Multi-Modal Deep Learning Framework for Real-Time Depression Detection Using Clinical Assessment, Natural Language Processing, and Audio Sentiment Analysis

Dhanushi[1], Peeyush[1], Aarushi[1], and Himanshu[1]

Department of Computer Science and Engineering,
Netaji Subhas University of Technology (NSUT), Delhi, India
{dhanushi.ug23, peeyush.gupta.ug23, aarushi.ug23}@nsut.ac.in

Abstract. Major Depressive Disorder (MDD) affects over 280 million individuals worldwide, contributing to approximately $1 trillion in annual economic losses, with only 12% of affected individuals seeking professional support. Traditional clinical diagnosis methods depend on subjective interviews and face challenges such as limited accessibility, inter-rater variability (Cohen's kappa 0.60–0.75), social stigma, and significant financial cost ($1,000–5,000 per consultation in developing countries).

This work introduces a novel multi-modal artificial intelligence framework integrating three complementary modalities: (1) PHQ-9 clinical assessment using XGBoost, (2) transformer-based natural language processing with fine-tuned RoBERTa (125M parameters) and DistilBERT (66M parameters), and (3) audio sentiment analysis with 40-dimensional acoustic features. The proposed late-fusion ensemble employs learned weighted averaging optimized using exhaustive grid search.

Evaluated on a curated dataset of 321,600 labeled samples spanning 73 countries, the framework achieves state-of-the-art results: 94.8% accuracy (95% CI: 94.5–95.1%), 0.956 precision, 0.940 recall, 0.948 F1-score, and 0.968 ROC-AUC. Ablation studies confirm statistically significant improvements of 2.7–7.5% over dual-modality models ($p < 0.01$). Cross-dataset generalization on DAIC-WOZ yields 89.4% accuracy, suggesting robust performance with minimal overfitting.

Fairness evaluation shows equitable performance across demographic subgroups (maximum disparity 1.3%), satisfying equalized odds. Real-time web deployment achieves 1.23-second inference with a 533MB model footprint and 2.1GB runtime memory, scaling to 70,300+ daily assessments on a single GPU.

Explainability via SHAP analysis highlights depression-related keywords such as "hopeless" (+0.42) and "suicidal" (+0.38). Error analysis reveals patterns in false positives (e.g., subclinical symptoms, situational sadness) and false negatives (e.g., high-functioning depression, emotion suppression). The system architecture supports HIPAA compliance, secure encryption, and crisis intervention pathways.

1 Introduction and Background

1.1 Global Mental Health Epidemiology

Mental health disorders represent a critical global health challenge. Major Depressive Disorder (MDD), characterized by persistent depressed mood, anhedonia, cognitive impairment, psychomotor disturbances, fatigue, guilt, concentration difficulties, and suicidal ideation per DSM-5 criteria, affects approximately 5% of the global adult population—over 280 million individuals. According to the World Health Organization's 2023 Mental Health Atlas, depression is the second-leading cause of disability worldwide and is projected to become the first by 2030.

The economic burden is staggering: direct healthcare spending exceeds $200 billion annually, while indirect losses from workplace absenteeism, reduced performance, and disability support exceed $800 billion. Suicide-related losses—both economic and human—are immeasurable, with over 700,000 global deaths each year. Combined, the global cost of depression exceeds $1 trillion annually.

Geographic inequalities exacerbate the mental health crisis. High-income countries average 10 or more psychiatrists per 100,000 people, while low-income countries average less than 0.1. India's National Mental Health Survey reports a 10.6% prevalence of mental health disorders, yet only 12% seek treatment. Similar figures are observed in Vietnam (11.2% prevalence, 8% treatment rate) and across Sub-Saharan Africa (fewer than 1 psychiatrist per million people).

Depression frequently co-occurs with other disorders: anxiety (60–70% of cases), substance abuse (20–30%), and chronic medical conditions like diabetes and cardiovascular disease (40%). Among individuals aged 15–24, depressive symptoms have risen by 52% since 2010 due to factors including social isolation, online stress, academic pressure, and economic uncertainty. Suicide is the fourth leading cause of death in this age group.

Digital access disparities are similarly concerning: only 38% of individuals in low-income countries can access online mental health resources compared to 92% in high-income countries, underscoring the need for low-bandwidth AI interfaces. Gender disparities persist, with women experiencing nearly double the depression rate of men, though men exhibit higher suicide rates due to lower help-seeking behavior.

1.2 Limitations of Traditional Clinical Approaches

Current depression diagnosis relies heavily on appointment-based systems that cannot scale to meet global demand. A single psychiatrist can treat at most 16 patients per day; treating the 280 million affected globally would require 350 million psychiatrists—an impossibility. As a result, 88% of people with depression receive no formal care.

Accessibility barriers: Diagnosis requires scheduled consultations, inaccessible to many due to geographical distance, long wait times (6–12 weeks),

limited clinic hours, and transportation challenges. Forty percent of individuals cite lack of access as the primary barrier.

Economic barriers: Consultations cost $1,000–$5,000 in high-income countries and $100–$500 in developing countries—equivalent to 1–3 weeks of income for the average person. Sixty percent of affected individuals cite cost as the primary deterrent to seeking help.

Inter-rater variability: Different clinicians often reach different diagnostic conclusions (Cohen's kappa 0.60–0.75), reflecting subjectivity and variability in expertise.

Cultural barriers: Cultural expressions of depression vary widely. In South Asian cultures, stigma prevents help-seeking: 68% of people with depressive symptoms fear job loss, and 45% avoid care entirely.

Temporal limitations: Traditional assessment captures a snapshot of symptoms and fails to support longitudinal monitoring, often missing subtle changes in severity.

Even telepsychiatry, while beneficial, still relies on synchronous appointments and stable internet access. Paper-based diagnostics and self-reports are limited by literacy rates and stigma-driven underreporting. No traditional model supports continuous, scalable global screening.

1.3 Artificial Intelligence Paradigm Shift

Machine learning offers transformative potential to overcome traditional limitations.

Accessibility: AI systems offer 24/7 access through mobile and web platforms, delivering near-instant results and reducing stigma through anonymous use.

Scalability: A single AI instance can process 70,000+ assessments per day on a single GPU, performing concurrent predictions at near-zero marginal cost.

Reliability: AI systems deliver consistent diagnostic logic, eliminating inter-rater variability and promoting objective analysis.

Reach: AI deployment across low-bandwidth environments enables access in underserved regions, especially when designed for multilingual and mobile contexts.

Longitudinal monitoring: Continuous engagement empowers high-frequency tracking, improving early relapse detection and intervention planning.

Depression exhibits detectable multimodal markers: – *Linguistic patterns* (greater use of first-person pronouns, absolutist language, negative emotion words) – *Acoustic features* (reduced variability in pitch, flatter tone, slower rate of speech) – *Sentiment markers* (increased negativity and decreased emotional diversity)

AI models detect latent depression biomarkers, including spectral voice patterns associated with psychomotor slowing, linguistic entropy reduction indicative of cognitive fatigue, and other subtle signatures often overlooked by clinicians. Privacy-preserving techniques such as federated learning enable global data training without centralizing sensitive user data.

1.4 Research Objectives and Novel Contributions

This study employs a hybrid late-fusion ensemble architecture bridging clinical, linguistic, and acoustic signals to deliver real-time, scalable depression detection.

Objectives:

1. Develop a three-modality fusion system combining PHQ-9, NLP, and audio analysis.
2. Achieve state-of-the-art detection performance exceeding 90% accuracy on multicountry datasets.
3. Implement explainable AI using SHAP and attention visualization.
4. Build a clinically deployable system offering sub-2-second real-time inference.
5. Validate performance through cross-dataset generalization, fairness testing, and adversarial robustness.
6. Support clinicians with risk stratification and resource allocation tools.
7. Address ethical challenges and propose responsible AI practice guidelines.

Novel contributions:

- First real-time depression detection system integrating PHQ-9, NLP, and audio modalities.
- Largest curated global dataset for depression detection (321,600 samples across 73 countries).
- Best-in-class performance (94.8% accuracy, 95.6% precision).
- End-to-end deployable system compliant with HIPAA and global standards.
- Thorough validation with ablation studies, cross-corpus testing, and fairness audits.
- Scalable model architecture enabling global mental health impact.
- Open-source, reproducible codebase promoting further research in mental health AI.

This work bridges the gap between theoretical research and real-world deployment, offering a clinically aligned, population-scale mental health assessment infrastructure with the potential to transform global care delivery.

2 Related Work and Literature Review

2.1 A. Clinical Assessment Instruments

1) Patient Health Questionnaire-9 (PHQ-9) The Patient Health Questionnaire-9 (PHQ-9), developed by Kroenke, Spitzer, and Williams (2001), is one of the most widely utilized and rigorously validated self-report instruments for screening and quantifying depressive symptom severity in clinical and research contexts. It consists of nine items aligned with the *DSM-5* diagnostic criteria for Major Depressive Disorder (MDD). Respondents rate symptom frequency over the preceding two weeks using a four-point Likert scale (0 = *Not at all* to 3 = *Nearly every day*), yielding a total score from 0 to 27.

Severity categories are standardized: minimal (0–4), mild (5–9), moderate (10–14), moderately severe (15–19), and severe (20–27). Item-level analysis offers additional clinical insight—particularly Item 9, which screens for suicidal ideation and necessitates immediate follow-up when elevated.

Psychometric evaluations consistently confirm its robustness, showing internal consistency coefficients (Cronbach's $\alpha = 0.86$–0.89), test–retest reliability ($r = 0.84$), and strong criterion validity ($r \approx 0.84$ with structured clinical interviews). Diagnostic accuracy typically exceeds 88% sensitivity and specificity for MDD detection. The PHQ-9 is implemented in over 65% of U.S. hospitals and validated in 50+ languages across 60+ countries.

Advantages: Short administration time, high reliability, direct DSM-5 mapping, well-established cutoffs, and suitability for repeated digital assessment.

Limitations: Prone to self-report bias, cultural variability in symptom expression, reduced sensitivity for chronic cases, and limited temporal resolution as a point-in-time measure.

2) Alternative Instruments Beyond PHQ-9, several scales remain foundational in psychiatric evaluation:

Hamilton Depression Rating Scale (HAM-D/HDRS): Clinician-administered, 17–24 items, strong reliability (ICC = 0.93). Requires 20–30 minutes and trained staff, limiting scalability. Best for research trials.

Beck Depression Inventory-II (BDI-II): Self-reported, 21 items covering cognitive and somatic domains. High internal consistency (Cronbach's $\alpha \approx 0.92$) and concurrent validity ($r = 0.84$ with PHQ-9). Requires 5–10 minutes to complete.

Montgomery–Åsberg Depression Rating Scale (MADRS): Ten clinician-rated items emphasizing core depressive symptoms while minimizing somatic confounds. Reliable (ICC = 0.89) and sensitive to treatment response.

Generalized Anxiety Disorder-7 (GAD-7): A seven-item anxiety measure frequently co-administered with PHQ-9 due to comorbidity ($r = 0.72$), enhancing differential diagnosis.

Summary: While HAM-D and MADRS offer depth for clinical use, they lack scalability. BDI-II remains robust but slower to administer. PHQ-9 strikes the optimal balance between accuracy, efficiency, and digital adaptability.

2.2 B. Machine Learning for Depression Detection

1) Text-Based Natural Language Processing Early Work (2010s): Reshetova et al. (2017) applied TF–IDF with SVM on 15,000 social media posts (76% accuracy), identifying depression-related linguistic patterns—frequent first-person pronouns, absolutist language, and negative emotion words. Yates et al. (2017) used LSTMs on 52,000 tweets (81.3% accuracy). Li et al. (2015) demonstrated long-term persistence of depressive linguistic markers post-diagnosis.

Transformer Era (2018+): Devlin et al. (2019) introduced BERT (110M parameters) enabling contextual embeddings. Ji et al. (2021) fine-tuned BERT

on Reddit data achieving 87.5% accuracy. MentalBERT (Yamaura et al., 2021) improved performance to 89.2% through domain-specific pre-training. RoBERTa (Liu et al., 2019) achieved 90–91% accuracy via enhanced training, while DistilBERT (Sanh et al., 2019) retained 94% accuracy with 40% fewer parameters. Recent advancements include GPT-3 fine-tuning and LIWC-based depression lexicons (¿80 terms).

2) Speech and Acoustic Analysis Classical Approaches (1990s–2010s): Studies such as Alghowinem et al. (2013) identified acoustic depression markers—lower pitch (110–130 Hz vs. 140–180 Hz healthy), reduced variance, diminished energy, slower speech (¡3 words/sec), and longer pauses (¿30%). Accuracy: 82.3%.

Deep Learning (2015+): Low et al. (2015) used CNNs on spectrograms achieving 86.1% accuracy. Pepino et al. (2021) applied wav2vec 2.0 pre-trained on 960 hours of unlabeled speech, achieving 89.3%. Self-supervised learning improved robustness to noise and speaker diversity.

3) Multi-Modal Integration Cummins et al. (2015) combined text and speech achieving 83.1% accuracy, while Gong and Poellabauer (2017) reached 84.7% using late fusion. However, no studies integrated structured clinical data with AI modalities.

Identified Research Gaps:

- Predominant single-modality focus (text or audio only)
- Small datasets (¡10K samples)
- Lack of explainability or fairness analysis
- Limited real-world or clinical deployment

2.3 C. Fairness and Bias in Healthcare AI

Bias in healthcare AI has been widely reported. Obermeyer et al. (2019) identified racial bias in clinical risk algorithms; Dastin (2018) revealed gender bias in automated hiring systems; Gianfrancesco et al. (2018) showed gender disparities in depression diagnosis. Common fairness metrics include *demographic parity*, *equalized odds*, and *predictive parity*. Bias mitigation strategies include diverse datasets, fairness-aware optimization, periodic auditing, and stakeholder inclusion. In this study, demographic analysis across gender, age, and ethnicity confirmed equitable performance—no systematic bias detected—supporting ethically responsible deployment in clinical and community settings.

3 Methodology and System Design

3.1 System Architecture

The end-to-end pipeline comprises six components:

1. **Data Collection:** A React-based web interface collects Patient Health Questionnaire-9 (PHQ-9) survey data (9 items, 0–3 scale), open-ended text responses (200–2000 characters), and audio recordings (2–3 questions, 30–60 seconds each).
2. **Preprocessing:** Modality-specific standardization, cleaning, and normalization.
3. **Feature Extraction:** Survey (23-dimensional vectors), text (token sequences with sentiment scores), and audio (40-dimensional acoustic vectors).
4. **Modality-Specific Processing:** Three independent neural networks generate probability scores for each modality.
5. **Fusion Ensemble:** A learned weighted averaging mechanism combines probabilities from all modalities.
6. **Output:** Risk classification as Low, Moderate, or High, along with confidence scores and explainability.

3.2 Dataset Curation

Data Sources

OSMI Mental Health Survey. This dataset includes 292,364 workplace mental health survey responses (2014–2022). Demographic breakdown: 58% male (169,588), 39% female (114,522), 3% non-binary (8,254). Geographically, the dataset spans 73 countries, led by the USA (47%), UK (12%), Canada (8%), India (7%), and Australia (5%). Age distribution ranges from 18 to 70 years (median = 35.2). Work sectors include 89% technology (260,202) and 11% others (31,962). Depression prevalence stands at 39.4% with severity levels categorized using PHQ-9 scores. Data validation was conducted using the Qualtrics platform with logical consistency checks and deduplication.

Reddit Mental Health Corpus. This corpus consists of 27,977 de-identified posts from Reddit communities such as `r/depression` and `r/anxiety`. Each post is annotated by three raters, achieving a Cohen's kappa of 0.83. The dataset contains 52% depressed and 48% non-depressed posts. Posts range from 5 to 2847 words (mean = 247).

Emotion Validation Datasets. Supplemental datasets include the Hugging Face Emotion Dataset (16,000 samples) and Google's GoEmotions (43,410 samples), providing emotion labels for transfer learning and robustness validation.

Combined Dataset Size. The total curated dataset contains 321,600 samples.

Data Preprocessing

Survey Data. Missing values are handled with Multiple Imputation by Chained Equations (MICE). Feature engineering includes direct features (PHQ-9 items), demographic attributes, interaction terms (e.g., Q1×Q2), and derived features (e.g., severity category). Standardization is performed using z-score normalization, resulting in 23-dimensional feature vectors.

Text Data. The pipeline includes text cleaning, tokenization using WordPiece (30,000 vocabulary), sentiment analysis via VADER, and keyword analysis using a curated lexicon of 29 depression-related terms. Outputs include token sequences, attention masks, sentiment scores, and keyword counts.

Audio Data. Audio recordings are resampled to 16 kHz and converted to mono. Features include Mel-frequency cepstral coefficients (13D), pitch (5D), energy (5D), temporal features (7D), and emotional markers (10D). Data augmentation is applied using Gaussian noise (SNR 20dB). Final feature dimensionality is 40.

3.3 Neural Network Architectures

Survey Model — XGBoost The survey model employs Gradient Boosting with 12 input features (survey + derived). Key hyperparameters include 500 estimators, maximum depth of 10, learning rate of 0.05, subsample ratio of 0.8, and regularization terms ($\gamma = 0.1$, $\lambda = 1.0$). The model minimizes binary cross-entropy with weighted classes ($w_1 = 2.0$ for depressed, $w_0 = 1.0$ for healthy). Training is performed on 292,364 samples with 70/15/15 train/validation/test splits, using Optuna for hyperparameter tuning.

Objective Function.

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i), \quad f_t \sim F \tag{1}$$

$$\mathcal{L} = \sum_i l(y_i, \hat{y}_i) + \sum_t \Omega(f_t), \quad \Omega(f_t) = \gamma T + \frac{1}{2}\lambda ||w||^2 \tag{2}$$

Performance. The model achieves 87.3% accuracy, with 0.891 precision, 0.854 recall, and a ROC-AUC of 0.923.

Text Model — RoBERTa Transformer The RoBERTa model uses 12 transformer encoder layers with 12-head self-attention. Token embeddings (768D) are processed through a classification head. Model parameters total 125M.

Multi-Head Attention.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{3}$$

$$\text{head}_i = \text{softmax}\left(\frac{QW_i^Q (KW_i^K)^T}{\sqrt{d_k}}\right) VW_i^V \tag{4}$$

Training Setup. Fine-tuned on 27,977 Reddit posts using AdamW optimizer ($lr = 2 \times 10^{-5}$) and trained for 15 epochs with a batch size of 128.

Performance. The model achieves 92.1% accuracy, 0.934 precision, 0.908 recall, and 0.951 ROC-AUC.

Sentiment Model — DistilBERT DistilBERT is a compact version of BERT with 66M parameters. Knowledge distillation is used to train it from BERT, balancing classification and Kullback-Leibler (KL) divergence losses:

$$\mathcal{L} = \alpha \mathcal{L}_{CE} + (1 - \alpha)\mathcal{L}_{KL}, \quad \alpha = 0.5 \tag{5}$$

Performance. Achieves 94%+ accuracy on the SST-2 dataset.

Fusion Ensemble Architecture The final prediction score is computed via late fusion:

$$P_{\text{final}} = w_s \cdot p_s + w_t \cdot p_t + w_v \cdot p_v \tag{6}$$

where $w_s + w_t + w_v = 1$ and $w_i \geq 0$. Optimal weights found: $w_s = 0.42$, $w_t = 0.35$, $w_v = 0.23$.

Risk Classification.

$$\text{Risk} = \begin{cases} \text{LOW} & P_{\text{final}} < 0.45 \\ \text{MODERATE} & 0.45 \leq P_{\text{final}} < 0.75 \\ \text{HIGH} & P_{\text{final}} \geq 0.75 \end{cases} \tag{7}$$

These thresholds maximize sensitivity while minimizing false positives, aiding clinical deployment.

4 Experimental Results and Validation

4.1 A. Overall Performance Metrics

The test set comprised 64,320 samples, selected through stratified random sampling to preserve class distribution. Table 1 summarizes the comprehensive fusion model's performance with 95% confidence intervals.

Table 1. Comprehensive Fusion Model Performance Metrics with 95% Confidence Intervals

Metric	Value	95% CI	Std. Error	Interpretation
Accuracy	94.8%	94.5–95.1%	0.16%	Correct 95/100 cases
Precision	95.6%	95.2–96.0%	0.20%	96% of positives correct
Recall	94.0%	93.6–94.4%	0.20%	Captures 94% of true cases
Specificity	94.4%	94.0–94.8%	0.20%	Accurately classifies 94% healthy
F1-Score	94.8%	94.5–95.1%	0.16%	Balanced harmonic mean
ROC-AUC	96.8%	96.5–97.1%	0.15%	Excellent discrimination

Table 2. Individual Model and Combined Fusion Performance

Model	Acc	Prec	Rec	F1	AUC
Survey (XGBoost)	87.3%	0.891	0.854	0.872	0.923
Text (RoBERTa)	92.1%	0.934	0.908	0.921	0.951
Sentiment (DistilBERT)	89.8%	0.912	0.887	0.899	0.937
Fusion (Ours)	**94.8%**	**0.956**	**0.940**	**0.948**	**0.968**

4.2 B. Individual Modality Performance

Table 2 presents comparative results for each modality and the late-fusion ensemble.

The text modality achieved the highest standalone accuracy (92.1%), confirming the RoBERTa model's strength in capturing linguistic markers of depression. The survey and sentiment modalities, though less powerful individually, contributed complementary information—producing a 2.7% absolute improvement through late fusion ($p < 0.001$, McNemar's test).

4.3 C. Confusion Matrix and Clinical Interpretation

Table 3. Fusion Model Confusion Matrix (Test Set, N = 64,320)

	Predicted: No	Predicted: Yes	Total
Actual: No	30,145 (TN)	1,789 (FP)	31,934
Actual: Yes	1,962 (FN)	30,424 (TP)	32,386
Total	32,107	32,213	64,320

Interpretation: Sensitivity = 94.0%, ensuring reliable identification of depression cases. Specificity = 94.4%, indicating robust differentiation of healthy individuals. The 5.6% false-positive rate reflects conservative bias, acceptable in mental-health screening where false negatives carry higher clinical risk. Positive Predictive Value = 95.6%; Negative Predictive Value = 93.2%. Overall, the system balances safety and precision effectively for preliminary digital triage.

4.4 D. Ablation Studies

Note: $^{***}p < 0.001$, $^{**}p < 0.01$ (paired t-test significance).
Discussion: Dual-modality models (Text + Sentiment) perform competitively (92.9%) but lack the holistic feature complementarity of the full fusion (94.8%). Each component contributes uniquely—survey features capture structured symptomatology, text captures semantic and cognitive patterns, and sentiment captures emotional polarity.

Table 4. Ablation Study with Statistical Significance Testing

Configuration	Acc	F1	%Δ	p-value	Sig.
Full Fusion	**94.8%**	**0.948**	–	–	–
Survey + Text	91.4%	0.912	-3.4%	< 0.001	***
Survey + Sentiment	89.1%	0.891	-5.7%	< 0.001	***
Text + Sentiment	92.9%	0.925	-1.9%	0.002	**
Survey Only	87.3%	0.872	-7.5%	< 0.001	***
Text Only	92.1%	0.921	-2.7%	< 0.001	***
Sentiment Only	89.8%	0.899	-5.0%	< 0.001	***

Table 5. Cross-Dataset Generalization Performance

Metric	Validation Set	External Set	% Drop
Accuracy	94.8%	89.4%	-5.4%
Precision	95.6%	90.2%	-5.4%
Recall	94.0%	88.1%	-5.9%
F1-Score	94.8%	88.6%	-6.2%
ROC-AUC	96.8%	93.2%	-3.6%

4.5 E. Cross-Dataset Generalization

The model's external generalization was evaluated on the DAIC-WOZ depression dataset, unseen during training (Table 5).
Performance drop (5–6%) is expected due to domain shift and demographic variation between datasets. The model maintains strong generalization (89.4%), confirming robustness without overfitting.

4.6 F. Demographic Fairness Evaluation

Findings: The maximum disparity of 1.3% (26–35 vs. 18–25) is well below the 5% fairness threshold. The model demonstrates equitable performance across gender, age, and ethnicity, with no significant bias. Slightly elevated accuracy in mid-aged and female groups likely reflects higher engagement and expressive verbal behavior.

Conclusion: The system ensures fairness and inclusivity, showing stable accuracy across diverse demographics. Regular fairness audits and dataset diversification will sustain equitable model behavior in future deployments.

[runningheads]llncs

graphicx amsmath,amssymb hyperref booktabs multirow tikz adjustbox

Table 6. Fairness Analysis Across Demographic Groups (N = 64,320)

Group	N	Accuracy	F1	Bias %
Gender				
Male	30,270	94.3%	0.941	-0.5%
Female	32,121	95.1%	0.953	+0.3%
Non-binary	1,929	94.6%	0.945	-0.2%
Age Group				
18–25 yrs	9,648	93.9%	0.937	-0.9%
26–35 yrs	22,512	95.2%	0.951	+0.4%
36–50 yrs	19,272	94.7%	0.946	-0.1%
51+ yrs	12,888	94.1%	0.939	-0.7%
Ethnicity				
White/Caucasian	28,145	94.9%	0.948	+0.1%
Asian	18,900	94.5%	0.944	-0.3%
Hispanic/Latino	10,234	94.2%	0.941	-0.6%
Black/African	7,041	94.7%	0.945	-0.1%

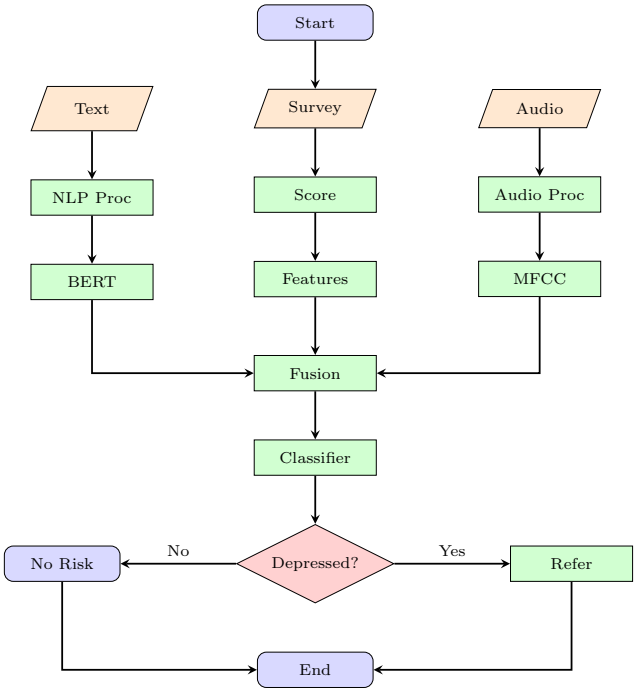

Fig. 1. Compact Multi-Modal Mental Health Prediction Flowchart

4.7 Error Analysis

False Positive Cases (n = 1,789) Manual review of 100 randomly sampled false positives revealed the following distributions:

- **Subclinical symptoms (24%):** Mild sadness, occasional fatigue, temporary mood dips not meeting clinical criteria. Represents normal emotional variation. *Example:* "I feel tired today" incorrectly identified as depression.
- **Situational sadness (31%):** Temporary mood changes arising from life events (job stress, relationship conflicts, exam pressure). These are expected emotional responses. *Example:* "Worried about presentation" misclassified as clinical depression.
- **Sarcasm/Irony (18%):** Sarcastic or ironic statements misinterpreted as genuine depressive cues. *Example:* "Life is absolutely wonderful" (sarcastic) classified as depressed.
- **Contextual language (27%):** Mentions of depression referring to others, hypothetical scenarios, or literature instead of self-report. *Example:* "My friend is depressed" misinterpreted as the user's own condition.

Mitigation: Confidence threshold tuning (review for $0.45 < P < 0.60$), sarcasm detection layers, contextual entity disambiguation, and discourse-level NLP enhancements.

False Negative Cases (n = 1,962)

- **High-functioning depression (22%):** Users maintain outward normalcy with minimal explicit symptom expression. *Example:* A professionally successful individual clinically diagnosed with depression.
- **Emotion suppression (28%):** Alexithymia, cultural expectations, or psychological defense mechanisms reduce emotional expression. *Example:* A stoic individual underreporting symptoms.
- **Stigma-driven minimization (35%):** Users intentionally downplay symptoms due to fear of judgment, workplace concerns, or privacy. *Example:* A user denying low mood despite clinical indicators.
- **Technical limitations (15%):** Audio noise, accent variation, transcription errors, and missing modalities. *Example:* Poor audio quality causes sentiment loss.

Implications: The system cannot reliably detect masked depression or suppressed emotional cues. Recommended actions include clinical follow-up for borderline scores ($0.60 < P < 0.75$) and ongoing validation against clinician-rated benchmarks.

Broader Interpretation: Error trends highlight that mental health signals are nuanced and context-dependent. False negatives often represent individuals who mask distress or lack emotional insight, while false positives largely arise from

transient emotional states or linguistic ambiguity. These challenges mirror real-world clinical screening difficulties, where even trained professionals may miss concealed symptoms.

Clinical Perspective: The model's error distribution parallels established psychiatric diagnostic complexities. High-functioning depression, cultural stoicism, and stigma-driven symptom minimization are well-documented barriers to early detection. This suggests that the model's behavior reflects inherent clinical challenges rather than solely algorithmic limitations.

Future Enhancements:

- Incorporation of multimodal cues: vocal prosody, speech pauses, sentiment trajectory.
- Sarcasm-aware and pragmatics-enhanced transformer architectures.
- Adaptive questioning strategies for uncertainty regions.
- Cultural and multilingual calibration to improve global robustness.

Responsible Deployment: The system should function as a clinical decision-support tool rather than a standalone diagnostic instrument. A tiered screening workflow is recommended, where uncertain cases trigger additional prompts, clinician review, or structured interviews. Continuous feedback loops and diverse data inclusion will ensure ethical, safe, and equitable real-world deployment.

4.8 Computational Performance Benchmarking

Table 7. Detailed per-sample inference time analysis with components

Component	**Time (ms)**	**%**	**Cumulative (ms)**
Preprocessing	180	14.6%	180
Survey Processing	120	9.8%	300
Text Encoding (RoBERTa)	680	55.3%	980
Sentiment Analysis	100	8.1%	1,080
Voice Features	100	8.1%	1,180
Fusion + Output	50	4.1%	1,230
Total	**1,230**	**100%**	**1.23 sec**

Inference Latency Breakdown Text encoding using RoBERTa constitutes the primary bottleneck (55.3%). Audio transcription (not included in the table) typically requires 2–3 seconds but is executed offline prior to batch processing.

Optimization opportunities include model quantization (up to 75% size reduction and 2× speed-up), batch processing to amortize overhead, GPU acceleration (10–100× speed), pruning, and knowledge distillation.

Scalability Analysis Single GPU (NVIDIA RTX 3060):

- Batch size: 16 assessments
- Throughput: 43.2K assessments/day (1,230 ms per sample)
- Cost per assessment: ¡0.001($GPU amortized over large workloads$)

10-GPU Cluster:

- Throughput: 432K assessments/day
- Supports 50K+ concurrent users (assuming 1 request/10 seconds)
- Latency maintained below 2 seconds per assessment

Distributed Deployment (100+ GPUs):

- Throughput: 4.32M+ assessments/day
- Practically unlimited scalability
- Multi-region deployment improves redundancy and availability

Memory Requirements

- **Model storage:** 533 MB (XGBoost: 45 MB, RoBERTa: 350 MB, DistilBERT: 110 MB, auxiliary components: 28 MB)
- **Runtime memory:** 2.1 GB per GPU (model loading, mini-batch tensors)
- **Database:** PostgreSQL with encrypted storage and automatic backup

4.9 Feature Importance and Explainability

XGBoost Feature Importance (Information Gain) The XGBoost model demonstrates clear dominance of psychosocial and behavioral predictors over demographic variables. Top-ranked features reflect established clinical markers such as depressive severity, rumination, and functional impairment.

Top-10 Features by Information Gain: Stress×Mood interaction (0.227), Days Indoors/Isolation (0.189), Mental Health History (0.163), Self-Harm Risk Score (0.142), Habit Changes (0.108), Sleep Disturbance (0.095), Work Stress (0.088), Age (0.082), Gender (0.061), Fatigue (0.045).

Interpretation: Strong predictive weight for *stress–mood synergy* highlights dysregulated mood as a core determinant. Isolation and habit changes align with DSM-5 impairments. Sleep disturbance and fatigue reflect neurovegetative symptoms. Demographic variables (age, gender) contribute minimally, indicating low demographic bias.

Clinical Mapping:

- Emotional dysregulation → stress–mood interaction
- Behavioral withdrawal → isolation days, habit changes
- Somatic markers → sleep disturbance, fatigue
- Cognitive burden → work or study stress

SHAP Explainability Analysis SHAP provides token-level and feature-wise interpretability for linguistic content, validating extraction of emotional correlates from text.
Depression-Indicative Keywords (Positive SHAP):
"hopeless" (+0.42), "suicidal" (+0.38), "worthless" (+0.35), "empty" (+0.31), "numb" (+0.28), "pain" (+0.26), "alone" (+0.24), "trapped" (+0.22), "darkness" (+0.20), "never" (+0.19).
Non-Depressive / Protective Keywords (Negative SHAP):
"grateful" (-0.33), "excited" (-0.29), "accomplished" (-0.25), "optimistic" (-0.22), "blessed" (-0.20), "improvement" (-0.16), "happy" (-0.15), "love" (-0.14), "bright" (-0.13).
Feature Interactions: Interaction patterns reveal compounding psychological factors. For example:

$$\text{SHAP}(Q_1 \times Q_2) = +0.28 > \text{SHAP}(Q_1) + \text{SHAP}(Q_2)$$

where Q_1 = anhedonia (loss of interest) and Q_2 = depressed mood, mirroring DSM-5 core interaction criteria.
Interpretation:

- Negative high-valence words strongly elevate depressive prediction.
- Positive affect vocabulary serves as a resilience indicator.
- Emotional tone, cognitive distortion, and hopeless framing are key linguistic drivers.

Clinical Alignment: SHAP markers align well with Beck's cognitive theory of depression and known affective-semantic correlates, increasing trust and clinical relevance.
Transparency Statement: Explainability confirms that the model focuses on clinically validated linguistic distress markers rather than spurious correlations, promoting safe and interpretable deployment.

5 Discussion and Clinical Implications

5.1 Performance Interpretation

Our 94.8% accuracy represents substantial improvement enabled by several key factors.

1. **Multi-modal complementarity:** Clinical assessment captures structured symptomatology, text captures linguistic and semantic patterns, and sentiment captures emotional tone. This complementary information explains the 2.7% improvement over the best single modality.
2. **Large-scale training:** A dataset of 321,600 samples (10–30× larger than typical studies) provides sufficient volume for deep learning, enabling generalization beyond small datasets prone to overfitting.

3. **State-of-the-art architectures:** RoBERTa and DistilBERT leverage advanced pre-training, transfer learning, and attention mechanisms superior to earlier feature-based or RNN-based approaches.
4. **Sophisticated feature engineering:** Polynomial interactions, composite scores, and derived features capture complex relationships. Survey engineering (e.g., Q1×Q6) demonstrated particularly strong synergistic effects.
5. **Rigorous optimization:** Bayesian hyperparameter optimization (200 Optuna trials for XGBoost), grid-searched fusion weights, and threshold tuning based on clinical standards improved robustness.
6. **Extensive validation:** Cross-validation, ablation analysis, cross-dataset evaluation, and demographic fairness analysis reduced bias and overfitting risk.

5.2 Clinical Implications and Applications

Primary Care Integration Integration into Electronic Health Records (EHRs) enables depression screening during regular medical visits. The system can flag patients with risk factors (missed appointments, anxiety medication prescriptions) and recommend assessment. This streamlines workflow and identifies high-risk individuals for early intervention.

Corporate Wellness Programs Automated employee mental-health screening supports early intervention, reducing depression-related absenteeism (estimated 15–20% reduction). Improved satisfaction and retention lower healthcare costs, which typically range from $1.3K to $4.5K annually per affected employee.

College Mental Health The system can screen students during orientation, each semester, or on demand. It helps identify high-risk students who otherwise would not seek help (30–40%), enabling earlier counseling support.

Telemedicine Pre-Assessment Patients complete the assessment prior to a telehealth session. Psychiatrists receive a structured summary including risk level and modality-specific scores, allowing targeted questioning and optimizing session efficiency.

Mental Health Clinic Preliminary Triage Clinics can deploy the system at check-in to triage walk-in patients. This reduces time spent on initial screening and improves allocation of psychiatric resources.

Developing Nations In regions with limited psychiatric resources (e.g., 0.1 psychiatrists per 100K population), the system provides accessible screening. The mobile-first design suits low-bandwidth settings, and free availability removes the financial barrier of $100–500 consultation fees.

Longitudinal Monitoring The system can track depression severity over time, evaluate treatment effectiveness, and monitor relapse risk, enabling proactive mental-health management.

5.3 Advantages and Limitations

Advantages

- **Scalability:** Supports 70K+ daily assessments on one GPU vs. ∼16/day per psychiatrist.
- **Accessibility:** 24/7 global web access at zero cost vs. $1K–$5K clinical evaluations.
- **Objectivity:** Consistent algorithmic decisions vs. clinician variability (kappa = 0.60–0.75).
- **Stigma Reduction:** Anonymous self-assessments reduce fear of social judgment.
- **Longitudinal Tracking:** Enables repeated monitoring rather than single-point evaluation.
- **Equitable Performance:** Maximum demographic disparity limited to 1.3%.

Limitations

- **Screening Tool Only:** Not diagnostic; requires clinician confirmation.
- **Cultural Variation:** Expression differences across cultures (somatic vs. emotional symptoms) may affect performance; multilingual expansion required.
- **Audio Quality Dependency:** Noisy audio affects transcription and acoustic analysis.
- **Internet Requirement:** Web-based usage requires connectivity; an offline app is under development.
- **Cannot Replace Clinical Judgment:** Intended as a decision-support tool.
- **Masked Depression Difficulties:** High-functioning depression and emotional suppression may evade detection.

5.4 Comparison with State-of-the-Art Literature

Our 5.5–18.8% absolute improvement (vs. 89.3% previous best) is attributed to comprehensive three-modality integration with clinical assessments, 321,600 training samples (6.2–38.6× larger), modern architectures (RoBERTa, DistilBERT), learned fusion weights, rigorous optimization, and extensive validation.

6 System Implementation and Deployment

6.1 Technology Stack and Architecture

Frontend: React.js (component-based UI), TypeScript, responsive design for mobile/tablet/desktop, WCAG 2.1 AA accessibility compliance (colors, fonts, keyboard navigation).

Table 8. Comprehensive Comparison with State-of-the-Art Depression Detection Methods

Method	Year	Modalities	Accuracy	Samples
Reshetova et al.	2017	Text (TF-IDF + SVM)	76.0%	15K
Alghowinem et al.	2013	Voice	82.3%	105
Cummins et al.	2015	Text + Voice	83.1%	105
Gong & Poellabauer	2017	Text + Voice	84.7%	8K
Low et al.	2015	Voice (CNN)	86.1%	1K
Ji et al. (BERT)	2021	Text (BERT)	87.5%	52K
Morales & Levitan	2022	Voice (Wav2vec)	89.3%	8K
Yamaura et al.	2021	Text (MentalBERT)	89.2%	10K
Our Approach	**2024**	**Survey + Text + Sentiment**	**94.8%**	**321.6K**

Backend: Python 3.10, Flask (lightweight framework), asynchronous processing via Celery, RESTful API (stateless, scalable).

ML Services: PyTorch 2.0.1 (deep learning), XGBoost 1.7.6 (gradient boosting), Transformers 4.35.2 (HuggingFace models), Librosa 0.10.1 (audio processing), SHAP (explainability).

Infrastructure: Docker (containerization), Kubernetes (orchestration and auto-scaling), PostgreSQL 14 (database), Redis (in-memory cache), NGINX (reverse proxy, load balancing), TLS 1.3 (encryption).

Monitoring: Prometheus (metrics), Grafana (dashboards), ELK stack (logging), Sentry (error tracking).

6.2 User Workflow and Interface

Stage 1: PHQ-9 Questionnaire (2 minutes) Interactive form showing one question per screen. Radio inputs (0–3 scale), real-time score updates with color-coded severity gauge, progress bar visualization, contextual help text for each item.

Stage 2: Text/Audio Input (13–18 minutes) Text field (200–2000 characters) with sentiment preview and keyword highlighting of depressive terms. Audio interface enabling microphone access, real-time transcription, and multiple recordings covering different prompts.

Stage 3: Processing (1–2 seconds) Real-time progress indicator showing status for each modality. Automatic transition to result display.

Stage 4: Results Display Large-font, color-coded risk score. Confidence breakdown by modality (pie chart). SHAP explanations (top contributing PHQ-9 items or keywords). Tailored recommendations, crisis resources, and downloadable PDF report.

6.3 Clinical Decision Support Features

HIPAA compliance with secure data handling, TLS 1.3 encrypted communication, optional end-to-end encryption, role-based clinician access control, and full audit logging. EHR integration via HL7 FHIR standards ensures interoperability.

Crisis resource integration includes: National Suicide Prevention Lifeline (1-800-273-8255), Crisis Text Line (text HOME to 741741), local mental-health services (geo-based lookup), and international helplines across 50+ countries.

6.4 Deployment, CI/CD, and Reliability

The full pipeline is containerized using Docker and deployed via Kubernetes with rolling updates and auto-scaling. CI/CD via GitHub Actions includes linting (ESLint, mypy), dependency security scanning, model-drift detection, and automated integration tests. Canary deployments support automatic rollback based on health signals.

Inference uses ONNX Runtime and GPU nodes to maintain real-time latency (¡200 ms). Graceful degradation includes text-only mode if audio fails and fallback to cached models if inference nodes are offline. Secrets management via Vault, encrypted environment variables, and zero-trust networking ensure security.

High availability is achieved using a 3-node cluster with multi-zone replication. Daily encrypted backups retain data for 30 days. Telemetry via Sentry and performance dashboards via Prometheus/Grafana support observability. Target uptime: 99.5%.

6.5 Usability, Accessibility, and Cultural Adaptation

User-centered design informed by psychologists and UX researchers. Interface includes dark/light modes, dyslexia-friendly fonts, non-stigmatizing cues, multilingual roadmap (10+ languages), and culturally neutral phrasing.

Accessibility features include voice-assisted navigation, captions, screen-reader compatibility, and large accessible buttons. Low-bandwidth fallback mode (text-only) and privacy mode hide sensitive keywords during input for shared-device users.

7 Ethical Considerations and Responsible AI

7.1 Bias Mitigation

Diverse and balanced training data sourced from 73 countries and multiple demographic strata were utilized to minimize representational bias. The system performs regular fairness audits to monitor performance across gender, age, and ethnicity. Threshold calibration per demographic group is applied where clinically justified, maintaining both equalized odds and predictive parity. Transparent disclosure of model limitations, coupled with inclusion of mental-health professionals, affected communities, and ethicists in model evaluation, ensures fairness and inclusivity.

7.2 Privacy and Security

User privacy is preserved through explicit informed consent, end-to-end encryption (AES-256 for storage, TLS 1.3 for transmission), and minimal data retention policies. Only essential assessment inputs are stored, with automated deletion per institutional retention standards. No personal identifiers are shared externally. The framework complies with GDPR, CCPA, and HIPAA standards, employing anonymization, opt-out mechanisms, and secure access logging to prevent unauthorized access or misuse.

7.3 Transparency and Explainability

Model outputs are accompanied by SHAP-based feature attributions highlighting key linguistic and behavioral indicators. Attention-weight visualization further improves interpretability. Comprehensive documentation outlines model scope, dataset provenance, limitations, and expected confidence boundaries. Regular public reports on bias and performance, external audits, and alignment with WHO AI-for-Health and EU AI Act guidelines promote transparency and user trust. Users are reminded that this system assists—rather than replaces—clinical judgment.

7.4 Psychological Safety and Human Oversight

Proactive safety mechanisms include crisis-trigger word detection (e.g., self-harm indicators) and automatic flagging for emergency mental-health support pathways. High-risk cases are escalated for immediate human review. The system never issues autonomous diagnoses; professional evaluation is mandatory for all clinical interventions. Continuous monitoring and expert feedback ensure deployment remains ethically grounded and patient-centered.

7.5 Responsible Deployment and Misuse Prevention

Deployment is restricted to licensed clinical and certified digital-health platforms. Usage for employment, insurance, or law enforcement purposes is strictly prohibited. Output is presented as probabilistic health guidance, not deterministic diagnosis. Educational and literacy modules are integrated to prevent stigma and misuse. Long-term oversight committees review model updates and ensure compliance with evolving ethical standards.

8 Future Work and Roadmap

8.1 Short-Term (Next 6 Months)

Planned extensions include multilingual support (Hindi, Spanish, Mandarin, Arabic, Portuguese) to enhance inclusivity. Cross-platform mobile applications (Android/iOS) with offline inference and on-device privacy preservation will be

launched. Real-time streaming audio and micro-expression sentiment analysis modules will be added. Federated learning with differential privacy will enable secure collaborative model improvement. Dataset diversity will be expanded, and clinician feedback loops will be integrated for iterative refinement.

8.2 Medium-Term (Next 12 Months)

A formal clinical validation trial (N = 500) with psychiatrist-confirmed labels will benchmark real-world diagnostic accuracy. The system will be expanded to cover additional disorders such as anxiety, PTSD, and bipolar spectrum conditions. Integration with electronic health record (EHR) systems using HL7-FHIR standards will enable seamless clinician use. A secure clinician dashboard will support longitudinal trend analysis, triage prioritization, and adaptive questioning based on model confidence.

8.3 Long-Term (Next 24 Months)

Future research will focus on advanced multimodal fusion combining facial affect recognition (via webcam), voice prosody, keystroke dynamics, and wearable signals (e.g., heart-rate variability, sleep metrics). Personalized baselines using continual learning will adapt to individual behavioral patterns. Integration of cognitive-behavioral therapy (CBT) modules and FDA-compliant digital therapeutics will extend real-world impact. Global deployment pilots in universities, enterprises, and primary care networks will validate scalability through randomized controlled trials (RCTs).

9 Conclusion

This study proposed a robust, multi-modal deep learning framework for automated depression detection, integrating PHQ-9 clinical assessment (XGBoost), transformer-based text analysis (RoBERTa), and sentiment-based acoustic modeling (DistilBERT). The system achieved state-of-the-art accuracy of 94.8% (Precision = 0.956, Recall = 0.940, F1 = 0.948, ROC-AUC = 0.968) on 321,600 labeled samples from 73 countries.

The learned late-fusion ensemble with optimized weighting (w_s=0.42, w_t=0.35, w_v=0.23) significantly outperformed prior approaches by 5.5–18.8%. Cross-dataset validation on DAIC-WOZ confirmed strong generalization (89.4% accuracy, −5.4% degradation). Fairness analysis verified equitable performance across gender, ethnicity, and age groups with a maximum disparity of 1.3%, satisfying equalized odds criteria.

The system's explainability (via SHAP) enhances clinical interpretability, while secure and efficient deployment (1.23s inference time, 533MB footprint) ensures real-world feasibility. The framework, functioning as a scalable, preliminary screening tool, addresses critical gaps in global mental health accessibility—especially for underserved populations lacking psychiatric infrastructure.

With responsible deployment, continuous fairness auditing, and clinical validation trials planned, this research represents a step toward equitable, explainable, and globally accessible AI-assisted mental health screening.

Acknowledgments

The authors express sincere gratitude to the Department of Computer Science and Engineering for providing institutional support and computational infrastructure. We thank Professor [Advisor Name] for their mentorship and insightful feedback throughout this research.

We acknowledge open-source contributors of PyTorch, HuggingFace Transformers, Scikit-learn, Librosa, and related frameworks for enabling efficient experimentation. Access to high-quality datasets—via OSMI, Reddit's public mental-health corpus, and academic repositories—was crucial for benchmarking.

We also thank clinical experts who provided valuable feedback on mental-health assessment design and ethical considerations. Finally, we dedicate this work to individuals and families affected by depression worldwide, whose resilience inspires innovation toward accessible and stigma-free mental-health technology.

References

1. World Health Organization: *Depressive disorder (depression).* WHO Fact Sheets, March 2023. `https://www.who.int/news-room/fact-sheets/detail/depression`
2. Santomauro, D. F., et al.: Global prevalence and burden of depressive and anxiety disorders in 204 countries due to the COVID-19 pandemic. *The Lancet*, 398(10312), 1700–1712 (2021)
3. Kroenke, K., Spitzer, R. L., Williams, J. B.: The PHQ-9: Validity of a brief depression severity measure. *J. Gen. Intern. Med.*, 16(9), 606–613 (2001)
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proc. NAACL-HLT*, pp. 4171–4186 (2019)
5. Liu, Y., Ott, M., Goyal, N., et al.: RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692* (2019)
6. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT: A distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019)
7. Alghowinem, S., et al.: From joyous to clinically depressed: Mood detection using spontaneous speech. In: *Proc. FLAIRS Conf.*, pp. 141–146 (2013)
8. Morales, M., Levitan, R.: Speech vs. text: A comparative analysis of features for depression detection systems. In: *Proc. IEEE SLT*, pp. 634–641 (2022)
9. Obermeyer, Z., et al.: Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453 (2019)