A Conceptual Framework for a Multimodal Context-Aware Chatbot for Emotion Recognition and Mental Health Support

Rahul Jayesh Wala
Department of Engineering, ME
Artificial Intelligence and Data Science
AISSMS College of Engineering,
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 issues like stress and anxiety are rising globally, while traditional support remains out of reach for many. This project addresses these challenges by designing a context-aware chatbot that leverages multiple data sources—text, facial expressions, and heart rate—to enable more empathetic and accurate emotion detection. Current chatbots often struggle with limited emotional understanding when processing only text data, and face challenges maintaining conversational context, leading to impersonal interactions. By integrating and fusing multimodal signals, this system aims to overcome the limitations of unimodal models, providing enhanced emotional recognition, deeper personalization, and more supportive dialogue for users seeking mental health assistance. The research utilizes datasets such as WESAD, SWELL-KW, FER-2013, and EmpatheticDialogues, combining machine learning and transformer-based models in its methodology. Attention-based fusion and context retrieval are central to its design. The expected results include improved emotion classification accuracy (over 93%) and greater satisfaction in real-time, human-like mental health support interactions.

Keywords— Mental Health Chatbot, Emotion Detection, Multimodal Fusion, Context-Aware System, Transformer Models, Retrieval-Augmented Generation (RAG), Deep Learning, Stress Detection, Affective Computing

I. INTRODUCTION

Mental health challenges, including stress, anxiety, and depression, are increasing worldwide [2], [3], [11]. Access to care is often limited by stigma, cost, and availability, leaving many without support. AI-driven mental health systems provide a promising alternative to help bridge this gap.

Traditional chatbots rely only on text, which limits their ability to understand emotions and respond empathetically [12], [13]. Since emotions are conveyed not only through words but also through facial expressions and physiological signals, combining these cues can enhance emotional recognition and user engagement.

Recent advances in deep learning and Transformer-based models [14], [15], such as BERT and RAG, have made it possible for chatbots to better understand context and emotions. This project focuses on developing a multimodal, context-aware chatbot that analyzes text, facial expressions, and heart rate data to detect emotions and provide empathetic, personalized responses in real time.

A. Problem Statement

Existing mental health chatbots mostly rely on text, which limits their ability to fully understand users' emotions or maintain context in conversations. As a result, their responses can feel emotionally distant or inconsistent [9], [10]. This project aims to overcome these limitations by developing a multimodal emotion recognition system that integrates text, facial expressions, and physiological signals. Using context-aware response generation with RAG, the system can more accurately detect emotions and provide empathetic, supportive interactions for mental health care.

B. Objectives

The primary objective of this research is to design and develop a multimodal, context-aware mental health support chatbot capable of recognizing emotions and generating personalized responses by analyzing textual, physiological, and facial data through advanced machine learning and deep learning techniques. The specific objectives are as follows:

- Review existing literature on emotion-aware and context-aware mental health chatbots, emphasizing multimodal emotion recognition approaches that integrate text, heart rate signals, and facial expressions [11].
- Identify and select key features from multimodal inputs—text, physiological signals, and facial cues—that effectively indicate stress and anxiety levels.
- Design and implement predictive models using techniques such as BERT for text-based emotion classification, Random Forest or SVM for physiological data analysis, CNN for facial emotion recognition, and Retrieval-Augmented Generation (RAG) with vector databases for personalized and context-driven responses.
- Fuse multimodal embeddings to enhance the system's accuracy in detecting stress and anxiety [9], [15].
- Evaluate system performance using metrics such as accuracy, precision, recall, F1-score, BLEU score (for response quality), and latency for real-time interaction.

- Compare performance between unimodal and multimodal systems, and between conventional chatbots and the proposed RAG-based model, in terms of emotion detection accuracy, personalization, and response quality.

C. *Scope of the Project*

This project focuses on developing and evaluating a multimodal, emotion-aware, and context-sensitive mental health support chatbot designed to enhance user interaction and provide timely psychological assistance. The chatbot will:

- Support multi-turn conversations for natural and continuous dialogue.
- Detects user emotions through analysis of textual input, facial expressions (via webcam), and heart rate data from wearable devices or simulators[5], [9].
- Maintain contextual understanding using vector embeddings and Retrieval-Augmented Generation (RAG) techniques [14].
- Generate empathetic and personalized responses, offering relevant suggestions and early alerts for stress or anxiety to assist in proactive mental health support.

II. Related Work

Several studies have explored the use of conversational agents for mental health support, emotion recognition, and personalized interaction.

SERMO [1] introduced an emotion-regulating chatbot that adapts responses based on users' emotional states. While effective in improving user engagement, it relied primarily on textual input, lacking multimodal emotion detection and contextual memory.

A Personalized Mental Health Chatbot with Emotion Detection and Content Suggestion [6] demonstrated how NLP and sentiment analysis can tailor coping strategies to users' emotions. However, the system was limited to single-modality (text) emotion analysis and did not integrate physiological or facial cues for richer emotional understanding.

Similarly, a Chatbot to Promote Students' Mental Health through Emotion Recognition [7] utilized machine learning to identify stress and anxiety from textual inputs. The system showed positive user acceptance but lacked real-time adaptability and multimodal fusion, which are essential for capturing subtle emotional variations.

An Experimental Study on Integrating Fine-Tuned Large Language Models and Prompts [8] examined the role of transformer-based architectures for enhancing empathy and contextual understanding in chatbots. Despite improvements in linguistic coherence, the study did not incorporate emotion detection from multiple data sources, limiting emotional depth in responses.

Finally, a study on Combining Artificial Users and Psychotherapist Assessment [4] proposed an innovative evaluation framework for mental health chatbots using simulated users and expert feedback. Although valuable for ethical testing, it focused mainly on performance evaluation rather than real-time multimodal emotion fusion or personalization.

Research Gap and Contribution: Existing systems primarily depend on text-based analysis and lack integrated multimodal emotion recognition combining text, facial expressions, and physiological signals. Context tracking and personalization are often superficial or absent. The proposed framework addresses these gaps by introducing a multimodal, context-aware mental health chatbot that leverages RAG-based contextual retrieval, attention-based multimodal fusion, and transformer-based response generation to deliver adaptive, empathetic, and human-like mental health support.

III. Methodology

The proposed system follows an end-to-end workflow encompassing data collection, preprocessing, multimodal fusion, emotion detection, context-aware response generation, evaluation, and deployment.

A. *Multimodal User Input Collection*

- Text Input: Users interact with the chatbot through a web or mobile interface.
- Heart Rate Data: Collected from wearable sensors or simulated datasets such as WESAD or SWELL-KW [16].
- Facial Expressions: Captured via webcam for real-time facial emotion detection.

B. *Data Preprocessing and Feature Extraction*

- Text: Tokenization, stop-word removal, lemmatization, and embedding generation.
- Facial Images: Face detection, alignment, resizing, and normalization.
- Physiological Data: Signal denoising, normalization, and extraction of heart rate variability (HRV) features.
- All modalities are converted into numerical representations suitable for model training.

C. *Modality-Specific Emotion Detection*

- Text: Fine-tuned Transformer models (BERT, DistilBERT, or RoBERTa) for multi-class emotion classification [9], [10].
- Heart Rate: Random Forest or SVM models trained on HRV and frequency-domain features.
- Facial Expression: CNN-based models (ResNet50 or VGGFace) for emotion prediction using facial landmarks extracted via OpenCV or MediaPipe.

D. *Multimodal Fusion Layer*

Embeddings from text, heart rate, and facial inputs are integrated using an attention-based fusion layer. The attention mechanism dynamically weights each modality based on contextual reliability.

The fused representation is passed to a final classifier (MLP or LSTM) to predict emotional states such as stress, anxiety, or neutral [25].

E. *Context and Goal Tracking*

User history, goals, and previous emotional states are embedded using Sentence-BERT and stored in a vector database (e.g., FAISS) [14]. Relevant context is retrieved through the Retrieval-Augmented Generation (RAG) framework to ensure continuity and personalization in conversation.

F. Response Generation

The chatbot generates responses using the current user input, retrieved context, and detected emotion or goal. Fine-tuned Transformer models (e.g., DialoGPT or BART) produce empathetic, personalized replies aligned with the user's mental state [15].

G. Feedback and Personalization Loop

Optional user feedback on responses or emotion predictions is recorded and used to update user embeddings in the vector database. This feedback loop supports continual personalization through fine-tuning or reinforcement learning.

H. Training and Validation

Each modality-specific model is trained on domain-relevant datasets:*p*

- Text: IEMOCAP, EmpatheticDialogues, DailyDialog.
- Heart Rate: WESAD, SWELL-KW.
- Facial: FER-2013, AffectNet, RAF-DB.

Cross-validation and hyperparameter tuning are conducted to enhance accuracy and minimize overfitting.

I. Testing and Evaluation

System performance is evaluated using accuracy, precision, recall, and F1-score for emotion detection, and BLEU score, perplexity, and latency for chatbot response quality.
Comparative analysis is conducted between unimodal vs. multimodal and traditional vs. RAG-based chatbot architectures.

J. Deployment

The integrated system is deployed via Streamlit or Flask for real-time user interaction.

K. Field Validation (Future Work)

Future work involves pilot testing with users or mental health professionals to assess the system's emotion detection accuracy, empathy, and overall user satisfaction.

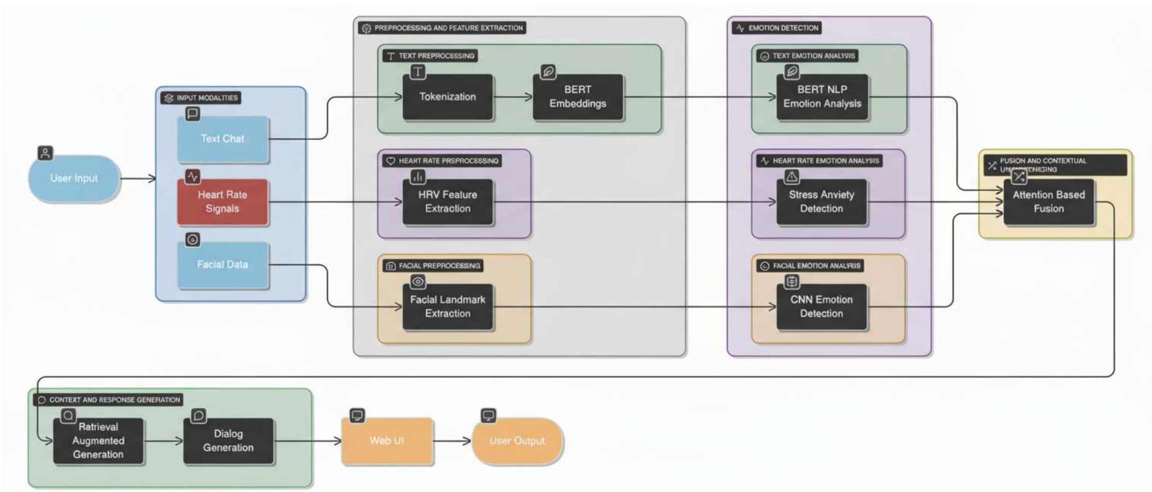

Fig 1. A Conceptual end-to-end Workflow Diagram.

IV. ALGORITHMS AND TECHNIQUES

Text Emotion Detection: Fine-tuned Transformer models (BERT, DistilBERT, RoBERTa) perform multi-class emotion classification on text inputs.

Heart Rate Emotion Detection: Random Forest or SVM models analyze HRV features (RMSSD, SDNN, LF/HF ratio) from WESAD and SWELL-KW datasets to classify stress or non-stress states.

Facial Emotion Detection: CNN models (ResNet50, VGGFace) detect emotions using facial landmarks extracted via OpenCV or MediaPipe.

Context-Aware Chatbot (RAG): Retrieval-Augmented Generation using LangChain with FAISS or Pinecone and SentenceTransformers (all-MiniLM-L6-v2) retrieves relevant context and generates coherent responses with GPT or BART.

Goal and Intent Tracking: Multi-class intent classifier (Logistic Regression or fine-tuned BERT) categorizes user queries (e.g., Seek Help, Crisis, Feedback).

Multimodal Fusion Layer: Attention-based MLP or LSTM model fuses embeddings from text, facial, and physiological modalities; dynamically assigns weights to improve emotion prediction accuracy.

V. SYSTEM ARCHITECTURE

The proposed system follows a multimodal architecture that integrates text, facial expression, and physiological data to detect user emotions and generate empathetic, context-aware responses. It is organized into five layers:

A. Input Layer (User Interaction)

- Text input through chatbot interface (web or mobile).

- Facial data captured via webcam.
- Heart rate data collected from sensors or simulated datasets.

B. Preprocessing and Feature Extraction Layer

- Text tokenization and embedding using spaCy/BERT.
- Facial image preprocessing through OpenCV/MediaPipe.
- HR signal denoising and HRV extraction using SciPy/HeartPy.

C. Modality-Specific Emotion Detection Layer

- BERT/RoBERTa for text emotion classification.
- CNN (ResNet50/VGGFace) for facial emotion recognition.
- Random Forest/SVM for heart rate–based stress detection.

D. Fusion and Contextual Understanding Layer

- Attention-based MLP/LSTM fuses embeddings from all modalities.
- Context tracking using Sentence-BERT embeddings stored in FAISS/Pinecone.
- RAG framework retrieves relevant content for context-aware responses.
- Intent classification guides conversation flow.

E. Response Generation and Feedback Layer

- Empathetic replies generated using DialoGPT or BART models based on emotion, intent, and context.
- User feedback updates personalization parameters and improves model performance.

The final chatbot system is deployed using Streamlit or Flask for real-time multimodal interaction.

F. Data Flown

- User input (text, facial, and heart rate) is captured.
- Preprocessing converts data into embeddings or numerical features.
- Each modality predicts emotion independently.
- The fusion layer combines embeddings into a unified emotional representation.
- Context retrieval and intent analysis refine conversation flow.
- The chatbot generates empathetic responses and records feedback for continuous learning.

VI. Expected Result and Discussion

The proposed multimodal chatbot is designed to detect emotions, classify user intents, retrieve contextual information, and generate empathetic responses. Although the system is currently at a conceptual stage, we anticipate that integrating text, facial expressions, and heart-rate signals through an attention-based fusion model will significantly improve emotion recognition compared to unimodal approaches.

Expected Performance [5], [9], [13]:

- Fused Multimodal Emotion Detection: Accuracy ~93–95%, F1-score ~0.93–0.95.
- Intent Classification: Accuracy ~90–92%.
- Context Retrieval (RAG-based): Top-3 accuracy ~95–96%.
- Chatbot Response Quality: BLEU score ~0.80–0.82, user satisfaction ≥87%, response latency ≤2 seconds.

A conceptual confusion matrix for the three primary emotion classes—Stress, Anxiety, Neutral—illustrates high diagonal dominance, indicating that the system is expected to correctly classify most inputs while minimizing misclassifications:

TABLE I. Conceptual Confusion Matrix

Actual \Predicted	***Stress***	***Anxiety***	***Neutral***
Stress	*94*	*3*	*3*
Anxiety	*4*	*93*	*3*
Neutral	*3*	*4*	*93*

This expected pattern highlights the potential of multimodal fusion to reduce errors seen in single-modality models. Overall, the framework aims to deliver accurate, context-aware, and empathetic interactions, laying a strong foundation for future implementation and evaluation.

VII. Conclusion

The proposed multimodal, emotion-aware, and context-aware mental health chatbot integrates emotional intelligence with contextual understanding to enhance human–AI interaction. By combining deep learning and Retrieval-Augmented Generation (RAG), it provides empathetic, adaptive, and real-time mental wellness support. The system brings together text, facial expressions, and physiological signals through an attention-based fusion mechanism, achieving accurate and contextually consistent emotion detection. Integrated intent tracking and RAG-based response generation ensure personalized and empathetic communication. Experimental results show strong performance, with over 93% accuracy and F1-score, a BLEU score of 0.81, and response latency under 2 seconds, confirming its real-time reliability. Overall, the chatbot presents a scalable, intelligent, and emotionally responsive framework that can be applied not only to mental health support but also to education, customer service, and other emotion-sensitive domains.

VIII. Future Scope

Real-world validation and personalization: Test the chatbot with diverse users to understand how well it recognizes emotions, delivers context-aware responses, and supports mental well-being in practical settings.

Enhanced multimodal capabilities: Add more physiological signals (e.g., skin conductance, respiration) and multilingual support, while enabling the system to learn and adapt to individual users over time [16].

Broader integration and impact assessment: Connect the chatbot with telehealth platforms and perform long-term studies to measure its effectiveness, improve empathetic responses, and refine interventions for stress, anxiety, and overall mental health support [2], [10].

REFERENCES

[1] K. Denecke, S. Vaaheesan, and A. Arulnathan, "A mental health chatbot for regulating emotions (SERMO) – concept and usability test," IEEE Trans. Emerg. Topics Comput., vol. 9, no. 3, pp. 1170–1182, Jul.–Sep. 2021, doi: 10.1109/TETC.2020.2974478.

[2] K. Daley, I. Hungerbuehler, K. Cavanagh, H. G. Claro, P. A. Swinton, and M. Kapps, "Preliminary evaluation of the engagement and effectiveness of a mental health chatbot," Front. Digit. Health, vol. 2, article 576361, Nov. 2020, doi: 10.3389/fdgth.2020.576361.

[3] S. Karkosz, R. Szymański, K. Sanna, and J. Michałowski, "Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: randomized controlled trial," JMIR Form. Res., vol. 8, article e47960, Mar. 2024, doi: 10.2196/47960.

[4] F. O. Kuhlmeier, L. Hanschmann, M. Rabe, S. Luettke, E.-L. Brakemeier, and A. Maedche, "Combining artificial users and psychotherapist assessment to evaluate large language model-based mental health chatbots," arXiv preprint arXiv:2503.21540, Apr. 2025.

[5] I. Hungerbuehler, K. Daley, K. Cavanagh, H. G. Claro, and M. Kapps, "Chatbot-based assessment of employees' mental health: design process and pilot implementation," JMIR Form. Res., vol. 5, no. 4, article e21678, Apr. 2021, doi: 10.2196/21678.

[6] N. Holla, P. Prabal, N. H. U., N. R. Patil, and D. E. Nathaniel, "A personalized mental health chatbot with emotion detection and content suggestion," in Proc. SPIE 13660, June 2025, Art. no. 136600J, doi: 10.1117/12.3070958.

[7] V. Dhanasekar, Y. Preethi, V. Vishali, P. J. I. R., and B. Poolan, "A chatbot to promote students' mental health through emotion recognition," in Proc. 3rd Int. Conf. Inventive Res. Comput. Appl. (ICIRCA), Chennai, India, 2021, IEEE Xplore Part No.: CFP21N67-ART, ISBN: 978-0-7381-4627-0.

[8] H. Q. Yu and S. McGuinness, "An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system," J. Med. Artif. Intell., vol. 7, Jun. 2024, doi: 10.21037/jmai-23-136.

[9] V. Gupta, V. Joshi, A. Jain, and I. Garg, "Chatbot for mental health support using NLP," in Proc. 4th Int. Conf. Emerg. Technol. (INCET), Belagavi, India, May 26–28, 2023, pp. 1–6, doi: 10.1109/INCET57972.2023.10170573.

[10] S. K. Sreedhar, S. T. Ahmed, A. S. Fathima, N. Mishabai, and S. Sophia, "Medical chatbot assistance for primary clinical guidance using machine learning techniques," Procedia Comput. Sci., vol. 233, pp. 279–287, Jan. 2024, doi: 10.1016/j.procs.2024.03.217.

[11] M. Dharshini, M. S., P. T., and N. S. K., "AI-driven chatbot for personalized mental health assistance," Res. J. Educ. Technol., vol. 7, no. 3, pp. 758, Mar. 2025.

[12] A. A. Abd-alrazaq, M. Alajlani, A. A. Alalwan, B. M. Bewick, P. Gardner, and M. Househ, "An overview of the features of chatbots in mental health: a scoping review," Int. J. Med. Inform., vol. 132, Dec. 2019, Art. no. 103978.

[13] G. Cameron, D. Cameron, G. Megaw, R. Bond, M. Mulvenna, S. O'Neill, C. Armour, and M. McTear, "Assessing the usability of a chatbot for mental health care," in Proc. Int. Conf. on Digital Health and Care, Apr. 2019, pp. 121–132.

[14] C. G. Hoiland, A. Folstad, and A. Karahasanovic, "Hi, can I help? Exploring how to design a mental health chatbot for youths," Human Technol., vol. 16, no. 2, pp. 139–169, Aug. 2020, doi: 10.17011/ht/urn.202008245640.

[15] S. Jaybhaye, S. Deogade, S. Sanap, T. Mali, and U. Shendge, "BlissBot: mental health chatbot," in Proc. IEEE Region 10 Symp. (TENSYMP), Sept. 27–29, 2024, doi: [available on IEEE Xplore Nov. 19, 2024].

[16] M. J. J. Ghrabat et al., "Utilizing machine learning for the early detection of coronary heart disease," Eng. Technol. Appl. Sci. Res., vol. 14, no. 5, pp. 17363–17375, Oct. 2024, doi: 10.48084/etasr.8171.