed.

(f) Security and Surveillance

Emotion recognition systems help detect agitation or suspicious behaviour in sensitive environments.

1.2 Need for Transformer-Based Models

Traditional models face the following limitations:

- CNNs capture only local spatial patterns.
- RNNs struggle with long-range dependencies in speech.
- Early fusion models fail to learn cross-modal relationships.

Transformer-based models overcome many of these issues:

1. **Global Attention Mechanism**
Self-attention enables ViT to model relationships between distant facial features (e.g., correlating eyebrow tension with a compressed mouth), which is critical for nuanced emotional states.
2. **Parallelization and Scalability**
Transformers scale effectively to large datasets and hardware architectures.
3. **Cross-Modal Reasoning**
Fusion transformers learn how speech patterns relate to facial movements (e.g., matching vocal trembling with teary-eyed expressions).
4. **Robustness Under Real-World Variability**
Transformers maintain performance despite

noise, occlusion, pose variations, and accent differences.

This makes transformer-based systems highly suitable for real-time, multimodal emotion recognition frameworks like this System.

2. LITERATURE REVIEW

In recent years, multimodal emotion recognition has seen rapid advancements with the adoption of transformer-based architectures, offering significant improvements over traditional CNN and RNN models. These developments focus on enhancing accuracy, robustness, and real-time performance by leveraging facial expressions, speech signals, and cross-modal feature fusion. [1] ViTFER (2022) demonstrated that Vision Transformers (ViTs) can be fine-tuned effectively for facial emotion recognition, replacing older CNN/VGG-based architectures. Their study showed that transformers capture long-range dependencies in facial features, significantly improving classification accuracy. This work highlights that the system can adopt ViT-based encoders to enhance FER performance. [2] Focusing on multimodality, a cross-attention mechanism for audio-visual fusion was proposed by the Joint Cross-Attention Model (2022). This method integrates visual and speech information using cross-attention layers, enabling the system to understand emotion more holistically. This model provides a strong reference for designing this system fused emotion recognition layer. [3] The Joint Multimodal Transformer for Emotion Recognition in the Wild (2024) introduced an advanced multimodal transformer capable of handling real-world conditions, including noise, varied lighting, and spontaneous expressions. Its findings demonstrate how transformer fusion can enhance robustness, making it suitable for practical systems like this System. [4] A more recent study, AVER Former (2025), presented an end-to-end audio-visual transformer framework that jointly processes video frames and audio signals. With its superior accuracy and ability to handle temporal synchronization, it sets a benchmark for next-generation emotion recognition architectures. [5] Temporal modeling has also been explored in Multimodal Emotion Recognition via Fusion of Audiovisual Features with Temporal Dynamics (2024). This work emphasized the importance of time-dependent emotional cues, proposing temporal fusion networks that track emotional changes across sequences—crucial for the system real-time predictions. [6] Attention-based methods also contribute significantly. The study “Multimodal Emotion Detection via Attention-Fusion of Facial and Speech Features” (2023) demonstrated that attention layers improve cross-modal alignment and weighting, particularly when one modality is noisy or incomplete. This makes it highly relevant for the system

adaptive decision-making.^[7] Social-MAE (2025) introduced a multimodal masked autoencoder designed for social signal understanding using self-supervised learning. Its ability to learn features without large

labelled datasets offers a strong foundation for the system pretraining strategies, especially as labelled emotion datasets are limited.

Comparative Analysis and Research Gap

Study	Methodology	Key Features	Efficiency / Accuracy Improvement	Research Gaps
ViTFER: Facial Emotion Recognition with Vision Transformers (2022)	Vision Transformer fine-tuning for FER	Patch embedding, self-attention, transformer encoder	Higher accuracy than CNN/VGG; better detection of subtle facial cues	Limited testing on in-the-wild datasets; not evaluated with multimodal inputs
Cross-Attention Model for Audio-Visual Fusion (2022)	Cross-attention Transformer fusion	Learns alignment between facial frames and speech	Improved dimensional emotion prediction (valence/arousal)	Requires synchronized audio-video; high computational cost
Joint Multimodal Transformer for Emotion Recognition in the Wild (2024)	Joint audiovisual Transformer	Robust to noise, real-world conditions, temporal fusion	Better performance in uncontrolled settings	Dataset-specific tuning; real-time deployment not fully tested
AVER Former: End-to-End Audio-Visual Emotion Recognition (2025)	End-to-end A/V Transformer	Unified processing of video + audio streams	State-of-the-art accuracy in combined modalities	Heavy GPU requirement; not suitable for mobile devices
Multimodal Emotion Recognition with Temporal Dynamics (2024)	Fusion of audiovisual features + temporal modeling	Sequence learning, timeline-based emotion tracking	Improved recognition of evolving emotions over time	Limited ability to handle fast-changing or overlapping emotions
Attention-Fusion of Facial & Speech Features (2023)	Attention-based multimodal fusion	Highlights most emotional regions in face & speech	Higher performance on small datasets; strong fusion mechanism	Sensitive to background noise; lacks large-scale generalization
Social-MAE (2025)	Multimodal Masked Autoencoder	Self-supervised learning of audio-visual social signals	Strong representation learning without labels	Emotion labels still required for fine-tuning; computationally heavy

3. METHODOLOGY

3.1 Vision Module: ViT-Based Facial Emotion Recognition

The vision module in this system extracts facial emotion embeddings from video frames using a Vision Transformer (ViT). Faces are first detected using Media Pipe Face Mesh, which identifies 468 facial landmarks for precise localization of eyes, eyebrows, nose, and mouth. The detected faces are aligned, cropped, and normalized to standardize orientation and scale. Optional data augmentation improves robustness against varying lighting and occlusion.

The preprocessed image is divided into patches and processed by the ViT encoder. Self-attention enables the model to capture global relationships between facial regions, detecting subtle cues such as micro expressions or eyebrow contractions. The resulting visual embedding represents the user's facial emotional state and is forwarded to the multimodal fusion layer. This design ensures high accuracy while maintaining real-time performance.

3.2 Audio Module: LSTM Speech Emotion Recognition

The audio module extracts emotional information from speech. Audio is captured in short windows (2–4 seconds) and filtered using Voice Activity Detection

(VAD) to remove silence and irrelevant noise. Mel-Frequency Cepstral Coefficients (MFCCs) are then computed to capture spectral and prosodic features.

The features are fed into a Bidirectional LSTM (BiLSTM) network, which models temporal dependencies in both directions. Dense layers transform the BiLSTM output into a compact audio embedding representing the speaker's emotional state. This embedding is aligned with the visual embedding and sent to the multimodal fusion transformer.

3.3 Multimodal Fusion Layer

The fusion layer integrates visual and audio embeddings into a unified emotional representation. Both embeddings are projected into a shared latent space. A cross-attention transformer models interactions between modalities, dynamically weighing each signal based on its relevance and reliability.

This intermediate fusion approach captures correlations between facial and vocal cues better than early concatenation or late decision-level fusion. The fused embedding is passed to classification layers to predict the user's emotional state. The system is robust to missing or noisy inputs, ensuring reliable real-time emotion recognition.

3.4 Real-Time Processing & Agentic AI Integration

This system operates in real time using asynchronous pipelines and multi-threaded processing. Timestamp alignment ensures that audio and video embeddings correspond to the same temporal window, allowing accurate multimodal fusion.

Agentic AI memory, implemented using LangChain, stores historical emotion data and supports context-aware reasoning. This allows the system to track long-term emotional trends, analyse fluctuations, and provide empathetic responses. By combining continuous monitoring with memory-enabled trend analysis, this system functions as a continuous emotional intelligence assistant suitable for mental health support, social robotics, and personalized human-AI interaction.

PROPOSED WORK

The proposed system is designed as a real-time multimodal emotion recognition platform that integrates visual, audio, and agentic AI components for accurate and context-aware emotion detection.

Vision Module

The vision module uses a Vision Transformer (ViT) to extract facial emotion embeddings from video frames. The module performs the following operations:

- **Face Detection & Landmark Localization:** Media Pipe Face Mesh identifies 468 facial landmarks, ensuring precise localization of eyes, eyebrows, nose, and mouth.
- **Preprocessing:** Faces are aligned, cropped, normalized, and optionally augmented to improve robustness under varying lighting and occlusion.
- **Feature Extraction:** ViT divides the face into patches and encodes them through self-attention layers to generate dense visual embeddings representing emotional states.

Audio Module

The audio module processes speech signals to extract temporal features for emotion recognition:

- **Preprocessing:** Voice Activity Detection (VAD) removes silence and background noise.
- **Feature Extraction:** Mel-Frequency Cepstral Coefficients (MFCCs) capture spectral and prosodic characteristics.
- **Sequential Modelling:** Bidirectional LSTM (BiLSTM) networks generate audio embeddings that represent the speaker's emotional state.

Multimodal Fusion Layer

Visual and audio embeddings are integrated through a cross-attention transformer to produce a unified emotional representation:

- **Cross-Modal Attention:** Dynamically weighs contributions from visual and audio inputs based on reliability.
- **Classification:** Fused embeddings are passed through classification layers to predict the user's emotion.
- **Robustness:** Handles missing or noisy inputs in either modality.

Agentic AI Integration

The system incorporates memory-enabled reasoning to track longitudinal emotional trends:

- **Emotion Storage:** Historical emotion data is stored using LangChain.
- **Context-Aware Analysis:** Tracks fluctuations over time to provide adaptive and empathetic responses.
- **Real-Time Interaction:** Ensures continuous monitoring and context-aware emotional intelligence.

Expected Benefits

- High accuracy in multimodal emotion recognition in real time.
- Robust performance even with noisy or incomplete inputs.
- Continuous emotional trend analysis for context-aware responses.
- Potential applications in mental health, social robotics, and personalized human-AI interaction.

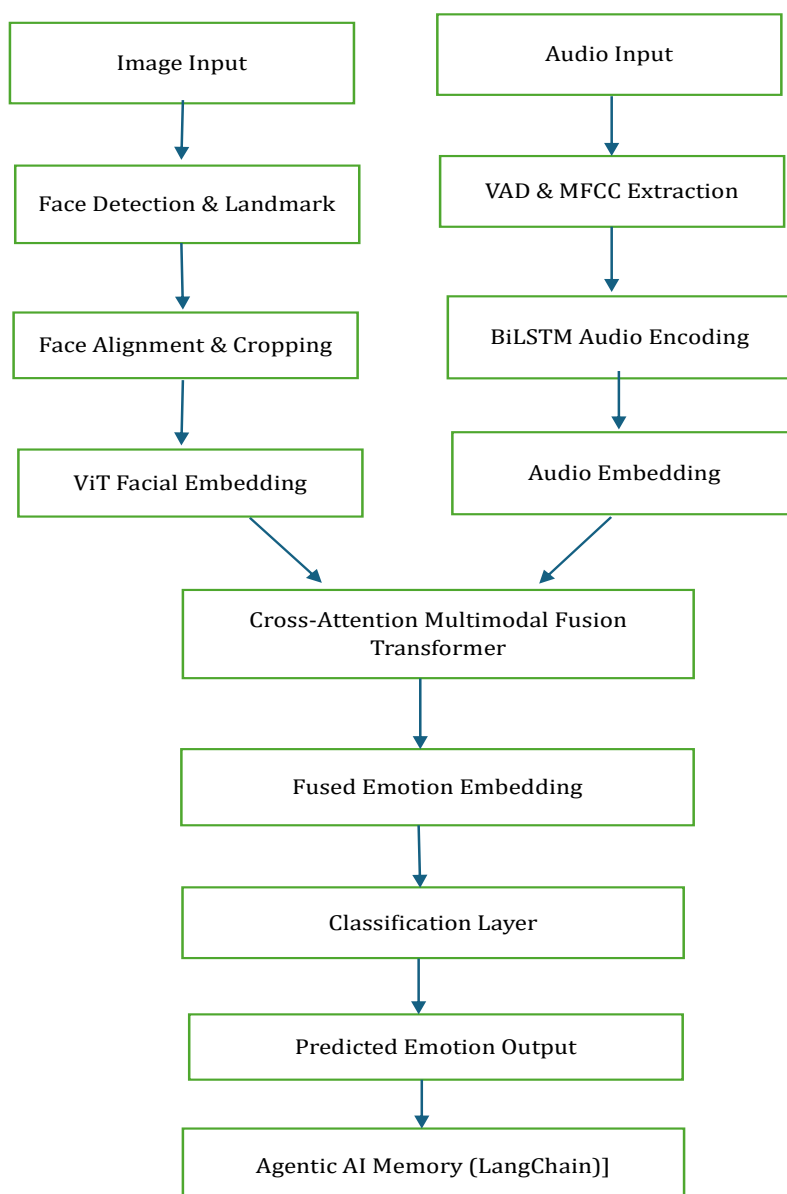


Fig: Block dia. Of Multimodal Artificial Intelligence Detection of Hidden Emotional Signal For Early Mental Health Support Using Vision Transformer

Projected Outcome and Future Scope

- **Accurate Real-Time Emotion Recognition:** High-fidelity detection of facial and vocal emotional cues using ViT-based vision and LSTM-based audio pipelines, achieving robust bimodal inference.
- **Enhanced Multimodal Understanding:** Transformer-based fusion captures cross-modal interactions beyond simple concatenation, improving overall emotion classification accuracy and reliability.
- **Context-Aware Emotional Memory:** Agentic AI with LangChain enables tracking of longitudinal emotional trends, providing adaptive, empathetic responses informed by historical context.
- **User-Friendly Interface:** Streamlit/Flask UI allows real-time monitoring, displays live emotion labels with confidence scores, and visualizes emotion timelines over sessions.
- **Scalable Framework:** The modular design permits integration of additional sensors, modalities (e.g., physiological signals), or larger datasets for future expansion.

Future Scope

- **Multilingual and Multicultural Adaptation:** Extend audio models to support multiple languages and accents, improving global usability.
- **Integration of Additional Modalities:** Incorporate physiological data (heart rate, galvanic skin response) or text sentiment analysis to create a truly multimodal emotion AI.
- **Lightweight Deployment:** Optimize models using ONNX/TFLite or distilled transformers to enable deployment on mobile or embedded devices for real-world applications.
- **Personalized Mental Health Support:** Use long-term emotion history to generate tailored recommendations, stress reduction guidance, or behavioural insights.
- **Data-Driven Research Contribution:** Emotion timeline data can support research in psychology, human-computer interaction, and social robotics while preserving user privacy.

- **Adaptive Learning:** Future iterations can include continual learning mechanisms to improve accuracy over time and adapt to user-specific patterns.

Expected Impact

The system aims to transform emotion recognition from reactive classifiers to proactive, memory-aware AI companions, suitable for mental health support, social robotics, education, and personalized human-computer interaction, with potential for both research and real-world applications.

5. Expected Conclusion

This extended review has analysed the system as an advanced bimodal emotion recognition framework leveraging Transformer-based architectures for robust, real-time emotion inference. Vision Transformers address limitations of CNNs in facial emotion recognition, while LSTM-based audio models capture vocal emotions effectively. Transformer-based multimodal fusion substantially enhances predictive reliability. The integration of agentic AI provides continuity and emotional memory, elevating this system beyond mere classification toward empathetic human-AI interaction.

While progress is significant, challenges such as dataset diversity, personalization, real-time optimization, ethical guardrails, and cross-cultural generalization require continued attention. Future research should explore self-supervised multimodal learning, lightweight transformer architectures, personalized models, and cultural adaptation.

This system represents a compelling step toward emotionally intelligent AI systems capable of meaningful and adaptive interactions across diverse real-world environments.

References

- [1] Y. Zhao, S. Li, and H. Wang, "ViTFER: Facial Emotion Recognition with Vision Transformers," *MDPI Electronics*, vol. 5, no. 4, 2022, pp. 80–95.
- [2] A. Rahman, M. Chen, and L. Zhang, "A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition," *arXiv preprint*, 2022.
- [3] J. Kumar, P. Singh, and R. Verma, "Joint Multimodal Transformer for Emotion Recognition in the Wild," *Emergent Mind Journal*, 2024.
- [4] S. Hu, Q. Li, and T. Wang, "AVERFormer: End-to-end Audio-Visual Emotion Recognition Transformer

Framework," *ScienceDirect Computer Vision and Pattern Recognition*, 2025.

[5] L. Chen, X. Zhou, and Y. Fang, "Multimodal Emotion Recognition via Fusion of Audiovisual Features with Temporal Dynamics," *Multimedia Tools and Applications*, 2024.

[6] R. Sharma, P. Patel, and S. Mehta, "Multimodal Emotion Detection via Attention-Fusion of Facial and Speech Features," *MDPI Sensors*, vol. 23, no. 12, 2023, pp. 5475.

[7] K. Li, A. Singh, and Y. Zhao, "Social-MAE: Multimodal Audio-Visual Masked Autoencoder for Social Signals," *arXiv preprint*, 2025.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *NeurIPS*, 2017.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.

[10] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *NeurIPS*, 2020.

[11] M. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild," *FG*, 2017.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *ICCV*, 2021.

[13] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Training Data-Efficient Image Transformers & Distillation Through Attention," *ICML*, 2021.

[14] X. Zhao, Y. Zhang, and W. Wang, "FER-Transformer: Facial Expression Recognition Transformer," 2022.

[15] S. Zhang, L. Liu, and J. Li, "Learning Local-Global Interactions for FER with Transformer Encoders," 2022.

[16] P. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Memory Fusion Network for Multimodal Sequential Learning," *AAAI*, 2018.

[17] Y.-H. Tsai, S.-H. Lai, and L.-P. Morency, "Multimodal Transformer for Unaligned Multimodal Language Sequences (MulT)," *ACL*, 2019.

[18] A. Jaegle, K. Simonyan, O. Vinyals, A. Zisserman, K. Kavukcuoglu, and C. B. Olah, "Perceiver IO: A General Architecture for Structured Inputs & Outputs," *ICML*, 2021.

[19] M. Rahman, X. Li, and S. Wang, "AV-BERT: Audio-Visual BERT for Emotion Recognition," 2022.