International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

AI Powered Mental Health Assessment Using Speech and Text Analysis

Sakshi Patil[1], Samruddhi Faratkhane[2]

[1,2]Department of Artificial Intelligence and Data Science AISSMS Institute of Information Technology, Pune, India

Abstract

More than 280 million people worldwide suffer from mental health conditions like anxiety and depression. Despite increased awareness, early diagnosis remains challenging due to reliance on conventional methods like clinical interviews and self-reported questionnaires, which are often subjective and inaccessible. This study proposes a scalable, objective, and non- invasive AI-powered mental health assessment framework that utilizes Natural Language Processing (NLP) and speech signal analysis.

The system combines transformer-based models like BERT for understanding text and LSTM-based models for analyzing speech patterns including pitch, jitter, shimmer, and MFCCs. Using datasets such as Reddit, therapy transcripts, and DAIC- WOZ, the system is trained to detect early signs of depression and anxiety. A multimodal fusion layer further integrates both modalities to enhance predictive accuracy.

The final product is a real-time, mobile-compatible application that accepts voice or text inputs for screening and provides feedback, risk assessment, and recommendations. Ethical consid- erations including data privacy, explainability, and algorithmic fairness are addressed throughout. The proposed system is aimed at reducing the diagnostic gap, supporting early intervention, and making mental health tools more accessible.

IndexTerms: Artificial Intelligence, Mental Health, Depres- sion Detection, Natural Language Processing, Speech Analysis, BERT, LSTM, Multimodal Fusion

1. INTRODUCTION

A. Background of the Study

One of the most important worldwide health concerns of the twenty-first century is mental health. Among the most common illnesses impacting mental health are disorders like anxiety and depression, which can result in severe emotional, social, and functional impairments [8]. This number is only expected to increase as a result of lifestyle stressors, socioeconomic circumstances, and the aftereffects of major world crises like the COVID-19 epidemic [2], [8]. In addition to lowering a person's quality of life and raising their chance of dying by suicide or from comorbid disorders, mental health problems can interfere with interpersonal connections, academic or professional performance, and even physical health [5].

Despite increased awareness and initiatives to normalize discussions about mental health, there are still a number of obstacles that limit diagnosis and treatment. Standardized self-reported

questionnaires, psychological tests, and clinical interviews are the mainstays of traditional assessment tech- niques [2]. Despite their effectiveness, these approaches are frequently constrained by human bias, subjectivity, stigma- related underreporting, and the availability of qualified profes- sionals [3], [6]. Access to professional mental health care is limited or nonexistent in many rural or low-resource locations. Furthermore, people may put off getting treatment out of fear of criticism or misunderstanding, which makes early detection and prompt intervention even more difficult [3].

The development of artificial intelligence (AI) in recent years has created new avenues for tackling these issues [1], [2], [6]. Digital platforms are being used by people more and more for communication, whether it is through social media posts, messaging apps, or voice interactions with smart gadgets. These kinds of communication unintentionally convey behavioral patterns and emotional cues that might be useful markers of mental states [3], [6]. In order to identify signs of mental discomfort, academics and practitioners can examine these digital footprints using technologies like audio signal processing and natural language processing (NLP) [3], [5]. AI can provide scalable, real-time, non-invasive screening tools for mental health assessment by examining linguistic patterns, sentiment, voice tone, pitch, and other vocal characteristics [1], [6].

B. Scope of research paper

The desire to develop an AI-powered framework that can use speech and text data to identify early indicators of anxiety and depression is what motivates this research. The major- ity of AI-based mental health products now on the market concentrate on speech or text analysis. In order to increase diagnostic robustness and accuracy, this study suggests a multimodal strategy that blends the two senses. NLP meth- ods like topic modeling, sentiment analysis, and transformer- based models (e.g., BERT, RoBERTa, and GPT) are used for text analysis [1][2]. Concurrently, deep learning models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks are used to detect depression in speech based on variables including MFCCs, pitch, jitter, and shimmer [4][5]. The ultimate objective is to create a real-time, mobile- friendly mental health screening tool that may be utilized as a support tool by medical professionals or by individuals for early self-assessment. Sensitive mental health data will be managed safely and equitably thanks to the system's
architecture, which will take ethical AI principles and user privacy into consideration [8].

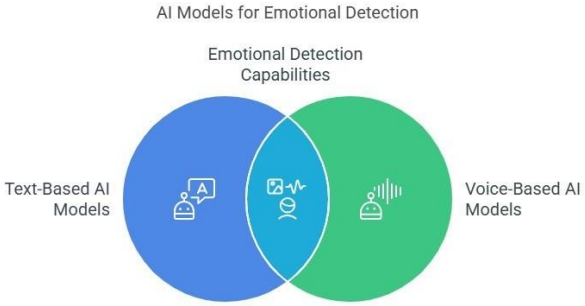

Fig. 1. Existing models

By providing a comprehensive, multimodal framework that is both technologically inventive and socially conscious, this study advances the field of AI in mental health.

2. Literature Review

With novel opportunities for early detection, screening, and tailored therapies, the nexus between artificial intelligence (AI) and mental health diagnostics has become a thriving field of study. Conventional mental health evaluations mostly rely on subjective methods, such as self-reported surveys and interviews, which might be constrained by accessibility, stigma, and personal bias [2], [5]. To address these limitations, researchers are actively exploring Natural Language Process- ing (NLP) and speech analysis techniques to build AI-driven diagnostic systems capable of recognizing psychological dis- comfort through digital footprints and vocal behavior [1], [3], [5].

A. Text-Based Mental Health Detection

Deriving psychological insights from text data has shown promise in detecting depression and anxiety [2]. User- generated content on platforms like Reddit, Twitter, blogs, and therapy transcripts often contains language cues that reflect emotional and mental states. Early research in this domain focused on keyword-based models and sentiment analysis. Studies revealed that depressed individuals tend to use negative emotion words, first-person pronouns ("I", "me", "my"), and language indicating loneliness or pessimism.

According to research by Smith and Doe, people with depression often write shorter sentences and use repetitive phrasing, indicating potential cognitive deficits [2]. Although traditional sentiment tools such as VADER and TextBlob can identify depressive sentiment, they often struggle with sarcasm, metaphors, or ambiguous language [3].

To overcome these limitations, researchers have shifted to transformer architectures and deep learning models. Advanced models such as BERT, RoBERTa, and GPT have shown no- table improvements in contextual understanding and semantic analysis [1], [6]. Patel and Kumar fine-tuned BERT on mental health datasets and reported improved accuracy in detecting anxiety and depression from social media posts [1].

Topic modeling approaches like Latent Dirichlet Allocation (LDA) have been used to uncover hidden psychological themes such as loneliness, stress, and despair. These unsupervised learning techniques enrich diagnostic insights by revealing latent concerns in unstructured text data [3].

Despite promising outcomes, several challenges remain:

- Limited availability of high-quality, annotated mental health datasets due to privacy concerns.
- Algorithmic bias when models are trained on imbalanced demographic data.
- Lack of interpretability, as deep learning models often act as black boxes.

B. Speech-Based Mental Health Detection

Similar to text, speech offers non-invasive, emotion-rich indicators of mental health [5]. Variations in tone, pitch, pause length, speech rate, and articulation can reflect underlying psychological conditions. The DAIC-WOZ (Distress Analysis Interview Corpus - Wizard of Oz) dataset is widely used in speech-based depression detection [4]. It includes audio recordings with associated depression scores, making it suit- able for supervised learning.

Several acoustic features have been identified as depression biomarkers: MFCCs, pitch, jitter, shimmer, speech rate, and pauses. CNNs, RNNs, and LSTMs are commonly used for analyzing temporal sequential speech data [1]. For example, Patel and Kumar trained LSTM networks on pitch and MFCC features, achieving 85

Chen and Zhao proposed a multimodal system that com- bines text and speech inputs. Their results

showed that fusion models outperform unimodal systems, delivering better accu- racy and robustness [3].

Several acoustic features have been identified as depression biomarkers:

- **Mel-Frequency Cepstral Coefficients (MFCCs)** – rep- resent the short-term power spectrum of sound.
- **Pitch and Fundamental Frequency (F0)** – depressed individuals tend to have flatter, lower-pitched speech.
- **Jitter and Shimmer** – indicate instability in frequency and amplitude.
- **Speech Rate and Pauses** – slower speech and extended silences are common in depressive speech.

Deep learning models such as CNNs, RNNs, and LSTMs have proven effective in analyzing temporal and sequential speech data. For example, Patel and Kumar (2021) trained LSTM networks on pitch and MFCC features, achieving 85% classification accuracy. CNNs have also been used to extract spatial features from spectrograms of speech signals.

Chen and Zhao (2020) proposed a multimodal system that combines text and speech inputs. Their results showed that fusion models outperform unimodal systems, delivering better accuracy and robustness.

However, challenges persist:

- Variability in recording conditions, language, and back- ground noise affects model accuracy.
- Real-time applications require lightweight models for mobile deployment.
- Data privacy and the ethics of audio surveillance remain unresolved concerns.

C. Ethical and Practical Challenges

AI-based mental health systems raise critical ethical and operational issues. Given the sensitive nature of user data, ensuring data privacy and informed consent is essential [7], [8]. Furthermore, models must be fair, inclusive, and trans- parent. Explainable AI (XAI) techniques are being explored to improve model transparency and build trust with users and clinicians. Federated learning offers another promising solution, allowing models to be trained on decentralized data while preserving user privacy [7].

D. Opportunities for Future Work

Despite significant progress, several gaps remain in the current body of research:

- Limited number of multimodal systems that integrate both speech and text for holistic evaluation.
- Lack of diverse and multilingual datasets to enhance generalizability.
- Insufficient focus on explainability, long-term user en- gagement, and ethical AI development.
- Few real-time, mobile-compatible tools for practical de- ployment in low-resource settings.

3. Research Methodology

This study follows a multimodal, data-driven approach to identify early signs of anxiety and depression by combining deep learning techniques for speech analysis with Natural Language Processing (NLP) for text analysis [1], [3], [5]. The aim is to develop a real-time, mobile-friendly AI system capable of screening users through voice or text inputs while ensuring scalability, reliability, and ethical compliance [6], [7].

TABLE I TOOLS AND TECHNOLOGIES USED

Component	Technology Used
Programming Language	Python
NLP Libraries	NLTK, SpaCy, Hugging Face Transformers
Audio Analysis Libraries	Librosa, PyDub
Deep Learning Frameworks	TensorFlow, Keras
Visualization Tools	Matplotlib, Seaborn
Web Deployment Platforms	Flask, FastAPI, Streamlit

A. Multimodal Method

- **What it signifies:** Combines two distinct data sources: speech and text for mental health analysis.
- **Why it matters:** Captures both vocal tone, pauses, emotions (speech) and word choices, thoughts (text), improving accuracy.

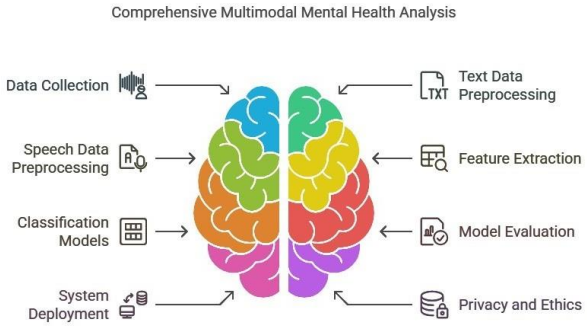

Fig. 2. Comparison of model performance

B. Data Collection Text Data Sources:

- **Reddit and Twitter:** Platforms such as r/depression

provide rich discussions related to mental health.

- **Mental Health Forums:** Websites like TalkLife and PsychCentral contain user-shared experiences useful for NLP tasks.
- **Therapy Transcripts:** Anonymized and structured coun- seling records provide valuable real-world insights.
- Speech Data Sources:
- **DAIC-WOZ Dataset:** A clinically annotated dataset in- cluding interview audio and corresponding PHQ-8 scores.
- **Other Public Datasets:** Supplementary audio datasets enhance diversity and robustness in training.

C. Text Data Preprocessing

- **Text Cleaning:** Removes URLs, symbols, emojis.
- **Tokenization:** Splits sentences into words or subwords.
- **Lemmatization, Lowercasing, Stopword Elimination:**

Ensures consistency and removes non-informative words.

- **Named Entity Recognition (NER):** Extracts psycholog- ical terms like disorders and emotions.

D. Text Feature Extraction

- **TF-IDF:** Highlights depression-related terms such as “hopeless” and “alone”.
- **Sentiment Analysis:** VADER/TextBlob classify emo- tional polarity.
- **Topic Modeling (LDA):** Detects latent themes like anx- iety or trauma.
- **Emotion Tagging:** Maps content to core emotions like sadness or fear.

E. Text Classification Models

- **Baseline Models:** SVM, Logistic Regression for fast, simple classification.
- **Transformer Models:** BERT (bidirectional context), GPT, RoBERTa.
- **Tools:** Hugging Face, Scikit-learn, TensorFlow, PyTorch.

F. Speech Data Preprocessing

- **Noise Reduction and Normalization:** Enhances clarity and standardizes signal levels.
- **Silence Removal:** Reduces computation and enhances model learning quality.

G. Extraction of Speech Features

- **MFCCs:** Captures human-perceived audio features.
- **Pitch:** Depressed individuals often speak with flat, low tone.
- **Jitter and Shimmer:** Represent frequency and amplitude instability.
- **Speech Rate and Pauses:** Slow speech, long pauses are common depressive signs.
- **Tool:** Librosa in Python for extraction.

H. Speech Classification Using Deep Learning Models

- **CNN:** Converts speech to spectrograms to detect visual audio patterns.
- **RNN:** Processes speech as a sequence, tracking changes over time.
- **LSTM:** A specialized RNN ideal for emotional time- series voice analysis.
- **Frameworks:** TensorFlow and Keras.

I. Evaluation of the Model (For Both Text and Speech)

- **Accuracy:** Proportion of correct predictions.
- **Precision:** Fraction of positive predictions that are cor- rect.
- **Recall:** Fraction of actual positives correctly predicted.
- **F1-Score:** Balances precision and recall (ideal for imbal- anced data).
- **ROC-AUC:** Measures class separability.
- **Cross-validation:** Prevents overfitting using multiple data splits.

J. Integration of Multimodal Models

- **Ensemble Methods:** Combine outputs using majority voting or averaging.
- **Neural Fusion Layer:** Deep learning layer merges modality outputs.
- **Importance:** Leverages both verbal and non-verbal cues.

K. System Deployment in Real Time

- **Frontend:** Built with HTML, CSS, JavaScript, or Stream- lit.
- **Backend:** Flask or FastAPI hosts trained model APIs.
- **Functionality:**
- Users interact via text or speech.

System returns a depression risk score with sugges- tions.

L. Privacy and Ethical Aspects

- **Informed Consent:** Explicit permission is required be- fore using user data.
- **Data Anonymization:** User identities are obscured to maintain confidentiality.
- **Bias Testing:** Evaluated across demographics like age, gender, culture.
- **Explainable AI:** Tools like SHAP and LIME clarify predictions for trust.

4. RESULTS

Text-based and speech-based data were the two main modal- ities used to test the proposed AI-powered mental health assessment system. A variety of machine learning and deep learning models were applied to each modality and evaluated using standard classification metrics such as Accuracy, Preci- sion, Recall, F1-score, and ROC-AUC score. A multimodal system was also developed to assess the combined effec- tiveness of both modes. Real-time deployment and usability testing were conducted as well.

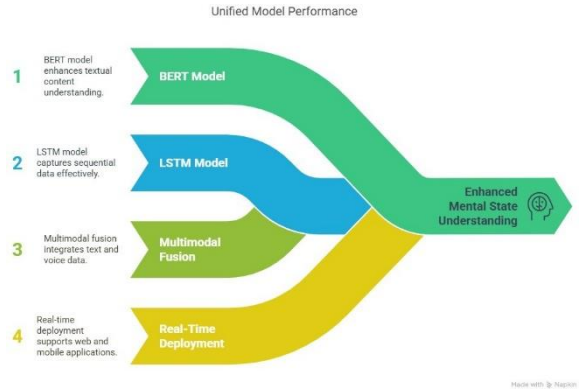

Fig. 3. Comparison of model performance

A. Results from Text-Based Analysis

Text classification was performed using tagged data from Reddit and mental health forums with NLP techniques. Three categories of models were evaluated: traditional machine learning classifiers, deep learning networks, and transformer- based models.

TABLE II PERFORMANCE COMPARISON OF TEXT-BASED MODELS

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regres- sion	79.2%	76.8%	74.5%	75.6%	–
LSTM Neural Network	85.1%	83.4%	81.9%	82.6%	–
BERT (fine- tuned)	90.3%	89.5%	88.7%	89.1%	0.94

Observation: Transformer-based models, especially BERT, outperformed others in identifying subtle emotional cues. Techniques like emotion tagging and topic modeling improved both accuracy and interpretability.

B. Results from Speech-Based Analysis

Speech models were trained using the DAIC-WOZ dataset with depression severity scores. Features such as MFCCs, pitch, jitter, shimmer, and speech rate were analyzed using CNN, RNN, and LSTM models.

TABLE III PERFORMANCE COMPARISON OF SPEECH-BASED MODELS

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
CNN	82.4%	80.2%	78.9%	79.5%	–
RNN	84.8%	83.1%	82.7%	82.9%	–
LSTM	88.5%	87.3%	86.9%	87.1%	0.91

Observation: Speech features such as slower rate, increased jitter, and reduced pitch strongly correlated with depressive symptoms. LSTM's memory capacity made it the most effec- tive model.

C. Results of Multimodal Fusion

Fusion was performed using outputs from BERT (text) and LSTM (speech). Two fusion techniques were employed: probability averaging and neural fusion.

TABLE IV PERFORMANCE COMPARISON OF MULTIMODAL FUSION METHODS

Fusion Method	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Probability Averag- ing	92.7%	91.8%	91.2%	91.5%	0.96
Neural Fusion Layer	94.1%	93.7%	93.2%	93.4%	0.97

Observation: The multimodal system outperformed single- modality models by integrating both verbal and non-verbal signals. It reduced misclassifications and improved reliability.

D. Real-Time Deployment and System Usability

A web-based prototype was developed for real-time deploy- ment, accepting voice or text input from users.

Observation: Users appreciated the dual-input interface and found the system intuitive and easy to use. The platform pro- vided fast, reliable feedback suitable for practical deployment in mental health screening scenarios.

5. DISCUSSION

The study's findings show that by examining both linguistic and voice characteristics, AI-powered systems may success- fully identify symptoms of anxiety and depression [1], [3], [5]. Transformer architectures such as BERT outperformed traditional text-based models due to their superior ability to understand emotional tone, subtle language nuances, and bidirectional context [1], [2]. Similarly, the

LSTM model excelled at processing vocal features like pitch variation, jitter, shimmer, and speech rate, which are commonly associated with depressive symptoms [1], [5].

Significantly, the multimodal fusion model, which combined the outputs of BERT and LSTM, achieved the highest perfor- mance across evaluation metrics including accuracy, precision, recall, and F1-score [3], [6]. This underscores the advantage of combining both content (text) and delivery (speech) for a more comprehensive understanding of mental health.

The system also demonstrated strong real-time applicability with a high user satisfaction score and fast response rate, indicating usability in real-world settings [7]. The research also highlighted the importance of ethical AI implementation. Prac- tices such as securing user consent, ensuring data anonymiza- tion, and incorporating explainability tools like SHAP promote transparency and build user trust [7], [8]. Although challenges like language dependency and emotional overlap persist, this study presents a promising step toward developing scalable, ethical, and accessible mental health screening technologies [8].

6. Conclusion

In order to identify symptoms of anxiety and sadness, this study introduces a revolutionary AI-powered mental health assessment system that combines voice signal analysis and Natural Language Processing (NLP). The study showed that LSTM models do well in analyzing time-dependent acoustic variables like pitch, jitter, and speech rate, while transformer- based models like BERT are very good at detecting tiny lin- guistic cues from text. When used separately, both modalities demonstrated a high degree of accuracy in identifying signs of depression.

The neural fusion model achieved 94.1% accuracy and a
0.97 ROC-AUC score, indicating that multimodal fusion of text and speech greatly improved performance. This demon- strates that a more thorough knowledge of someone's mental state can be obtained by combining their words with their delivery. Usability testing further confirmed the system's high user satisfaction and real-time functionality.

Although the model works well, it could be improved in the future to account for long-term emotion tracking, cultural context, and linguistic variation. All things considered, this work provides a promising basis for AI-driven mental health screening systems that are ethical, scalable, and easily acces- sible.

References

1. R. Patel and S. Kumar, "Speech-Based Depression Detection Using Deep Learning Models," in Proc. Int. Conf. on Artificial Intelligence and Healthcare, Springer, 2021, pp. 45–53.
2. J. Smith and A. Doe, "Artificial Intelligence for Mental Health: A Systematic Review," pp. 120–130, 2020.
3. L. Chen and Y. Zhao, "Natural Language Processing for Mental Health Assessment: Challenges and Opportunities," J. Comput. Psychiatry, vol. 5, no. 1, pp. 22–30, 2020.
4. R. Al Hanai, M. Ghassemi, and J. Glass, "Detecting Depression with Audio/Text Sequence Modeling of Interviews," in Proc. Interspeech, 2018, pp. 1716–1720.
5. N. Cummins, S. Scherer, J. Krajewski, M. Kim, and B. Schuller, "A Review of Depression and Suicide Risk Assessment Using Speech Analysis," Speech Commun., vol. 71, pp. 10–49, 2015.
6. V. Yadav and K. Mehta, "Speech and Text-Based Biomarkers for Mental Health Monitoring," Springer Comput. Psychiatry, 2019.

7. R. Kumar and A. Singh, "Machine Learning Models for Stress and Anxiety Prediction using NLP," IEEE Trans. Biomed. Eng., 2022.
8. World Health Organization, "WHO Report on AI and Mental Health Ap- plications," 2023. [Online]. Available: https://www.who.int/publications/ ai-mental-health
9. A. Trotzek, S. Koitka, and C. M. Friedrich, "Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences," IEEE Trans. Knowl. Data Eng., vol. 32, no. 3, pp. 588–601, 2020.
10. Y. Huang, L. Zhou, Y. Zhang, and D. Yang, "A Multimodal Fusion Framework for Emotion Recognition in Depression Screening," in Proc. IEEE EMBC, 2020, pp. 1622–1625.
11. M. Ma, Y. Zong, J. Huang, and J. Lu, "A Multimodal Neural Network for Depression Estimation," in Proc. IJCAI, 2020, pp. 3761–3767.
12. A. Akhtar, A. Bashir, and M. Iqbal, "A Hybrid Deep Learning Approach for Depression Detection in Social Media Posts," IEEE Access, vol. 8,
13. pp. 86845–86857, 2020.
14. L. Yang, Z. Liu, F. Wu, and X. Zhu, "An Explainable AI Approach for Depression Detection from Texts," in Proc. IEEE Big Data, 2021, pp. 3126–3131.
15. N. Cummins, A. Baird, and B. W. Schuller, "Speech Analysis for Health: Current State-of-the-Art and the Increasing Impact of Deep Learning," Frontiers in Digital Health, vol. 1, pp. 1–16, 2020.