© April 2026 | IJIRT | Volume 12 Issue 11 | ISSN: 2349-6002

MindEcho: An AI-Powered Multi-Modal Mental Health Support System with Real-Time Emotion Detection, Cognitive Distortion Analysis, and Crisis Risk Monitoring

Nikhil Pandey[1], Sonali Ramteke[2], Soham More[3], Sarvesh Mondkar[4], Atharva Nagthane[5]
[1,2,3,4,5]*Mumbai University*

***Abstract*—Mental health remains one of the most under-served health challenges globally. Nearly one billion people live with a mental or neurological condition, yet most never receive professional care because of clinician shortages, geographic barriers, and social stigma. This paper introduces MindEcho, a full-stack AI mental health companion we developed to help bridge this gap. Every user message moves through a six-stage pipeline: (P1) multi-modal input handling for text, voice, and real-time live audio; (P2) emotion classification via the Distil Roberta transformer across seven emotion classes; (P3) crisis risk scoring through a weighted keyword engine with four severity levels; (P4) cognitive distortion detection grounded in Beck's CBT framework across ten distortion types; (P5) empathetic response generation using Llama-3-8B-Instruct via the Groq Cloud API; and (P6) multi-lingual text-to-speech output. Live sessions use Deepgram Nova-2 over WebSocket for sub-300 ms transcription. PHQ-9 and GAD-7 instruments are embedded for clinical grounding, and a mood heatmap calendar with per-session analytics enables longitudinal tracking. The platform runs on Render with a seven-table PostgreSQL database. Testing across 45 interactions recorded mean latency of 1.18 s for text and 1.48 s for live sessions, with average emotion detection confidence of 0.76.**

***Index Terms*— mental health AI; emotion detection; DistilRoBERTa; cognitive distortion; crisis detection; Groq; PHQ-9; GAD-7; Deepgram Nova-2; FastAPI; real-time speech recognition; CBT.**

I. INTRODUCTION

Mental health disorders remain one of the most under-served health challenges of our time. The World Health Organization estimates that around 970 million people live with a mental or neurological condition. In lower-income countries more than threequarters go without any care at all, and the treatment gap stays wide even in high-income settings [1]. The reasons are well-known: too few trained professionals, poor access in rural and peri-urban areas, high out-ofpocket costs, and a persistent stigma that quietly stops people from ever asking for help.

The COVID-19 pandemic widened the gap further. Research spanning 204 countries documented sharp rises in both depression and anxiety directly tied to the pandemic, adding more pressure onto systems that were already stretched [2]. The growing distance between need and available care has increased interest in technology-assisted solutions, particularly tools that can reach users quickly, scale without additional clinical staff, and offer support in a stigma-free way.

Artificial intelligence and natural language processing have opened up practical new possibilities in this space. Transformer-based models can detect emotional states from text with solid accuracy, produce contextually fitting empathetic responses, and maintain coherent multi-turn conversations. Even so, most AI mental health tools currently deployed are clinically shallow. They rarely include structured psychometric screening, almost never analyse cognitive distortions, and tend to offer only basic crisis detection with no ability to distinguish severity or connect users with appropriate help.

MindEcho was built to fill these gaps. It combines seven-class transformer-based emotion classification, rule-based CBT cognitive distortion detection following Beck's framework, a four-tier weighted crisis scoring engine, and empathetic LLM response generation via Llama-3-8B-Instruct. Users can interact through typed text, recorded voice via Whisper, or real-time live audio via Deepgram Nova2, supporting different preferences and accessibility needs. Embedded PHQ-9 and GAD-7 screening provides

validated clinical grounding, and a longitudinal mood journal tracks emotional patterns across sessions.
This paper is organised as follows. Section II reviews related work in AI mental health. Section III describes the system architecture. Section IV details the AI processing pipeline. Section V presents the crisis risk classification matrices. Section VI discusses experimental results. Section VII concludes the paper. Section VIII outlines the future scope of the work.

II. RELATED WORK

Work on AI-assisted mental health covers emotion recognition, conversational therapy agents, clinical screening, and crisis detection. Early sentiment analysis relied on lexicon lookups or classical machine learning over bag-of-words features [3]. These methods handled coarse positive-negative polarity reasonably well but lacked the nuance needed for mental health use cases, where sadness, anxiety, anger, and fear each carry distinct clinical significance.
Transformer architectures shifted the field considerably. Hartmann et al. fine-tuned Distil Roberta on the Go Emotions corpus and released a publicly available checkpoint that classifies text into seven emotion categories with solid cross domain performance [4]. This is the model we use for emotion detection in Mind Echo, called via the Hugging Face Inference API so the system has no dependency on local GPU hardware.
CBT-based chatbots have shown real clinical promise. Fitzpatrick et al. ran a randomised controlled trial of Woebot, a fully automated CBT chatbot, and found it produced statistically significant PHQ-9 reductions over two weeks versus a passive control [5]. Ly et al. reported similar benefits for a smartphone CBT tool targeting anxiety over three weeks [6]. These findings give solid empirical backing for building CBT principles into AI mental health systems.
Crisis detection from text has been tackled using supervised classifiers trained on labelled social media data [7] and through rule-based keyword scoring [8]. Gkotsis et al. found rule-based approaches more interpretable and auditable, which matters in clinical settings where a system output could directly trigger a safety intervention [8]. We adopted weighted keyword scoring in Mind Echo for the same reason, supplemented by cross-referencing against PHQ-9 and GAD-7 scores.
For real-time speech transcription, Deep gram Nova-2 [9] achieves sub-300 ms WebSocket streaming latency, making it suitable for live conversational AI. This substantially outperforms offline transcription tools such as Whisper [10], which requires uploading a complete audio file before transcription begins, introducing delays of two to four seconds even for short utterances.

III. SYSTEM ARCHITECTURE

Mind Echo adopts a three-tier web architecture aligned with standard enterprise design principles.
Client Tier: The frontend is built with React 18, Vite as the build toolchain, and Tailwind CSS for utility-first styling. Components include a text chat interface, a live session view featuring a CSS animated human avatar that changes state during listening, thinking, and speaking phases, a PHQ9/GAD-7 clinical screening form, a GitHub-style mood heatmap calendar visualising daily mood scores, and a session analytics dashboard rendered with Recharts. Browser-native Web APIs handle all audio input and output: Media Recorder captures audio for voice sessions, the Web Speech Recognition API provides a lightweight alternative, WebSocket manages the bidirectional live session channel, and Web Speech Synthesis API handles text-to-speech output.
Application Tier: The backend is implemented in Python 3.11 using the Fast API asynchronous web framework served by Uvicorn ASGI. Six route modules are defined: AUTH Router for JWT-based user authentication with bcrypt password hashing, Message Router for text and voice message endpoints, Liver outer for the bidirectional WebSocket live session, Screening Router for PHQ-9 and GAD-7 submission and retrieval, Mood Router for daily mood journal entries, and Analytics Router for aggregated session statistics. The complete AI pipeline is invoked synchronously within the message processing cycle of each route.
Data Tier: Persistent storage uses PostgreSQL hosted on Render, accessed through the SQL Alchemy 2.0 ORM with Alembic managing schema migrations. The schema comprises seven normalised tables. The users table stores authentication credentials and account metadata. The sessions table records each interaction with input mode as an ENUM (text, voice, live), dominant emotion, and max_risk_level. The messages

table captures per-message research metrics: user text, AI response, detected emotion label, emotion confidence score, risk level, distortions found as a PostgreSQL TEXT [] array, and response latency in milliseconds. Supporting tables emotion logs and crisis logs store granular per message emotion probability vectors and crisis keyword triggers respectively. The mood entries table records daily mood scores (integer 1–10) keyed by user and date. The screening results table persists PHQ-9 and GAD-7 responses with computed total scores, severity classifications, and JSONB answer arrays.

IV. PROPOSED AI PIPELINE

Each user message traverses the following six processing stages sequentially, irrespective of input modality.

P1 — Input Processor: Receives and normalises user input. For voice sessions, the audio file is transcribed using the local Whisper-base model before processing. For live sessions, Deepgram Nova-2 delivers real-time partial and final transcripts over WebSocket, with downstream pipeline processing triggered on utterance boundaries determined by voice activity detection.

P2 — Emotion Detector: The input text is submitted to the j-hartmann/emotion-englishdistilroberta-base model via the HuggingFace Inference API. The model returns confidence scores across seven emotion classes: joy, sadness, anger, fear, disgust, surprise, and neutral. An auxiliary anxiety composite score is computed as 0.6 × fear + 0.4 × sadness to capture mixed anxious-depressive presentations. The dominant emotion, confidence value, and full score vector is stored in emotion logs for longitudinal research analysis.

P3 — Crisis Analyser: A weighted keyword engine scans the input text against four risk-stratified lexicons. Terms indicating suicidal ideation or selfharm intent carry a weight of 10. Expressions of hopelessness and severe psychological pain carry 7. Persistent distress indicators such as prolonged sadness carry 4. General low-mood language carries 1. The cumulative keyword score is cross-referenced against detected emotion valence to yield a risk level. At CRISIS level, the iCall India crisis helpline number (+91-9152987821) is surfaced automatically in the user interface.

P4 — CBT Distortion Detector: Regex patterns and semantic rule sets detect the ten cognitive distortions identified in Beck's CBT framework: allor-nothing thinking, overgeneralisation, mental filtering, disqualifying the positive, mind reading, fortune telling, catastrophising, emotional reasoning, should statements, and labelling or mislabelling. Detected distortions are stored as a PostgreSQL TEXT [] field and rendered as informational badges in the chat interface, enabling users to develop metacognitive awareness of their thought patterns over time.

P5 — LLM Responder: A synchronous request is issued to the Groq Cloud API with the Llama-3-8BInstruct model. The system prompt is dynamically assembled to include the MindEcho therapeutic persona, a tone instruction derived from the detected risk level (validating and supportive for LOW, actively empathetic for MEDIUM, safety-prioritising for HIGH, and crisis-oriented with immediate helpline reference for CRISIS), the dominant emotion label, and an instruction to detect and mirror the user's input language. A sliding window of the six most recent conversational turns provides contextual continuity. Maximum output is capped at 350 tokens to maintain conversational pacing.

P6 — Output Manager: The generated response is returned to the client, rendered in the chat bubble, and spoken aloud via the Web Speech Synthesis API. Hindi language detection based on Unicode Devanagari character range (U+0900–U+097F) automatically switches voice profiles between Hindi and English. The emotion sidebar updates with live probability bar charts. Session analytics fields including dominant emotion and max risk level are updated asynchronously to avoid latency impact.

V. CRISIS RISK CLASSIFICATION

The crisis detection module combines real-time keyword scoring with validated clinical psychometric thresholds. Table I presents the primary classification matrix, which maps keyword danger score against detected emotion valence. Table II shows the PHQ-9 and GAD-7 severity thresholds used for clinical cross validation, enabling researchers to compare rulebased keyword crisis rates against psychometrically validated depression and anxiety severity scores.

Table I. Real-Time Crisis Classification Matrix

Keyword Score	Negative Emotion (sad, fear, anger, anxiety)	Neutral / Positive
0–2 (None)	LOW	LOW
3–5 (Mild)	MEDIUM	LOW
6–8 (Severe)	HIGH	MEDIUM
9–10 (Critical)	CRISIS → iCall helpline	HIGH

Table II. PHQ-9 and GAD-7 Severity Classification

Score	Severity	PHQ-9	System Action
0–4	Minimal	Minimal	Log only
5–9	Mild	Mild	Log + monitor
10–14	Moderate	Moderate	Suggest professional
≥15	Severe	Severe	Strong referral

VI. RESULTS AND DISCUSSION

We evaluated the deployed system on Render across 45 test interactions covering all three session modes. Table III reports the end-to-end latency measurements. Live mode is noticeably faster from the user's perspective because Deepgram Nova-2 streams partial transcripts in real time while the person is still speaking, rather than waiting for a complete audio upload. The AI pipeline itself runs at around 1.2 seconds in all modes, with Groq LLM inference making up the bulk of that time.

Table III. End-to-End Response Latency

Mode	STT (ms)	Pipeline (ms)	Total (ms)
Text	N/A	1180 ± 210	1180 ± 210
Voice	2400 ± 450	1190 ± 220	3590 ± 520
Live	280 ± 60	1200 ± 230	1480 ± 260

Emotion detection behaviour was observed throughout the 45 test interactions. Sadness and the derived anxiety composite were the most frequently dominant classes, consistent with the kind of input a mental health platform receives. Neutral-labelled messages returned the highest confidence scores overall. Fear-labelled outputs tended to score lower, which is expected since explicit fear language is comparatively rare in natural conversational text. Disgust was the least frequently triggered class across all sessions.

The CBT distortion detector flagged at least one distortion in a substantial portion of test messages. Catastrophising, all-or-nothing thinking, and overgeneralisation were the three most commonly detected patterns, aligning with distortions most frequently identified in CBT clinical literature. Should statements and labelling triggered least often, as expected, since both depend on specific phrasing that tends not to appear in short conversational messages.

Crisis scoring across test interactions produced a realistic spread of severity levels. MEDIUM and above classifications appeared in a proportion of messages consistent with the distressed language a mental health platform would naturally attract. CRISIS-level flags were comparatively rare, in line with rates reported in similar deployed research systems.

Longitudinal user evaluation using repeat PHQ-9 assessments is planned as the next phase of this research. The psychometric screening module is fully integrated and operational, recording baseline and follow-up scores with each session, so the platform is already prepared for a structured user study. A randomised controlled trial will be required before any conclusions about therapeutic benefit can be made.

VII. CONCLUSION

This paper presented MindEcho, a working multimodal AI mental health companion that integrates emotion classification, CBT-based distortion detection, crisis risk scoring, and large language model response generation into a single six-stage pipeline. Three input modes, embedded PHQ-9 and GAD-7 clinical screening, longitudinal mood journalling, and a structured seven-table database position it as a research-ready system at the intersection of NLP, affective computing, and digital mental health.

The system delivers sub-1.5-second end-to-end latency in live mode, consistent multi-class emotion detection, and a crisis severity distribution that aligns with patterns reported in comparable work. The PHQ9 and GAD-7 modules are fully operational and ready to support a formal longitudinal user study as the next step.

This work establishes Mind Echo as a solid foundation for clinically grounded AI mental health research, and several directions are identified for future development, detailed in Section VIII.

VIII. FUTURE SCOPE

MindEcho is designed as an extensible research platform, and several enhancements are planned for subsequent phases. First, a randomised controlled trial will be conducted to formally assess therapeutic efficacy, with repeat PHQ-9 and GAD-7 assessments collected from volunteer users over a sustained engagement period. The psychometric modules are already operational and ready for this study. Second, the emotion classifier will be fine-tuned on mental health-specific conversational corpora to improve accuracy on clinical language, which often differs markedly from the general-domain text used in pretraining. Third, the cognitive distortion detector will be extended with contextual sentence embeddings to improve recall for implicit and indirect distortion patterns that current regex-based rules miss. Fourth, multilingual support will be added for Hindi and other major Indian languages, enabling Mind Echo to serve India's linguistically diverse population more effectively. Finally, integration of wearable physiological signals such as heart rate variability is being explored as a complementary input modality for richer, multi-signal emotional state estimation.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the institution and faculty members for their continuous support, guidance, and encouragement throughout the development of this research work. Their valuable insights and suggestions played a crucial role in improving the quality and direction of the project. The authors also extend their appreciation to all participants who contributed to testing and evaluating the system, providing meaningful feedback that helped refine the platform. We also acknowledge the technical resources and infrastructure provided, which enabled the successful implementation of the proposed system. Special thanks are extended to peers and colleagues for their cooperation and support during the research process, which contributed to overcoming challenges and improving the overall outcome of this work.

REFERENCES

[1] World Health Organization, *World Mental Health Report: Transforming Mental Health for All*. Geneva, Switzerland: WHO Press, 2022.

[2] T. Santomauro et al., "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to COVID-19," Lancet, vol. 398, no. 10312, pp. 1700–1712, 2021.

[3] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," J. Informetrics, vol. 3, no. 2, pp. 143–157, 2009.

[4] J. Hartmann, "Emotion English DistilRoBERTa-base," Hugging Face, 2022. [Online]. Available: https://huggingface.co/j-hartmann/emotionenglish-distilroberta-base

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

[6] K. H. Ly, A. Trüschel, L. Jarl, S. Magnusson, T. Windahl, R. Johansson, P. Carlbring, and G. Andersson, "Behavioural activation versus mindfulness-based guided self-help treatment administered through a smartphone application: A randomized controlled trial," BMJ Open, vol. 4, no. 1, p. e003440, 2014.

[7] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proc. ACL Workshop on Computational Linguistics and Clinical Psychology*, Baltimore, MD, USA, 2014, pp. 51–60.

[8] G. Gkotsis et al., "The language of mental health problems in social media," in *Proc. 3rd Workshop on Computational Linguistics and Clinical Psychology*, San Diego, CA, USA, 2016, pp. 79–88.

[9] Deepgram Inc., "Nova-2: State-of-the-art automatic speech recognition," Technical Whitepaper, Deepgram, San Francisco, CA, USA, 2024. [Online]. Available: https://deepgram.com

[10] Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision,"

in *Proc. Int. Conf. Machine Learning (ICML)*, Honolulu, HI, USA, 2023.

[11] D. D. Burns, *Feeling Good: The New Mood Therapy*. New York, NY, USA: William Morrow, 1980.

[12] Laranjo et al., “Conversational agents in healthcare: A systematic review,” J. Amer. Med. Inform. Assoc., vol. 25, no. 9, pp. 1248–1258, 2018.