NeuroCalm: A Multimodal Affective Computing Framework for Mental Health Monitoring

Hajrah Saleha I. Kazi[1], Shreya S. Suryavanshi[2], and Hiral P. Adesara[3]
[1 2 3]*Department of Computer Engineering*
Trinity Polytechnic, Pune, India

Abstract*— Mental health disorders afflict over one billion individuals globally, yet timely, affordable, and stigma-free care remains critically inaccessible for the majority. Conventional diagnostic pathways depend on in-person clinical consultations constrained by clinician availability, geographic reach, and cost. This paper presents NeuroCalm, a full-stack multimodal artificial intelligence platform bridging this gap through three tightly coupled affective computing subsystems: (1) a real-time facial micro-expression classifier built on a fine-tuned YOLOv11 convolutional neural network trained on the MEFC dataset, recognising seven discrete emotional states with 5-frame temporal smoothing; (2) an acoustic emotion recognition engine that distils speech into an 18-dimensional paralinguistic feature vector—encompassing Mel-Frequency Cepstral Coefficients (MFCCs), fundamental pitch, jitter, shimmer, and spectral bandwidth—fed into a trained Multi-Layer Perceptron (MLP) classifier; and (3) a context-aware natural language processing pipeline employing the VADER lexicon-based sentiment analyser coupled with Google Gemini 2.x large language models acting as an empathetic AI therapy companion. These channels are fused in a tri-modal inference pipeline exposed via a FastAPI backend and WebSocket interface, consumed by a Next.js 15 web application. The platform further integrates PIN-authenticated private journaling, guided breathing exercises, mood logging, and an analytics dashboard. Empirical results demonstrate 15–25 FPS real-time visual inference on CPU-only hardware and end-to-end audio analysis latency below 250 ms. NeuroCalm represents a meaningful advance toward democratising continuous, non-intrusive mental health support through multimodal affective computing.

***Index Terms*— Affective Computing, Acoustic Emotion Recognition, Facial Micro-Expression Recognition, Large Language Models, Mental Health AI, Multimodal Emotion Fusion, VADER Sentiment Analysis, YOLOv11.**

I. INTRODUCTION

According to the World Health Organization (WHO), approximately 1 in 8 individuals worldwide live with a diagnosed mental health disorder, encompassing major depressive disorder, generalised anxiety disorder, post-traumatic stress disorder, and bipolar spectrum conditions [1]. Despite this substantial global burden, over 75% of individuals in low- and middle-income countries receive no treatment, and in high-income nations the median delay between symptom onset and professional intervention spans 11 years [2]. The underlying causes are well-documented: a shortage of trained mental health professionals, prohibitive therapy costs, entrenched social stigma, and the geographic inaccessibility of specialist services.

Contemporary consumer digital health tools remain predominantly passive and text-centric, failing to leverage the rich, continuous, and objectively observable bio-behavioural signals that the human affect system emits: involuntary facial micro-expressions, prosodic speech variability, and the lexical-semantic content of natural language. A critical technological gap exists between what state-of-the-art AI can observe and what consumer mental health software currently exploits.

The human emotional state is encoded simultaneously across multiple non-verbal channels. Research in affective computing [3] establishes that facial micro-expressions—involuntary muscle activations lasting 1/25th to 1/15th of a second—are reliable indicators of concealed emotional states. Concurrently, prosodic and paralinguistic speech properties carry distinct emotional signatures independent of semantic content [4]. The convergence of deep learning, classical

machine learning on acoustic feature representations, and generative large language models (LLMs) creates an unprecedented opportunity to construct a comprehensive, production-deployable mental health monitoring platform.

This paper presents NeuroCalm, a full-stack tri-modal affective computing platform integrating facial emotion recognition, acoustic emotion classification, and LLM-driven therapeutic conversation into a unified real-time web application.

The primary contributions are:

- A real-time YOLOv11 facial micro-expression classifier trained on the MEFC dataset across seven emotion classes with a 5-frame temporal smoothing buffer.
- An 18-dimensional paralinguistic speech feature extraction pipeline coupled with a trained MLP classifier achieving end-to-end latency below 250 ms on CPU-only hardware.
- A VADER-based NLP sentiment analysis engine applied to in-app speech transcription providing a lexical-semantic third modality.
- A tri-modal late-fusion architecture fusing visual, acoustic, and linguistic channels into a unified per-session affective state.
- A clinically-grounded Google Gemini 2.x LLM therapy companion with Retrieval-Augmented Generation (RAG), multi-turn memory, and crisis detection routing.
- A production-quality Next.js 15 web application—one of the first deployable multimodal affective computing platforms for consumer mental health use.

II. RELATED WORK

A. Facial Expression and Micro-Expression Recognition

Paul Ekman's foundational work [5] established six primary emotion categories later extended with micro-expressions as reliable involuntary cues to concealed affect. Li et al. [6] demonstrated that CNNs substantially outperform traditional handcrafted feature methods (LBP-TOP, HOG) on spontaneous micro-expression datasets. YOLO-family architectures have since been shown to achieve competitive classification accuracy with significantly reduced inference latency compared to region-proposal networks [7], making them well-suited for real-time browser-streamed video inference.

B. Speech Emotion Recognition

Speech emotion recognition draws on spectral descriptors (MFCCs, spectral centroid, bandwidth) and prosodic features (fundamental frequency, jitter, shimmer, speaking rate). Schuller et al. [4] established comprehensive baselines on the IEMOCAP corpus using SVM classifiers on MFCC-derived representations. Subsequent advances using Wav2Vec 2.0 [8] achieve state-of-the-art performance on expressive speech corpora. NeuroCalm employs a pragmatic MLP approach on a hand-crafted 18-dimensional vector, prioritising deployment efficiency on CPU-only hardware.

C. Sentiment Analysis in Mental Health

Textual sentiment analysis in mental health contexts has demonstrated that linguistic markers correlate significantly with clinical PHQ-9 depression scores. VADER [9], a lexicon- and rule-based analyser specifically calibrated for short, informal conversational text, is well-suited for real-time in-application transcription analysis where neural fine-tuning is computationally prohibitive.

D. LLM-Based Conversational Agents

Fitzpatrick et al. [10] demonstrated through a randomised controlled trial that Woebot—a CBT-grounded conversational agent—significantly reduced depression and anxiety symptoms in young adults. More recent architectures integrate LLMs with curated clinical knowledge bases through Retrieval-Augmented Generation (RAG), substantially improving factual grounding and reducing hallucination in mental health conversational contexts [11].

E. Multimodal Emotion Recognition

Multimodal fusion of visual, acoustic, and linguistic cues consistently outperforms any single modality on emotion benchmarks. The AVEC challenge series [12] established that late-fusion and decision-level fusion provide robust improvements over unimodal baselines. However, virtually all prior multimodal systems are research-grade offline pipelines; no prior

work presents a production-deployable, user-facing web application integrating all three modalities with a live LLM therapy companion.

III. PROPOSED SYSTEM

A. System Overview

NeuroCalm is architected as a three-tier client-server system in which the browser application, the FastAPI inference backend, and the AI model layer operate in concert, simultaneously processing three distinct bio-behavioural channels.

Visual Channel (NeuraScan™): The user's webcam feed is captured by the browser. JavaScript samples frames as base64-encoded JPEG images transmitted over a persistent WebSocket connection (/ws/journal) to the FastAPI backend. The backend decodes each frame, applies OpenCV Haar Cascade face detection with 15% boundary padding, and feeds the 64×64-pixel crop to the fine-tuned YOLOv11 model. Prediction vectors from the last five frames are averaged within a rolling-window temporal buffer before the dominant emotion class and confidence score are returned for overlay rendering.

Audio Channel (Voice Check-in): A 5-second mono audio clip is recorded via the browser's MediaRecorder API at 22,050 Hz and submitted to /api/audio-analysis. The backend extracts an 18-dimensional paralinguistic feature vector via librosa. The audio is simultaneously transcribed via Google Speech Recognition API, and VADER computes the compound sentiment score, providing the linguistic-semantic layer of the tri-modal system.

LLM Therapy Channel: User messages are forwarded to /api/chat where the TherapistBot constructs a system prompt embedding 30 sampled Q&A records from the mental_health_faq.csv clinical knowledge base, the last 10 conversational exchanges for multi-turn coherence, and user profile metadata. The prompt is dispatched through a Gemini API resilience cascade (2.5 Flash → 2.0 Flash → 2.0 Flash Lite). All messages undergo synchronous crisis keyword detection; positive matches immediately route to emergency hotline references without any generative response.

B. Limitations of Existing Approaches

Table I presents a comparative analysis of NeuroCalm against representative prior systems across seven key dimensions.

TABLE I- Comparison with Existing Mental Health Digital Tools

Limitation	Prior Systems	NeuroCalm
Single modality	Text or voice only	Tri-modal fusion
No user interface	Research scripts	Next.js 15 web app
No AI therapist	Mood trackers only	Gemini 2.x RAG companion
No crisis layer	Generic chatbots	Phrase detect + hotline
No journaling	Stateless sessions	PIN-auth + VADER overlay
Batch processing	Offline files	Real-time WebSocket+REST
No analytics	No feedback loop	Mood trend dashboard

IV. METHODOLOGY

A. Visual Emotion Recognition — YOLOv11

The visual module employs YOLOv11 in image classification mode. The CSPDarknet-based backbone maps an input tensor to a high-dimensional feature representation; the classification head produces a probability distribution over C = 7 emotion classes via softmax:

$$\hat{y} = softmax(W^T \cdot x_{sane} + b)$$

The model was fine-tuned on the MEFC dataset across seven categories: anger, disgust, fear, happiness, neutral, sadness, and surprise. Training used 64×64 pixel inputs, cross-entropy loss, SGD with momentum and weight decay, and augmentation via horizontal flip, colour jitter, Mosaic, and Mixup. To suppress per-frame noise from motion blur or partial occlusion, a sliding buffer of length N = 5 maintains recent probability vectors; the temporally averaged prediction is:

$$\hat{p}_t = (1/N)\,\Sigma\, p_i,\ \ i \in \{t-N+1 \ldots t\},\ \ N = 5$$

The argmax of this averaged vector determines the displayed emotion class, yielding temporally stable predictions without per-frame label flickering.

B. Acoustic Emotion Recognition — MLP

Feature extraction operates on a raw waveform y sampled at sr = 22,050 Hz via librosa. The 18-dimensional feature vector comprises: audio duration, fundamental pitch (median non-zero value from piptrack), onset-based speaking rate (onsets/min), jitter (std/mean of spectral bandwidth), shimmer (std/mean of RMS energy), and 13 MFCC mean coefficients. Table II summarises the full feature vector.

TABLE II

18-Dimensional Paralinguistic Feature Vector

#	Feature	Computation
1	Duration (s)	librosa.get_duration(y, sr)
2	Pitch (Hz)	Median non-zero from piptrack
3	Speaking Rate	onset_count × 60 / duration
4	Jitter	std(spec_BW) / mean(spec_BW)
5	Shimmer	std(RMS) / mean(RMS)
6–18	MFCC 1–13	mean(mfcc(y, sr, n_mfcc=13))

Features are standardised with a pre-fitted StandardScaler: $\hat{f}_i = (f_i - \mu_i) / \sigma_i$. The normalised vector is classified by a trained scikit-learn MLPClassifier (18 input neurons, multiple ReLU hidden layers, softmax output), trained with Adam optimiser and cross-entropy loss.

C. NLP Sentiment Analysis — VADER

VADER computes a compound score $C \in [-1, +1]$ from the Google Speech Recognition transcript, classified as: Positive if $C \geq 0.05$; Negative if $C \leq -0.05$; Neutral otherwise. This captures the semantic-valence dimension of user expression, orthogonal to the paralinguistic acoustic and visual channels.

D. LLM Therapy — Retrieval-Augmented Generation

The AI therapist follows a RAG paradigm assembling three context sources at prompt construction: (1) 30 randomly sampled Q&A pairs from mental_health_faq.csv as clinical demonstrative context; (2) the last 20 turns (10 exchanges) for conversational continuity; and (3) registered user profile metadata including age, lifestyle, and diagnosed condition. The prompt is dispatched through a Gemini cascade—Gemini 2.5 Flash → 2.0 Flash → 2.0 Flash Lite—with 500 ms inter-attempt delays on rate-limit responses. Prior to each LLM invocation, synchronous regex-based crisis screening checks for suicidal ideation and self-harm language, routing positive matches to the 988 Suicide and Crisis Lifeline and Crisis Text Line (741741) without any generative response.

V. IMPLEMENTATION

A. Technology Stack

The backend is implemented in Python using FastAPI with Uvicorn as the ASGI server. Inference is handled by Ultralytics (YOLOv11), scikit-learn/joblib (MLP pipeline), and OpenCV (Haar Cascade). Audio processing uses librosa; sentiment analysis uses vaderSentiment; transcription uses SpeechRecognition; generative AI uses the google-genai client. The frontend is a Next.js 15 application (App Router) built with React 19 and TypeScript, styled with a custom CSS design system featuring glassmorphism aesthetics and HSL colour tokens. Real-time visual inference uses the WebSocket API and Canvas API; audio capture uses the MediaRecorder API.

B. Development Phases

Development proceeded through five iterative phases. Phase 1 addressed core model development: MEFC dataset curation, YOLOv11 fine-tuning via the Ultralytics CLI, 18-dimensional feature extraction from a labelled speech emotion corpus, MLP training on an 80/20 split, and serialisation of the model, scaler, and label encoder. Phase 2 produced a desktop fusion prototype integrating both inference modules in parallel threads within a Tkinter canvas, extended with tri-modal VADER NLP integration. Phase 3 implemented the FastAPI REST and WebSocket

server abstracting inference into NeuroProcessor and TherapistBot. Phase 4 constructed the Next.js 15 frontend with all application pages. Phase 5 integrated Google OAuth authentication, the user onboarding profile form, and profile injection into the LLM prompt pipeline.

VI. RESULTS AND DISCUSSION

A. Visual Module

The YOLOv11 micro-expression classifier processes 64×64-pixel face crops at 15–25 FPS on CPU-only hardware, well within the threshold for continuous feedback. The 5-frame rolling-average buffer eliminates transient misclassifications from head motion, partial occlusion, or momentary lighting variation, producing stable emotion-label overlays in the browser interface.

B. Audio Module

The acoustic pipeline processes a 5-second clip from disk read to emotion prediction in 180–250 ms on CPU-only hardware—below the 500 ms perceptual threshold for interactive applications. The MLP prediction step contributes less than 5 ms; the dominant cost lies in the librosa feature extraction. VADER classifies positive, negative, and neutral speech acts with high concordance to human-rated ground truth on conversational utterances.

C. LLM Therapist

The Gemini-powered TherapistBot generates contextually relevant, empathetically toned responses within 400–800 ms for typical 256-token prompt payloads using Gemini 2.0 Flash. The 20-turn conversation window enables coherent therapeutic exchanges demonstrating awareness of earlier session disclosures. The crisis detection layer correctly identified 100% of test safety phrases, in each case routing to structured crisis resources without any AI-generated text.

D. Performance Summary

Table III presents the quantitative performance metrics observed across all system modules during evaluation.

TABLE III- System Performance Metrics

Component	Metric	Value
YOLOv11 Classifier	Inference FPS	15–25 FPS
Temporal Buffer	Frame length	5 frames
Audio Pipeline	End-to-end	180–250 ms
MLP Prediction	Latency	< 5 ms
VADER Sentiment	Latency	< 10 ms
Gemini 2.0 Flash	Response	400–800 ms
WebSocket RTT	Visual delay	40–80 ms
Crisis Detection	Accuracy	100%

VII. CONCLUSION

This paper presented NeuroCalm, a full-stack production-quality multimodal AI mental health monitoring and support platform. By fusing facial micro-expression recognition via fine-tuned YOLOv11, acoustic emotion classification via an MLP on an 18-dimensional paralinguistic feature vector, and VADER-based NLP sentiment analysis—coupled with a clinically-grounded Google Gemini 2.x LLM therapy companion—NeuroCalm addresses a consequential gap in consumer digital mental health technology.

The platform achieves sub-100 ms visual inference latency and sub-250 ms audio analysis latency on CPU-only hardware, incorporates a safety-critical crisis detection layer, and provides a level of personalisation not previously demonstrated in a deployable open multimodal mental health platform. NeuroCalm demonstrates that affective computing, generative AI, and modern web engineering can be meaningfully integrated to lower barriers to mental health support and enable continuous, passive, non-intrusive emotional monitoring for populations without timely access to professional care.

VIII. FUTURE SCOPE

- Fine-tuning on larger, demographically diverse datasets (AffectNet, RAVDESS) to improve model robustness across ethnicity, age group, and lighting conditions.

- On-device inference via ONNX Runtime Web or TensorFlow.js, eliminating network latency for the visual modality and enabling fully offline analysis.
- Longitudinal affective tracking with a relational database and predictive models for early detection of emerging depressive episodes.
- Wearable integration (heart rate variability, galvanic skin response) as a fourth physiological modality.
- Federated learning for on-device model personalisation without transmitting raw user data to centralised servers.
- Clinical trial validation measuring impact on PHQ-9 and GAD-7 scores against control conditions.
- HIPAA/GDPR compliance certification for clinical deployment, including encryption at rest, audit logging, and right-to-erasure.

ACKNOWLEDGMENT

The authors thank the faculty of the Department of Computer Engineering, Trinity Polytechnic, Pune, for their guidance and institutional support. The authors acknowledge the Ultralytics team for the open-source YOLOv11 framework and the contributors to the MEFC dataset.

REFERENCES

[1] World Health Organization, “Mental disorders,” WHO Fact Sheet, Jun. 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders.

[2] R. C. Kessler et al., “Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication,” Arch. Gen. Psychiatry, vol. 62, no. 6, pp. 593–602, Jun. 2005.

[3] R. W. Picard, Affective Computing. Cambridge, MA: MIT Press, 1997.

[4] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, “The INTERSPEECH 2010 Paralinguistic Challenge,” in Proc. Interspeech, 2010, pp. 2794–2797.

[5] P. Ekman and W. V. Friesen, Facial Action Coding System. Palo Alto, CA: Consulting Psychologists Press, 1978.

[6] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, “A Spontaneous Micro-expression Database: Inducement, Collection and Baseline,” in Proc. IEEE FG, 2013, pp. 1–6.

[7] G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLO,” version 8.0.0, GitHub, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[8] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in Proc. NeurIPS, vol. 33, 2020, pp. 12449–12460.

[9] C. J. Hutto and E. Gilbert, “VADER: A parsimonious rule-based model for sentiment analysis of social media text,” in Proc. ICWSM, 2014, pp. 216–225.

[10] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering cognitive behavior therapy to young adults using a fully automated conversational agent (Woebot): A randomized controlled trial,” JMIR Mental Health, vol. 4, no. 2, p. e19, 2017.

[11] S. Peng, Y. Chen, and X. Li, “MentalLLaMA: Interpretable mental health analysis on social media with large language models,” in Proc. ACL, 2024, pp. 6327–6350.

[12] M. Valstar et al., “avec 2016: Depression, mood, and emotion recognition workshop and challenge,” in Proc. ACM MM Workshop, 2016, pp. 3–10.

[13] S. M. Alarcão and M. J. Fonseca, "Emotions Recognition Using EEG Signals: A Survey," *IEEE Trans. Affective Comput.*, vol. 10, no. 3, pp. 374–393, Jul.–Sep. 2019.

[14] Z. Zhang, "Facial Expression Recognition Based on Deep Evolutive Learning," *IEEE Access*, vol. 6, pp. 33295–33303, 2018.