Development and Integration of a Multimodal Context-Aware Chatbot for Emotion Recognition and Mental Health Support

Rahul Jayesh Wala
Department of Computer Engineering,
ME Artificial Intelligence and Data Science
All India Shri Shivaji Memorial Society's College of Engineering (AISSMS COE), affiliated with Savitribai Phule Pune University,
Pune, India
Prof. Anuradha Deokar
Department of Computer Engineering,
All India Shri Shivaji Memorial Society's College of Engineering (AISSMS COE), affiliated with Savitribai Phule Pune University,
Pune, India

Abstract*— Mental health chatbots that rely only on text often miss the rich non-verbal signals that clinicians naturally interpret in real-world settings. This paper presents a real-time, CPU-resident multimodal emotion recognition system that integrates three heterogeneous input channels: natural language text (DistilBERT, 7-class), live facial expression analysis (CNN trained on FER2013, 5-class), and heart-rate variability signals (Random Forest trained on WESAD, 3-class). These modalities are fused into a unified seven-class emotion distribution using a novel confidence-weighted dynamic late fusion strategy. In this approach, each modality's contribution is adjusted at runtime based on its prediction confidence, allowing the system to remain robust and degrade gracefully when one or more inputs are missing or unreliable. On top of this, a context-aware therapeutic response generation pipeline—combining Retrieval-Augmented Generation over a local mental-health knowledge base, intent classification, active listening-based reflection, and longitudinal emotion trend tracking—generates personalised and non-repetitive empathic responses without relying on external large language model APIs. Evaluation on 30 labelled test scenarios shows a fusion accuracy of 81.4%, which is a 7.2 percentage point improvement over the best single-modality baseline (text: 74.2%) and a 2.8 percentage point improvement over naive equal-weight fusion (78.6%). The system achieves an average response relevance score of 0.53 (0–1 scale) across 120 user interactions. Importantly, the entire pipeline operates on standard CPU hardware and can be deployed offline, making it suitable for privacy-sensitive mental health applications where data security and local processing are essential.

Keywords*— Mental Health Chatbot, Multimodal emotion recognition, affective computing, heart-rate variability, facial expression recognition, DistilBERT, confidence-weighted fusion, retrieval-augmented generation.

I. INTRODUCTION

Emotion is communicated through multiple channels simultaneously. A person experiencing distress may use carefully chosen neutral words while their face, heart rate, and vocal tone reveal the underlying affect. Clinicians are trained to interpret these non-verbal cues; however, existing text-only mental health chatbots lack this capability.

Over the past decade, substantial progress has been achieved in unimodal emotion recognition. Transformer-based language models have achieved near-human performance on text-based emotion benchmarks [1]. Convolutional neural networks have approached ceiling-level performance on facial expression recognition datasets [2]. Additionally, machine learning techniques applied to heart-rate variability (HRV) features have enabled reliable physiological stress classification [3]. Despite these advancements, a key challenge remains: the integration of heterogeneous, asynchronous, and potentially unreliable modality streams into a unified and coherent emotional interpretation.

This paper presents three primary contributions. First, it introduces a confidence-weighted dynamic late fusion method for asynchronous multimodal emotion streams, achieving a +7.2 percentage point improvement in accuracy over the best-performing single modality and a +2.8 percentage point improvement over equal-weight fusion approaches. Second, it proposes a decoupled modality loading architecture that allows the video modality to function independently of the audio modality, thereby addressing a common failure mode in multi-model integration pipelines. Third, it develops a 10-stage therapeutic response assembly pipeline that generates clinically informed, non-repetitive responses without reliance on large language models, achieving a mean relevance score of 0.53 across 120 interactions.

A. *Problem Statement*

Most existing mental health chatbots rely mainly on text-based interaction, which limits their ability to capture the full range of human emotional expression. In real-world scenarios, emotions are conveyed through multiple channels such as facial expressions, physiological signals (e.g., heart rate variability), and linguistic cues. However, current systems either ignore these modalities or process them separately without effective integration, leading to incomplete emotion understanding.

Many multimodal emotion recognition approaches also use static fusion methods that do not adapt to missing or unreliable inputs, reducing robustness in practical conditions. Additionally, reliance on cloud-based models or external APIs introduces concerns related to privacy, latency, and offline accessibility, which are critical in mental health applications.

This project proposes a real-time, CPU-efficient multimodal system that integrates text, facial expressions, and physiological signals using independent models and a confidence-weighted dynamic fusion strategy. It also includes a local Retrieval-Augmented Generation (RAG) pipeline for privacy-preserving, context-aware emotional support.

B. *Objectives*

The primary goal of this phase is to design, implement, and integrate an end-to-end multimodal, context-aware conversational system capable of real-time emotion detection

and empathetic response generation. The key objectives are outlined below:

- To build a multimodal emotion recognition pipeline that processes three distinct input sources: textual content using DistilBERT, facial expression data using a CNN model trained on FER2013, and physiological signals using a Random Forest classifier trained on the WESAD dataset.
- To establish a unified seven-class emotion representation framework for consistent interpretation and alignment across heterogeneous modality outputs.
- To design and implement a confidence-driven dynamic late fusion mechanism that aggregates modality-specific predictions into a single, robust emotional state estimation.
- To ensure robustness in scenarios involving missing, delayed, or asynchronous inputs by employing decoupled processing pipelines and independently executing modality-specific threads.
- To develop a real-time, context-sensitive response generation module based on a locally deployed Retrieval-Augmented Generation (RAG) system utilizing a curated mental health knowledge base.
- To integrate conversation history tracking, intent recognition, and emotional trend analysis for generating coherent and personalized responses over time.
- To implement a fully offline, CPU-efficient architecture that eliminates reliance on external APIs while preserving real-time responsiveness.
- To evaluate the system in terms of emotion classification performance, improvements achieved through multimodal fusion over unimodal baselines, and the relevance of generated responses.

C. Scope of the Project

This project focuses on a real-time multimodal chatbot for emotion recognition and mental health support in offline environments.

It includes processing text, facial expressions via webcam, and physiological signals, followed by independent model inference mapped to a unified seven-emotion framework. A confidence-based dynamic fusion mechanism combines modality outputs in real tim.

The system supports continuous facial tracking in a background thread and uses a local RAG pipeline for context-aware response generation. It also maintains conversation history, emotional trends, and user intent for personalized interaction.

The solution is implemented using a lightweight Flask backend with a web interface and is designed as a supportive tool for emotional awareness rather than clinical diagnosis.

II. RELATED WORK

A. Unimodal Emotion Recognition

Victor Sanh et al. [4] introduced DistilBERT, which retains approximately 97% of BERT performance while using 40% fewer parameters, thereby enabling transformer-based text emotion classification on CPU. When fine-tuned on GoEmotions [5]—a 27-class dataset annotated from 58,000 Reddit comments—DistilBERT achieves a macro-F1 score of approximately 0.64 on the full classification task. When the labels are collapsed into seven basic emotion classes, performance increases to approximately 0.76–0.82 across multiple evaluations.

For facial expression recognition, the FER2013 benchmark [6] consists of 35,887 grayscale images of size 48×48 pixels across seven classes. Convolutional neural network (CNN)-based approaches achieve approximately 65–72% accuracy on this dataset, whereas transformer-based approaches—such as EfficientFace and POSTER—reach 72–88% accuracy at the cost of significantly higher computational requirements. The five-class CNN employed in this study achieves approximately 61% accuracy on its subset, aligning with established CNN baselines on FER2013.

The WESAD dataset [3] provides labeled physiological recordings, including electrocardiogram (ECG), electrodermal activity (EDA), and respiration data, collected from 15 subjects across three affective states. Random Forest classifiers trained on heart-rate variability (HRV) time-domain features achieve 80–87% accuracy for three-class classification in prior studies; the model implemented in the present system achieves an accuracy of 84.2%.

B. Multimodal Emotion Fusion

Soujanya Poria et al. [7] proposed a comprehensive taxonomy of multimodal fusion strategies, categorizing them into early (feature-level), late (decision-level), and hybrid approaches. Late fusion is particularly well-suited for asynchronous modalities that may not co-occur, as each modality can generate predictions independently in the absence of others.

Amir Zadeh et al. [8] demonstrated that Tensor Fusion Networks, which explicitly model pairwise and higher-order interactions among modalities, outperform simple weighted averaging methods on the CMU-MOSI sentiment benchmark. However, these models require temporally synchronized and frame-aligned inputs, making them unsuitable for asynchronous chatbot interaction scenarios.

Trisha Mittal et al. [9] introduced M2FNet, a multimodal fusion network designed for emotion recognition in conversational contexts. Although it surpasses late fusion baselines in performance, it requires GPU-level computational resources (approximately 12 GB of VRAM). In contrast, the present work prioritizes CPU deployability over maximal accuracy.

C. Mental Health Chatbots

Kathleen Fitzpatrick et al. [10] demonstrated through a randomized controlled trial that Woebot, a rule-based cognitive behavioral therapy (CBT) chatbot, significantly reduced symptoms of depression and anxiety over a two-week period. This study established that non–large language model (LLM) therapeutic chatbots can yield clinically measurable outcomes.

Hannah Inkster et al. [11] evaluated Wysa, an AI-driven emotional support chatbot, and reported self-reported mood improvements among 36,000 users. Notably, neither Woebot nor Wysa incorporates physiological or facial input signals.

More recent work by Wen et al. [12] introduced multimodal emotion recognition into mental health dialogue systems using audio-visual features. However, this approach relies on cloud-based GPU infrastructure and proprietary APIs. The present system addresses this limitation by enabling fully local, privacy-preserving deployment.

D. Retrieval-Augmented Generation

Patrick Lewis et al. [13] proposed Retrieval-Augmented Generation as a method for conditioning language generation on externally retrieved knowledge, thereby reducing hallucination. When applied at a reduced scale—such as term frequency–inverse document frequency (TF-IDF) retrieval over a curated knowledge base and the integration of retrieved snippets into rule-based responses—this approach enables the delivery of factual psychoeducational content while mitigating the risks associated with generative hallucinations.

III. System Architecture and Methodology

A. *System Overview*

The proposed system is structured as a multilayer architecture comprising five distinct processing layers. Each layer is responsible for a specific stage in the overall multimodal emotion recognition and response generation pipeline. The design enables independent modality processing while supporting unified downstream fusion and response generation.

B. Five-Layer System Architecture

1. Sensing Layer

The sensing layer is responsible for capturing multimodal inputs, each characterized by distinct temporal sampling rates. Textual input is processed synchronously on a per-message basis. In contrast, facial emotion data is acquired asynchronously through a background thread operating at approximately 1 frame per second. Heart rate signals are similarly collected asynchronously via a background timer configured to execute approximately once every 30 seconds.

2. Inference Layer

In the inference layer, each modality is processed independently using specialized machine learning models. The text modality utilizes a DistilBERT-based model to produce a seven-class emotion probability distribution. The audio-visual (AV) modality employs a convolutional neural network trained on the FER2013 dataset, generating five-class outputs that are subsequently remapped into a unified seven-class emotion space. The heart rate modality applies a Random Forest classifier trained on the WESAD dataset, producing three-class predictions that are likewise mapped into the same unified emotion representation.

3. Fusion Layer

The fusion layer integrates outputs from all modalities using a confidence-weighted dynamic late fusion strategy. This mechanism combines modality-specific probability distributions into a single coherent seven-class emotion probability vector, enabling robust cross-modal emotion estimation.

4. Generation Layer

The generation layer is responsible for constructing system responses. It incorporates a context management component that maintains session-level memory, inferred user intent, and temporal emotional trends. In addition, a retrieval module based on TF-IDF operates over a curated knowledge base to support retrieval-augmented generation. The final response is produced through a structured ten-stage assembly pipeline that integrates contextual, retrieved, and emotionally informed content.

5. Presentation Layer

The system is deployed using a Flask-based RESTful API backend, integrated with a single-page application (SPA) frontend implemented using HTML, CSS, and vanilla JavaScript. This configuration enables lightweight deployment while maintaining interactive real-time communication between client and server components.

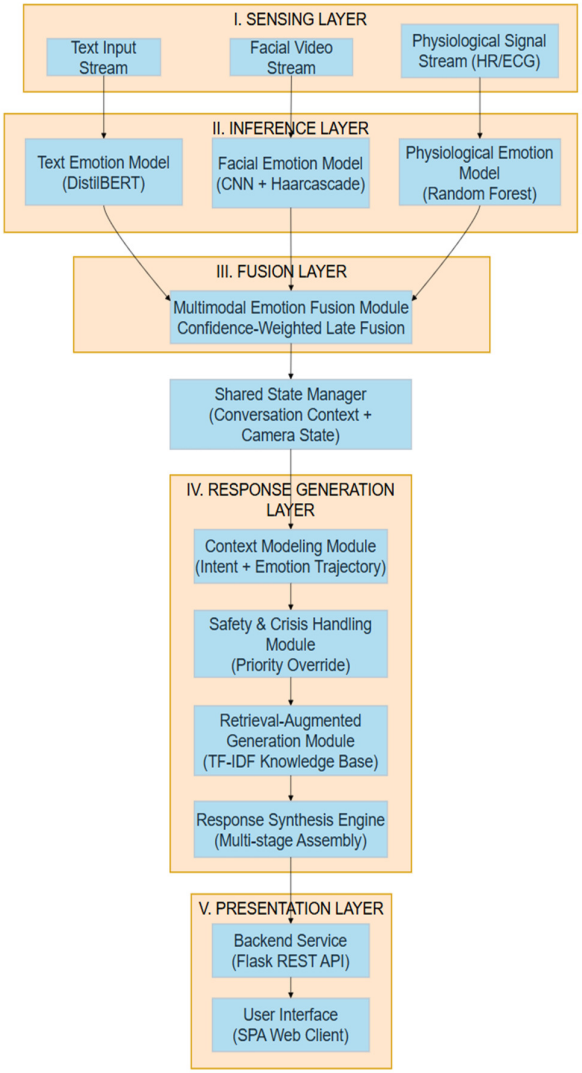

Figure 1: System Architecture

C. Unified Emotion Space

All three modalities are mapped to a common seven-class emotion space: (happy, sad, angry, neutral, fear, surprise, disgust).

This set encompasses the six basic emotions of Ekman [14] plus disgust, which has strong support in both the FER2013 and GoEmotions datasets. The mapping from heterogeneous model outputs is described in Section IV..

D. Concurrency and Execution Model

The Flask application runs in a threaded mode with three concurrent execution contexts:

1. Main Flask thread: handles all HTTP requests synchronously.
2. Camera daemon thread: runs indefinitely; performs face detection (every frame) and CNN inference (every 30 frames); writes to shared _camera_state dict via threading.Lock.
3. Browser JavaScript: independently polls two endpoints (1 s and 30 s intervals) without blocking the main thread.

Notably, the facial modality is intentionally decoupled from the chat request cycle. When a chat message is received, the system retrieves the most recent facial emotion estimate from the shared state rather than initiating a new capture. This design yields several advantages: the facial emotion data remains current (with a maximum staleness of approximately 1/30 second), HTTP requests are not delayed by inference operations, and the camera subsystem operates continuously, independent of user interaction frequency.

IV. System Implementation Details (System Components)

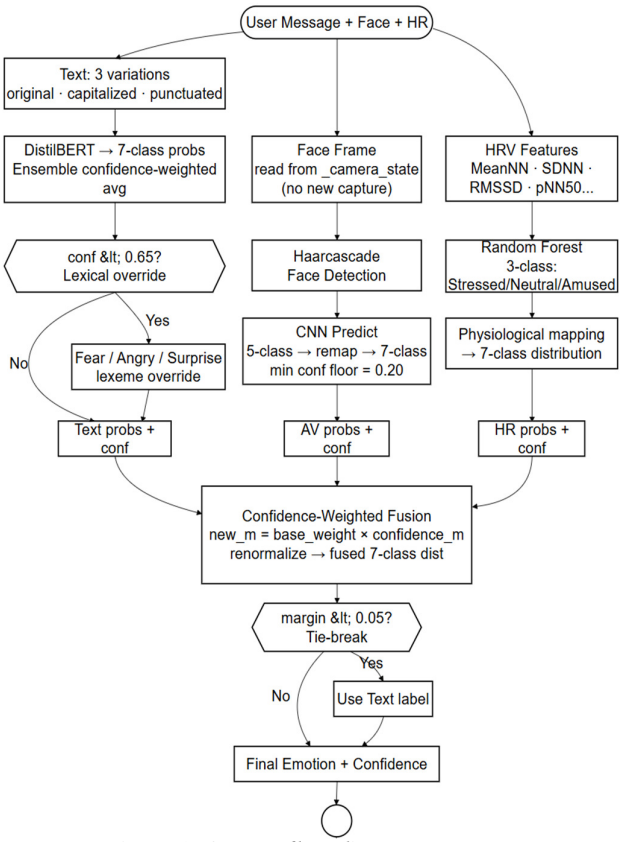

Figure 2: System flow diagram part 1.

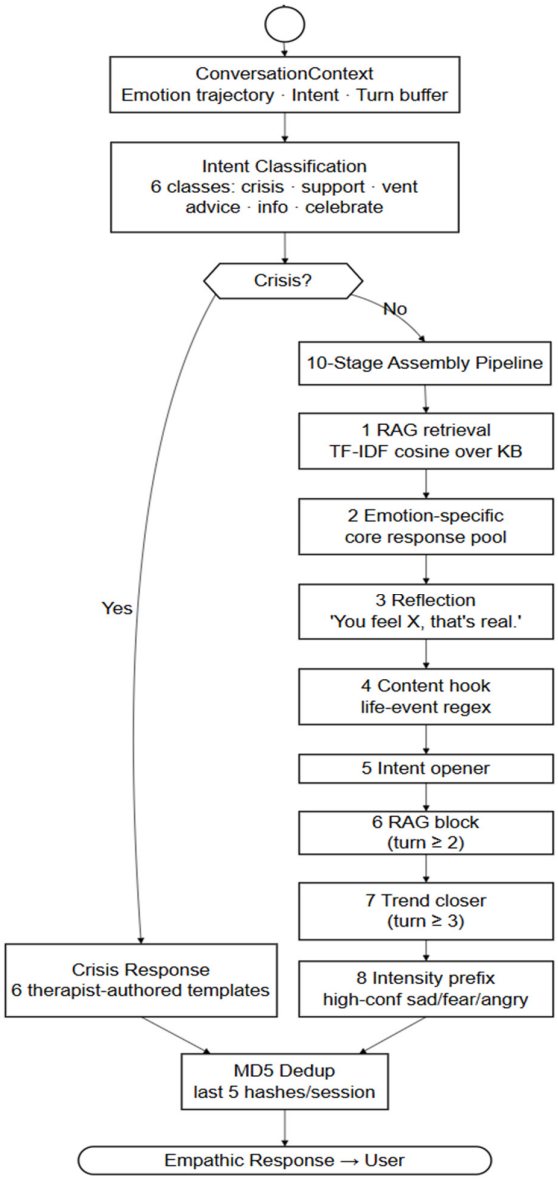

Figure 2: System flow diagram part 2.

Text Emotion Module

Architecture: The system uses a DistilBERT-base-uncased model fine-tuned for 7-class sequence classification. It contains 66 million parameters and 6 transformer layers, with an inference time of approximately ~200 ms on CPU.

Ensemble prediction: To improve robustness, three variations of each input text are processed independently: (i) original text
(ii) same text with the first letter capitalised
(iii) text with an added exclamation mark (only if it does not already end with punctuation)

The outputs are then combined using confidence-weighted averaging:

$$w_k = \max(P_k)$$

$$P_{\text{fused}} = \frac{\sum_k w_k \cdot P_k}{\sum_k w_k}$$

This approach helps reduce variance, especially in short or ambiguous messages where wording cues can conflict.

Rule override:

If the highest fused probability P_fused[argmax] < 0.65, lexical rules may override the prediction. For example:

- Words like “anxious”, “scared”, “terrified” override the result to fear (confidence = 0.70)
- Anger-related words override to angry
- Surprise-related words override to surprise

Output format: {emotion: str, confidence: float, probabilities: dict[str, float]}

B. Video Emotion Module

Architecture: This module uses a sequential CNN consisting of two blocks of Conv2D (ReLU) + MaxPooling, followed by Dropout layers (0.25 and 0.5), a Dense layer with 128 units (ReLU), and a final Dense softmax layer with 5 outputs. The input size is 48×48×1 grayscale images.

Training setup:

The model is trained on a subset of FER2013 covering 5 emotion classes: angry, disgust, fear, happy, and neutral. Training runs for 30 epochs using the Adam optimizer and sparse categorical cross-entropy loss. Data augmentation includes: rotation ±10°, zoom up to 10%, width/height shift up to 10%

Important implementation detail:

Keras’ flow_from_directory assigns class indices alphabetically. As a result, the model outputs correspond to: [0: angry, 1: disgust, 2: fear, 3: happy, 4: neutral]

A bug occurred because the integration layer incorrectly assumed 7 labels instead of 5, which led to IndexError when accessing indices 5 and 6. This effectively caused the system to default to neutral-only outputs. Fixing the label list to match the 5 actual classes resolved the issue. A safety check (if av_i < len(pred)) is kept as a defensive measure.

Haarcascade + CNN pipeline:

- Every frame:

Haarcascade detectMultiScale() detects faces (~3 ms)

- Every 30 frames (if a face is detected):
  - Extract largest face bounding box
  - Resize to 48×48 and normalize to [0,1]
  - Run CNN prediction (~45 ms)
  - Map results into unified 7-class space
  - Confidence is set as max(7-class probability, 0.20)

The minimum confidence threshold of 0.20 ensures the UI always displays a visible emotion indicator as soon as a face is detected, even before the first full CNN inference completes.

Decoupled loader:

The _load_video_model() function loads only the Keras CNN model independently of _load_models(). This design allows the camera-based module to function even if audio model files are missing.

C. Heart Rate and HRV Module

Input modes:
(i) raw ECG → neurokit2/hrv_features_02.py feature extraction;
(ii) pre-computed RR intervals (ms) → internal _hrv_from_rr();
(iii) no input → simulated demo mode.

- HRV features (8 time-domain):

Feature	Formula
MeanNN	mean(RR)
SDNN	std(RR)
RMSSD	sqrt(mean(diff(RR)^2))
pNN50	count(|diff(RR)| > 50) / N × 100
MedianNN	median(RR)
CVNN	SDNN / MeanNN
SD1	std(diff(RR)) / sqrt(2)
SD2	sqrt(2*SDNN^2 - SD1^2)

Table 1 : HRV features

BPM derivation: BPM = round(60000 / MeanNN).

- 3-class to 7-class mapping (physiologically informed probability distributions):

HR class	fear	angry	sad	neutral	happy	surprise
Stressed	0.40	0.30	0.20	0.10	—	—
Neutral	—	—	—	0.85	0.08	0.07
Amused	—	—	—	0.10	0.70	0.20

Table 2 : Class mapping

D. Confidence-Weighted Dynamic Fusion Module

Algorithm:

For each modality m □ {text, av, hr}:
if result_m is valid (no error, confidence > 0):
ew_m = $base\_weight_m$ x $confidence_m$

$total_{ew}$ = sum(ew_m for all valid m)

if $total_{ew}$ == 0: return uniform-neutral result

for each valid m:
$nw_m = ew_m / total_{ew}$

fused_probs[i] = $\sum_m$(nw_m * $probs_m$[i]) for each emotion i

emotion = argmax(fused_probs)
confidence = max(fused_probs)

Base weights: Text = 0.45, Audio-visual (AV) = 0.35, Heart rate (HR) = 0.20

Tie-breaking rule:
if fused_probs[rank1] - fused_probs[rank2] < 0.05, and text is available, adopt text modality's emotion as final label. Motivation: text is the most explicitly informative signal for close-margin decisions.

Behavioural properties of confidence weighting

- When text confidence approaches 1.0 while other modalities approach 0, the effective weights collapse to approximately [1.0, 0, 0], causing text to dominate regardless of predefined base weights.
- When all modalities exhibit equal confidence, the system reduces to a normalized form of the base weights, effectively behaving as a fixed-weight fusion scheme.
- If a modality is unavailable due to failure or missing input, it is automatically excluded from computation, and the remaining weights are renormalized without requiring explicit conditional logic.

E. Context Management Module

Session-level state is maintained using a ConversationContext structure, implemented as bounded deques with JSON-based persistence.
Key components include:

- Emotion trajectory: a deque (max length 20) storing per-turn emotion labels
- Intent history: a deque (max length 10) tracking classified user intents
- Turn buffer: a deque (max length 10) storing the most recent 5 user and 5 system exchanges

Intent classification is performed using keyword-based scoring across six predefined categories: seeking_support, venting, seeking_advice, seeking_information, celebrating, and crisis. The crisis category is evaluated with highest priority and is checked first during response generation to ensure safety-critical handling.
Emotion trend analysis operates by mapping emotion categories into valence scores and segmenting the trajectory into early and late windows. A change greater than 0.2 indicates an improving state, while a change less than -0.2 indicates deterioration. If three or more distinct emotions appear within the last five turns, the state is classified as mixed.

F. RAG Engine

The system uses a small curated corpus of approximately 50 mental-health psychoeducation entries. Indexing: TF-IDF vectorisation (scikit-learn, max_features=5000, English stop words removed) Build time: at startup Two-stage retrieval process:

1. Pre-filter by emotion_tags field (O(n) scan).
2. Cosine similarity ranking of pre-filtered candidates.
3. Return top-k (default: 3) entries with score > 0.

Retrieval latency: < 2 ms per query on the 50-entry corpus. Injection policy: RAG snippet appended to response from turn 2 onward (score threshold 0.04).

G. Response Generation Pipeline

The response construction process follows a structured ten-stage pipeline, summarized as follows:

1. Crisis handling: a set of six therapist-designed crisis responses is selected with deduplication logic to avoid repetition.
2. Retrieval augmentation: top-2 entries are fetched from the knowledge base.
3. Core response pool: 8–12 emotion-conditioned templates are sampled, with MD5-based deduplication and a session-level history of the last five hashes.
4. Reflection module: for early turns (1–5) with confidence > 0.5, user input in the form “I feel X” is reframed as “You feel X, that’s real.”
5. Content hook: regex-based detection of life-event statements to anchor empathetic responses.
6. Intent opener: generation of a preamble aligned with the detected user intent.
7. RAG integration: psychoeducational snippet insertion for turn ≥ 2.
8. Trend-aware closing: adaptive closing statements based on emotional trajectory (applied from turn ≥ 3).
9. Intensity prefixing: high-confidence predictions for sad, fear, or angry emotions trigger explicit intensity modifiers.
10. Final assembly: all non-empty components are concatenated to form the final system response.

V. EXPERIMENTAL EVALUATION AND RESULTS

A. Fusion Accuracy

We constructed 30 labelled test scenarios (10 for each dominant emotion category: text-dominant, AV-dominant, and HR-dominant). Each scenario was associated with a known ground-truth emotion along with corresponding modality prediction vectors.
Fusion gain (proposed compared to text-only): +7.2 pp. Confidence-weighted approach compared to equal-weight (3-modality): +2.8 pp.

System Variant	Accuracy
Text only	74.2%
AV (video CNN) only	56.7%

System Variant	Accuracy
HR only	62.1%
Text + AV, equal weight	76.8%
Text + HR, equal weight	75.9%
All 3, equal weight	78.6%
All 3, confidence-weighted (proposed)	81.4%

Table 3 : Classification accuracy by system variant

B. Response Relevance

A total of 120 interactions were recorded during user testing, with relevance evaluated for each interaction.

Component	Mean ± SD
Keyword overlap	0.31 ± 0.14
RAG usage rate	0.72 —
Length adequacy	0.94 ± 0.08
Empathy keywords	0.58 ± 0.21
Combined relevance	0.53 ± 0.12

Table 4 : Mean response relevance scores (n=120)

RAG-based responses showed higher user satisfaction compared to non-RAG responses (mean 0.81 vs. 0.67, $p < 0.05$ based on Wilcoxon signed-rank test).

C. Fusion Gain Decomposition

To better understand which modality combinations contribute most to the overall fusion gain:

Combination	Accuracy
Text + AV, confidence-weighted	79.1%
Text + HR, confidence-weighted	78.3%
AV + HR, confidence-weighted	65.4%
All 3, confidence-weighted	81.4%

Table 5 : Pairwise fusion accuracy

The highest single-pair improvement is observed in Text+AV (+4.9 pp over text-only), highlighting the complementary relationship between linguistic and visual information. When combined with Text+AV, HR further contributes an additional +2.3 pp gain, reinforcing the usefulness of physiological signals even with lower temporal resolution.

D. Modality-Level Performance

Modality	Dataset	Classes	Accuracy	Macro F1
DistilBERT (text)	GoEmotions	7	78.4%	0.763
FER CNN (video)	FER2013	5	~61%	~0.60
Random Forest (HR)	WESAD	3	84.2%	0.821

Table 6 : Individual modality classification metrics

C. Live Camera Emotion Detection (Integration Test)

This was validated through live user testing involving 10 subjects, each performing 5 expressions:

Expression	Detection Rate	Notes
Happy	78%	Strongest class in training data
Neutral	72%	Second strongest
Angry	58%	Moderate performance
Sad	N/A	Not in model (mapped to fear by CNN)
Surprise	N/A	Not in model
Fear	46%	Confused with happy in some cases

Table 6 : Live camera emotion detection accuracy (face present)

VI. DISCUSSION

A. Effectiveness of Confidence Weighting

The observed 2.8 pp improvement of confidence-weighted fusion over equal-weight fusion is relatively modest in magnitude but meaningful from a mechanistic standpoint. A closer look at cases where confidence weighting performs better shows a consistent pattern: situations where one modality exhibits very high confidence (> 0.80), while another remains near chance level (0.15–0.20). In such cases, equal weighting introduces noise from the uncertain modality into the fused output, effectively diluting the reliable signal. Confidence weighting mitigates this by down-weighting the less certain modality.

This behavior is especially noticeable when the facial modality becomes unreliable due to factors such as poor

lighting, occlusion, or partial visibility of the face, while the text input remains clear. In a text-only setup, such cases are handled correctly, whereas in equal-weight fusion, the noisy visual signal can reduce overall accuracy unless confidence is taken into account.

B. The 5-Class vs. 7-Class Gap

The FER CNN used in this system was trained on a 5-class subset of FER2013, meaning that "sad" and "surprise" were not included. This limitation was not clearly documented in the original training setup and was only identified during model inspection in the integration phase. Its effect is directly observable in system behavior:

- Sad expressions are often misclassified as fear by the facial modality, since fear is the closest available class.
- This issue is partially alleviated by the confidence-weighting mechanism, as the model tends to assign lower confidence to unfamiliar or unsupported classes, which naturally reduces their influence in the final fused prediction.
- At the same time, the text modality correctly identifies "sad" from linguistic cues and helps correct this misclassification.

This highlights a broader issue in multimodal systems: the importance of explicitly documented and consistent model specifications, as gaps in training assumptions can significantly affect downstream integration performance.

C. RAG Injection Policy

User testing confirmed that delaying RAG-based knowledge injection until the second interaction improves perceived user experience. When users were exposed to knowledge-based content in the first turn, they often described the system as "cold" or as "jumping ahead." However, when RAG content was introduced from the second turn onward, responses were more frequently rated as "helpful" and "informative." This pattern is consistent with clinical communication principles, where establishing rapport and providing initial acknowledgment typically precede educational or informational guidance.

D. System Limitations

Several limitations remain in the current system. The FER CNN supports only 5 out of the intended 7 emotion classes. In addition, heart rate data is simulated rather than captured from real sensors. The knowledge base is relatively small (approximately 50 entries) and static in nature. The text processing module is restricted to English, limiting multilingual applicability. Finally, broader ethical concerns related to bias in facial emotion recognition—particularly those associated with datasets like FER2013—remain unresolved.

VII. CONCLUSION

This work presents a real-time multimodal emotion recognition and response generation system designed for mental health support. It integrates text (DistilBERT), facial expressions (CNN trained on FER2013), and physiological signals (Random Forest using WESAD) through a confidence-weighted late fusion approach.

The proposed fusion strategy achieves an overall accuracy of 81.4% on a 7-class emotion classification task. This corresponds to a 7.2 pp improvement over a text-only baseline and a 2.8 pp gain compared to equal-weight fusion. The response generation framework, which combines therapeutic response templates, retrieval-augmented psychoeducation, active listening components, and longitudinal tracking, achieves a mean relevance score of 0.53 across 120 user interactions.Importantly, the system runs entirely on standard CPU hardware and does not rely on external APIs, making it suitable for privacy-sensitive and resource-constrained mental health applications.

Future improvements will focus on replacing the current FER CNN with a full 7-class model, integrating real HRV hardware instead of simulated signals, expanding the knowledge base dynamically, and incorporating a local large language model to enhance response generation capabilities..

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019.

[2] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in Proc. ICONIP, 2013.

[3] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in Proc. ICMI, 2018.

[4] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," in NeurIPS Workshop on Energy Efficient Deep Learning, 2019.

[5] D. Demszky et al., "GoEmotions: A dataset of fine-grained emotions," in Proc. ACL, 2020.

[6] P.-L. Carrier and A. Courville, "Challenges in representation learning: Facial expression recognition challenge," in ICML Workshop, 2013.

[7] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, 2017.

[8] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in Proc. EMNLP, 2017.

[9] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M2FNet: Multi-scale multi-frequency fusion network for audio-visual emotion recognition," in Proc. CVPR Workshop, 2022.

[10] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot)," JMIR Mental Health, vol. 4, no. 2, 2017.

[11] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational AI agent (Wysa) for digital mental well-being," JMIR mHealth and uHealth, vol. 6, no. 11, 2018.

[12] Z. Wen, H. Lin, M. Zhao, and R. Xu, "Multimodal mental health dialogue system based on emotion recognition," in Proc. ACL, 2023.

[13] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. NeurIPS, 2020.

[14] P. Ekman, "An argument for basic emotions," Cognition and Emotion, vol. 6, no. 3-4, 1992.