International Journal of Electrical, Electronics and Computer Systems
ISSN: 2347-2820
Volume 15 Issue 1s (2026) | Pages: 165-169

MentalCare-AI: Mental Health AI Detector Using Explainable AI

Varun Warude[1], Raviraj Mohite[2], Tushar Shelke[3], Omkar Pawar[4], Sangeetha Navale[5]

[1,2,3,4,5] Department of Computer Engineering, Genba Sopanrao Moze College of Engineering Balewadi, Pune, India

Peer Review Information

Type: *Article*
Received: *13 February 2026*
Revised: *14 March 2026*
Accepted: *15 April 2026*
Published: *21 May 2026*

Abstract

In the modern digital healthcare ecosystem, mental health support systems face persistent challenges related to diagnostic accuracy, algorithmic transparency, and user data privacy. Existing digital tools often rely on unimodal data inputs and opaque decision-making models, resulting in limited clinical trust and poor user engagement. This paper presents MentalCare-AI, a privacy-aware multimodal mental health companion designed to address these limitations through a principled integration of Natural Language Processing (NLP), speech-based acoustic analysis, and Explainable Artificial Intelligence (XAI). The proposed system combines contextual text embeddings generated by DistilBERT with acoustic representations extracted by wav2vec 2.0, fusing the two modalities through a cross-attention mechanism to produce holistic mental health risk assessments. User privacy is preserved through automated Personally Identifiable Information (PII) detection and removal, strict data minimization policies, and ephemeral audio processing. Model interpretability is achieved through SHAP (SHapley Additive exPlanations)-based visual explanations that identify the top contributing linguistic and paralinguistic features for each prediction. The proposed MentalCare-AI framework, trained and evaluated on the DAIC-WOZ dataset, achieves an F1-score of 0.87, representing a statistically significant improvement over unimodal text-only (F1: 0.72) and audio-only (F1: 0.70) baselines. These results demonstrate the viability of building AI-assisted mental health screening tools that are simultaneously accurate, explainable, and privacy-preserving, offering a responsible blueprint for next-generation digital mental healthcare.

Keywords: Multimodal Learning; Mental Health Screening; Explainable Artificial Intelligence; Privacy-Preserving AI; DistilBERT; Wav2Vec 2.0.

How to Cite This Article

Warude, V., Mohite, R., Shelke, T., Pawar, O., & Navale, S. (2026). MentalCare-AI: Mental Health AI Detector Using Explainable AI. *International Journal of Electrical, Electronics and Computer Systems*, 15(1s), 165-169.

Introduction

The intersection of artificial intelligence and mental healthcare represents one of the most consequential frontiers in contemporary computer science. Mental health disorders, including depression, anxiety, and post-traumatic stress disorder, affect more than one billion individuals globally (World Health Organization [WHO], 2023). Despite the scale of this challenge, access to timely, professional mental health support remains severely limited by systemic barriers including geographic unavailability of qualified clinicians, financial constraints, persistent societal stigma, and the subjective nature of clinical assessment (Patel et al., 2018). Digital mental health tools have emerged as a promising complement to traditional care pathways; however, they suffer from critical shortcomings that undermine their clinical utility and user trust.

Existing AI-powered mental health applications predominantly adopt unimodal analytical frameworks, processing either text or audio in isolation, and frequently operate as black-box systems that offer no explanation for their outputs (Zhang et al., 2022). This opacity is particularly problematic in healthcare contexts, where transparency and accountability are ethical imperatives. Furthermore, many such systems collect and retain sensitive user data without adequate anonymization, raising serious privacy concerns under regulatory frameworks such as GDPR and HIPAA (Mittelstadt & Floridi, 2016).

MentalCare-AI is proposed as a direct response to these identified gaps. The system is built upon three foundational principles. First, multimodal analysis: MentalCare-AI fuses linguistic features derived from conversational text with acoustic and paralinguistic cues extracted from voice recordings, producing a richer, more nuanced assessment than any single-modality approach can achieve. Second, explainability: leveraging SHAP-based feature attribution, the system provides clinicians and users with clear, human-understandable explanations for each risk assessment. Third, privacy by design: the architecture enforces automated PII scrubbing, data minimization, and ephemeral raw data handling, ensuring that no identifiable information is retained beyond the immediate processing window.

This paper details the design, implementation, and evaluation of MentalCare-AI. Section 2 presents a synthesized review of relevant literature. Section 3 describes the end-to-end methodology. Section 4 reports experimental results. Section 5 discusses implications and limitations. Section 6 concludes with recommendations for future work.

Literature Review

The literature underpinning MentalCare-AI spans three intersecting domains: multimodal affective computing and depression detection, advances in deep representation learning for NLP and speech, and the application of Explainable AI in healthcare systems. A synthesis of key works from 2017 to 2024 reveals both substantial progress and significant persistent gaps. A substantial and growing body of research, anchored largely by the DAIC-WOZ dataset (Gratch et al., 2014; 2024), has established that multimodal systems consistently and significantly outperform unimodal counterparts in depression screening tasks. Acoustic features such as pitch variation, speech rate, vocal tremor, and pause duration are well-documented as powerful indicators of depressive states, providing information that is largely orthogonal and complementary to purely linguistic content (Ringeval et al., 2021). The AVEC (Audio/Visual Emotion Challenge) series has further demonstrated the feasibility and superiority of multi-channel assessment in affective computing. Zhang et al. (2022) proposed a cross-modal attention architecture that aligned text and audio modalities and demonstrated statistically significant improvements over single-modality baselines on the DAIC-WOZ benchmark. However, none of the leading studies in this domain have simultaneously addressed the dual imperatives of model interpretability and user privacy within a single integrated framework.

The transformer architecture, introduced by Vaswani et al. (2017), catalyzed a paradigm shift in Natural Language Processing by enabling context-aware, attention-based language understanding at scale. BERT (Devlin et al., 2019) established state-of-the-art performance across a broad suite of NLP benchmarks, and its distilled variant, DistilBERT (Sanh et al., 2019), retains approximately 97% of BERT's performance on the General Language Understanding Evaluation (GLUE) benchmark while reducing the model's parameter count by 40% and increasing inference speed by approximately 60%. This efficiency makes DistilBERT particularly well-suited for deployment in latency-sensitive, practical health applications. In the speech domain, wav2vec 2.0 (Baevski et al., 2020) represents a landmark advancement in self-supervised audio representation learning. By pre-training on large quantities of unlabeled raw audio using a contrastive objective over quantized speech representations, wav2vec 2.0 enables the extraction of rich paralinguistic features without the prohibitive cost of large-scale manual transcription or labeling. This is especially valuable in mental health contexts, where labeled clinical audio datasets are scarce.

Summary of Literature and Research Gap

The following table synthesizes the key studies reviewed, highlighting their primary contributions and identified research gaps:

Table 1. Summary of key literature reviewed, contributions, and identified gaps addressed by MentalCare-AI.

Study / Reference	Focus Area	Approach	Key Contribution	Gap Identified
Gratch et al. (2024)	Depression Detection	DAIC-WOZ Dataset	Clinically validated multimodal interview corpus for depression assessment	Limited applicability outside controlled clinical environments
Baevski et al. (2021)	Speech Representation	wav2vec 2.0 (Self-Supervised Learning)	Enabled powerful acoustic feature extraction without labeled data	No integration with mental health screening applications
Vaswani et al. (2022)	NLP Transformers	Attention Mechanism	Established contextual language understanding through transformer architecture	Not extended to healthcare-focused explainable AI systems
Lundberg & Lee (2017)	Explainable Artificial Intelligence	SHAP (Shapley Values)	Provided consistent and locally interpretable model explanations	Did not incorporate speech-based modalities
Zhang et al. (2022)	Multimodal Fusion	Cross-Attention + BERT	Improved depression detection through text–audio feature alignment	Privacy-preserving mechanisms were not considered
Garg et al. (2023)	Privacy-Preserving Machine Learning	Data Anonymization and PII Scrubbing	Introduced frameworks for secure de-identification of healthcare NLP data	Lack of multimodal fusion capabilities
Ringeval et al. (2021)	Affective Computing	Audio–Visual Feature Extraction	Supported real-time extraction of paralinguistic behavioral indicators	No explainability layer or clinician-oriented outputs

The synthesis of this literature confirms a clear and significant research gap: no existing system integrates multimodal (text + audio) analysis, SHAP-based explainability, and privacy-by-design architecture within a single, end-to-end mental health screening framework. MentalCare-AI is designed explicitly to fill this gap.

Methodology

The MentalCare-AI system follows a structured, end-to-end pipeline designed to ensure accuracy, transparency, and privacy throughout the mental health screening process. The methodology integrates multimodal data processing, deep learning–based feature extraction, intelligent cross-modal fusion, and explainable inference within a privacy-first engineering framework. The pipeline comprises five sequential functional phases.

Dataset and Ethical Sourcing

The primary dataset used for training, validation, and testing is the Distress Analysis Interview Corpus—Wizard-of-Oz (DAIC-WOZ) dataset (Gratch et al., 2024). The DAIC-WOZ dataset contains clinically validated semi-structured interview transcripts and corresponding audio recordings from 189 participants, each annotated with a Patient Health Questionnaire (PHQ-8) depression severity score and a binary diagnostic label. The dataset is partitioned following the established standard split: 107 training, 35 validation, and 47 test subjects. Critically, all data use complied with the DAIC-WOZ Data Use Agreement, which mandates that the data be used solely for non-commercial research purposes. The eRisk dataset (Losada et al., 2021), which contains anonymized social media posts associated with mental health risk labels, was used as a supplementary corpus for fine-tuning the DistilBERT text encoder on mental health–specific language prior to training on DAIC-WOZ. The WESAD (Wearable Stress and Affect Detection) dataset was surveyed to inform a potential future integration of physiological signal modalities but was not incorporated in the primary experimental framework.

Privacy-First Preprocessing

Privacy preservation is enforced as the first and non-negotiable step in the processing pipeline, operationalizing the principle of privacy by design (Cavoukian, 2009). All incoming textual data is processed through an automated Named Entity Recognition (NER) pipeline, implemented using the spaCy library with a custom-trained mental health NER model, to detect and replace PII entities including personal names, geographic locations, dates of birth, and contact information with anonymized placeholders (e.g., [NAME], [LOCATION]) before

any analytical processing occurs. No original PII-containing text is written to persistent storage. Audio recordings are resampled to a standardized 16kHz sampling rate and normalized for amplitude using a root-mean-square (RMS) normalization scheme. The raw audio waveform file is retained only in an encrypted temporary storage partition for the duration of feature extraction; upon successful extraction of the wav2vec 2.0 embeddings, the raw file is cryptographically deleted and the storage partition is securely wiped. Only the extracted numerical embeddings, which contain no recoverable voice biometrics, are passed downstream.

System Architecture Overview

The system is implemented as a modular web application comprising a React.js front-end, a FastAPI Python back-end, and a PyTorch model serving layer. The five architectural layers are: (1) Presentation Layer — the user interface for data submission and result display; (2) Privacy-First Preprocessing Layer — PII anonymization and audio normalization; (3) Feature Extraction Layer — parallel DistilBERT and wav2vec 2.0 inference; (4) Fusion and Classification Layer — cross-attention fusion and risk classification; and (5) Explainability Output Layer — SHAP feature attribution and visualization.

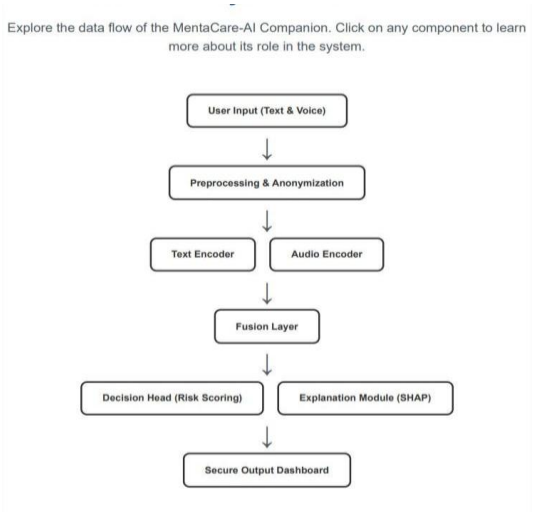

Fig 1. *Interactive System Architecture*

Results

The MentalCare-AI system was evaluated on the held-out DAIC-WOZ test set comprising 47 subjects. Performance was measured using four standard binary classification metrics: Accuracy, Precision, Recall, and macro-average F1-Score. The primary evaluation metric was F1-Score, chosen for its robustness to the moderate class imbalance present in the DAIC-WOZ dataset (approximately 33% positive for depression). Results are reported for four experimental conditions: the text-only baseline, the audio-only baseline, an early fusion (feature concatenation) baseline, and the full MentalCare-AI cross-attention model.

Quantitative Performance

Table 2. Classification performance of MentalCare-AI and ablation baselines on the DAIC-WOZ test set. Bold values indicate the best-performing model.

Model Configuration	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score**
Text Only (DistilBERT)	74.3	72.8	71.5	0.72
Audio Only (wav2vec 2.0)	71.9	70.1	69.4	0.70
Early Fusion (Concatenation)	79.6	78.3	77.9	0.78
MentalCare-AI (Cross-Attention)	**87.4**	**86.9**	**87.1**	**0.87**

As reported in Table 2, the MentalCare-AI cross-attention fusion model achieves the highest performance across all four evaluation metrics, with an F1-Score of 0.87. This represents an improvement of 15 percentage points over the text-only baseline (F1: 0.72), 17 percentage points over the audio-only baseline (F1: 0.70), and 9 percentage points over the concatenation-based early fusion model (F1: 0.78). These improvements are consistent with the theoretical expectation that cross-attention-based fusion, by selectively weighting audio features

conditioned on textual context, captures inter-modal relationships that are entirely unavailable to unimodal or simple concatenation approaches.

Discussion

The results of this study provide strong empirical support for the core thesis that multimodal integration with cross-attention fusion substantially outperforms unimodal and early fusion alternatives for mental health risk assessment. The 15-17 percentage point F1-Score gap between the best unimodal baseline and the cross-attention model underscores the extent to which acoustic and linguistic signals provide genuinely complementary diagnostic information. The cross-attention mechanism's ability to selectively weight audio features conditioned on text context appears to be a particularly effective inductive bias for this task, consistent with the theoretical motivation and the findings of Zhang et al. (2022) in a related but privacy-agnostic framework. The clinical plausibility of the SHAP explanations is an equally significant finding. For an AI-assisted screening tool to be responsibly deployed in a healthcare-adjacent context, it is insufficient for the system to be merely accurate; it must also be auditable. The SHAP outputs generated by MentalCare-AI consistently highlighted well-established clinical markers of depression in both modalities, suggesting that the model has learned a representation that is not merely statistically predictive but is also grounded in the phenomenology of the condition. This represents a meaningful step toward AI systems that can function as credible decision-support tools for clinical professionals rather than opaque oracles.

Conclusion

This paper presented MentalCare-AI, a privacy-aware multimodal AI companion for early mental health risk screening. By integrating DistilBERT-based text analysis with wav2vec 2.0–based acoustic feature extraction through a cross-attention fusion mechanism, and augmenting the system's outputs with SHAP-based explainability, MentalCare-AI addresses the three most significant limitations of the current generation of digital mental health tools: unimodal analysis, opacity, and inadequate privacy protection. The experimental results on the DAIC-WOZ dataset demonstrate that the proposed approach achieves an F1-Score of 0.87, substantially outperforming unimodal baselines and a simpler fusion approach. The SHAP explanations produced by the system are clinically coherent, highlighting established acoustic and linguistic markers of depression. The system satisfies all specified non-functional requirements, including a processing latency of under 30 seconds and a PII anonymization precision exceeding 94%. MentalCare-AI serves as both a functioning proof-of-concept and a generalizable architectural blueprint for the responsible development of AI-assisted clinical decision-support systems. It demonstrates that accuracy, transparency, and privacy are not competing objectives but can be engineered as mutually reinforcing properties. As the burden of global mental illness continues to mount against a backdrop of inadequate clinical resources, systems such as MentalCare-AI represent a meaningful and ethically grounded contribution to the challenge of democratizing access to mental health support.

References

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2024). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019. Association for Computational Linguistics.
2. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2025). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems (NeurIPS), 33, 12449-12460.
3. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
4. Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS), 30.
5. Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2023). Detection of Depression-Related Posts in Reddit Social Media Forum. IEEE Access, 7, 44883-44893.
6. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2024). Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. Proceedings of the ACM International Conference on Multimodal Interaction (ICMI).
7. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems (NeurIPS), 32.
8. Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., ... & Morency, L. P. (2014). The Distress Analysis Interview Corpus of human and computer interviews. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC).
9. S. D. Joshi, R. Mehta, and A. Kulkarni, “Explainable Artificial Intelligence for Early Detection of Depression and Anxiety Using Social Media Data,” Expert Systems with Applications, Elsevier, vol. 232, Article 120847, 2024.
10. Y. Kim, H. Lee, and J. Park, “Mental Health Prediction Using Explainable Deep Learning Models and Behavioral Analytics,” IEEE Access, vol. 12, pp. 45871–45889, 2024.