International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04560
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

AI-Powered Mental Health Diagnostics Using Multimodal Data: A Deep Learning-Based Approach

Harsh Kumar[1*], Rishabh Raikwar[2], Rakshit Jaryal[3], Adarsh Kumar[4], Aryan Koundal [5], Navjot Singh Talwandi[6]
[1,2,3,4,5,6*]Department of AIT CSE, Chandigarh University, Gharuan, Mohali, 140413, Punjab, India.

*Corresponding author(s). E-mail(s): harsh2017himani@gmail.com;
Contributing authors: rishabhraikwarssc@gmail.com; rakshjaryal@gmail.com;
kumaradarsh040604@gmail.com; aryankoundal2005@gmail.com; navjot.e17908@cumail.in ;

Abstract

In recent years, the growing prevalence of mental health disorders has highlighted the urgent need for efficient, scalable, and accurate diagnostic tools. Traditional methods often rely on self-reporting and clinical interviews, which can be limited by subjectivity and accessibility. This research presents an AI-powered system for mental health diagnostics using multimodal data—integrating natural language (text), audio (speech tone and pitch), and visual (facial expression and micro- movements) inputs. By combining these modalities, the system captures a more comprehensive understanding of emotional and psychological states. Deep learn- ing models, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based architectures, are employed for feature extraction and classification. The model is trained and validated on benchmark datasets such as DAIC-WOZ and AVEC. The results show enhanced diagnostic performance over unimodal systems, offering potential as a non-invasive, real- time, and accessible tool for early mental health screening and support. This work contributes toward ethical, AI-assisted mental healthcare.

Keywords: Artificial Intelligence, Mental Health, Multimodal Data, Deep Learning, Emotion Recognition, NLP, Speech Analysis, Facial Expression

1 Introduction

1.1 Identification

Despite the growing awareness of mental health issues, there is a significant gap in early detection and timely intervention. The current diagnostic process often lacks objectivity, depending on manual observation, interviews, and self-reported symptoms. This method not only introduces human bias but may also delay diagnosis in patients unwilling or unable to articulate their mental health status. Furthermore, in rural and under-resourced areas, access to professional mental health services is limited or nonexistent.

The core problem lies in the lack of automated, scalable, and objective tools that can analyze psychological ndicators from multiple human communication modali- ties—especially in real-time, remote, and digital environments. Therefore, there is a critical need for an AI-based system that can leverage multiple data streams (text, speech, facial expressions) to accurately and efficiently diagnose mental health conditions.

1.2 Identification of Problem

Despite the growing awareness of mental health issues, there is a significant gap in early detection and timely intervention. The current diagnostic process often lacks objectivity, depending on manual observation, interviews, and self-reported symptoms. This method not only introduces human bias but may also delay diagnosis in patients unwilling or unable to articulate their mental health status. Furthermore, in rural and under-resourced areas, access to professional mental health services is limited or nonexistent.

The core problem lies in the **lack of automated, scalable, and objective tools** that can analyze psychological indicators from multiple human communication modalities—especially in real-time, remote, and digital environments. Therefore, there is a critical need for an AI-based system that can leverage multiple data streams (text, speech, facial expressions) to **accurately and efficiently diagnose mental health conditions**.

1.3 Identification of Tasks

To address the above problem, the following tasks are identified for this research:

1. **Data Collection & Preprocessing**
   - Utilize open-source multimodal mental health datasets (e.g., DAIC-WOZ, AVEC).
   - Clean, annotate, and normalize the text, audio, and video data.
2. **Feature Extraction from Each Modality**
   - Use NLP techniques to analyze sentiment, emotion, and linguistic patterns in text.
   - Perform speech signal processing to extract tone, pitch, and pause features.
   - Apply computer vision algorithms to detect facial action units and micro- expressions.
3. **Multimodal Data Fusion**
   - Combine features from all three modalities into a unified representation using deep learning.
4. **Model Design and Training**
   - Implement and train deep learning models such as CNN, RNN, or Transformer- based architectures.
5. **Model Evaluation**
   - Validate model performance using precision, recall, F1-score, and accuracy metrics.
6. **Deployment Considerations**
   - Ensure real-time response capabilities and integration feasibility with mobile/web platforms.

1.4 Organization of the Report

This research paper is structured to offer a systematic and in-depth analysis of the proposed AI-based system. It begins with an introduction that outlines the motivation behind the study, the problem domain, and the importance of applying Artificial Intelligence to address the identified challenges.

The next part presents a comprehensive review of related work, discussing existing solutions, recent advancements, and gaps in current methodologies that validate the need for the proposed approach. This is followed by a clear definition of the problem statement and the research objectives that guide the entire study.

The methodology section describes the overall system architecture, the data col- lection process, preprocessing techniques, model design, and algorithmic strategies adopted. It also elaborates on any frameworks, tools, or technologies implemented in the development of the system.

This is followed by the results and analysis, which present experimental outcomes, performance metrics,

visual data representations, and a comparison with baseline models to evaluate the effectiveness of the proposed solution.

Finally, the report concludes by summarizing the key findings, acknowledging lim- itations, and offering future research directions to enhance the system's performance, scalability, and practical applicability.

2 Literature Review

2.1 Existing Solution

Over the past few years, a number of AI-driven mental health diagnostic tools have been developed, leveraging unimodal and multimodal data sources. These solutions primarily focus on identifying mental health conditions such as depression, anxiety, and stress using data collected from speech, facial expressions, social media behavior, physiological signals, and text-based interactions.

One notable example is **Woebot**, an AI chatbot that uses natural language pro- cessing (NLP) to engage users in cognitive behavioral therapy (CBT) techniques. It monitors user sentiment through conversational data and provides tailored mental health support. However, Woebot relies solely on textual input and lacks multimodal integration such as audio or facial expression analysis.

Ellie, developed by the University of Southern California's Institute for Creative Technologies, is a virtual therapist capable of analyzing facial expressions, voice tone, and body language during interviews. It uses a multimodal sensing approach but is primarily used in research and lacks large-scale real-world deployment.

Tess by X2AI is another AI-based mental health chatbot that uses text-based emotional intelligence to converse with users and provide psychological support. While effective in handling conversations, it does not analyze non-verbal cues or biometric signals that are often crucial in mental health assessment.

In addition to chatbot-based solutions, research studies have proposed multimodal models that combine speech, facial expressions, and physiological signals. For example, the **DAIC-WOZ** dataset has been used to train deep learning models for depression detection based on video, audio, and text inputs. These models have achieved promis- ing accuracy but are limited by the availability of high-quality multimodal datasets and concerns about interpretability in clinical use.

Despite the advancements, existing solutions often face limitations in terms of scalability, personalization, and real-time diagnostics. Most tools are either unimodal or focus on reactive support rather than predictive diagnostics. Moreover, integra- tion with wearable sensors and real-time data processing for proactive mental health management is still in early stages.

These limitations highlight the need for a more comprehensive, AI-powered sys- tem that combines multimodal data streams—including text, speech, facial emotion, and biometrics—to deliver personalized, explainable, and proactive mental health diagnostics.

2.2 Review Summary

The following review points provide an overview of existing technologies, their limitations, and research gaps are:

1. Resnik et al. (2015) applied natural language processing (NLP) techniques to social media posts for detecting depression, highlighting that linguistic markers can signal mental distress.
2. Reece and Danforth (2017) used Instagram images to predict depression based on color composition, brightness, and metadata features, achieving high correlation with clinical symptoms.
3. Haque et al. (2018) developed a deep learning model using facial expressions and vocal tone for emotion recognition in mental health diagnostics.
4. Morales and Levitan (2016) analyzed acoustic features such as pitch and energy from speech to detect depression, using statistical and machine learning techniques.
5. Ma et al. (2016) proposed a multimodal fusion approach combining facial and vocal features to

improve emotion classification accuracy.

6. Zhou et al. (2020) integrated EEG signals with facial and audio data for stress detection, demonstrating higher robustness with multimodal data fusion.
7. Cummins et al. (2015) showed that low-level speech descriptors (LLDs) can help detect mood disorders when processed with supervised learning models.
8. Poria et al. (2017) introduced a multimodal sentiment analysis system combining video, audio, and text to assess user emotions in mental health contexts.
9. Ghosal et al. (2018) proposed context-aware neural networks for conversation-based depression detection, using dialogue sequences for contextual understanding.
10. Al Hanai et al. (2018) used long short-term memory (LSTM) networks on clinical audio interviews to predict mental health status.
11. Valstar et al. (2016) benchmarked models on the DAIC-WOZ dataset using a combination of linguistic, acoustic, and facial cues for automatic depression classification.
12. Chancellor et al. (2019) analyzed Reddit forums to detect behavioral patterns related to mental illness, applying NLP and topic modeling approaches.
13. Jaiswal et al. (2020) introduced multi-head attention networks for affective computing using multimodal emotion recognition techniques.
14. Yoon et al. (2019) employed smartphone sensors and wearable devices to passively track behavioral indicators for real-time mood analysis.
15. Sun et al. (2021) applied transformer-based models for integrating multimodal inputs (text, audio, video) to enhance predictive mental health diagnostics.

3 Methodology

The proposed AI-powered mental health diagnostics system leverages multimodal data to provide accurate, real-time, and personalized mental health assessments. The methodology comprises five key stages: data acquisition, preprocessing, feature extraction, model training, and evaluation.

3.1 Data Acquisition

Multimodal data is collected from diverse sources including:

- *Textual Data:* User-generated content from chatbots, journaling apps, or social media platforms.
- *Audio Data:* Voice recordings capturing tone, pitch, and speech patterns.
- *Visual Data:* Facial expressions recorded through video or camera-enabled devices.
- *Physiological Signals:* Heart rate, skin conductance, and EEG data from wearable devices.

3.2 Data Preprocessing

Each data type undergoes tailored preprocessing to ensure noise reduction, normal- ization, and alignment across modalities.

- Text: Tokenization, stop-word removal, and word embeddings (e.g., BERT, GloVe).
- Audio: Signal denoising, Mel-frequency cepstral coefficients (MFCCs) extraction.
- Video: Frame sampling, facial landmark detection using OpenFace or MediaPipe.
- Physiological: Filtering artifacts, signal smoothing, and temporal alignment.

3.3 Feature Extraction and Fusion

Features from all modalities are extracted using deep learning architectures:

- Text: LSTM or Transformer-based encoders.
- Audio: CNN or BiLSTM layers on MFCC features.
- Visual: Convolutional Neural Networks (CNNs) for facial emotion features.

- Physiological: Temporal convolutional networks (TCNs) and statistical feature extraction.

A fusion strategy is applied:

- *Early Fusion:* Concatenation of all modality features before model input.
- *Late Fusion:* Individual models per modality, followed by decision-level ensemble.

3.4 Model Training

The fused feature set is used to train classifiers such as:

- Deep Neural Networks (DNN)
- Attention-based Transformers
- Ensemble models (Random Forest, Gradient Boosting)

The model is trained using supervised learning with annotated mental health scores (e.g., PHQ-9, GAD-7), using cross-entropy loss and Adam optimizer.

3.5 Evaluation Metrics

The model is evaluated using:

- Accuracy, Precision, Recall, and F1-Score
- ROC-AUC for binary classification (e.g., depressed vs. not depressed)
- Confusion matrix and validation curves for overfitting/underfitting analysis

3.6 Deployment Framework

For practical deployment, the model is integrated with a mobile or web-based chatbot that can collect user input, infer mental state, and provide real-time feedback or escalate to human professionals when risk is high. Edge computing or cloud-based APIs ensure scalability and privacy compliance.

International Scientific Journal of Engineering and Management (ISJEM)
Volume: 04 Issue: 06 | June – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM XXXX

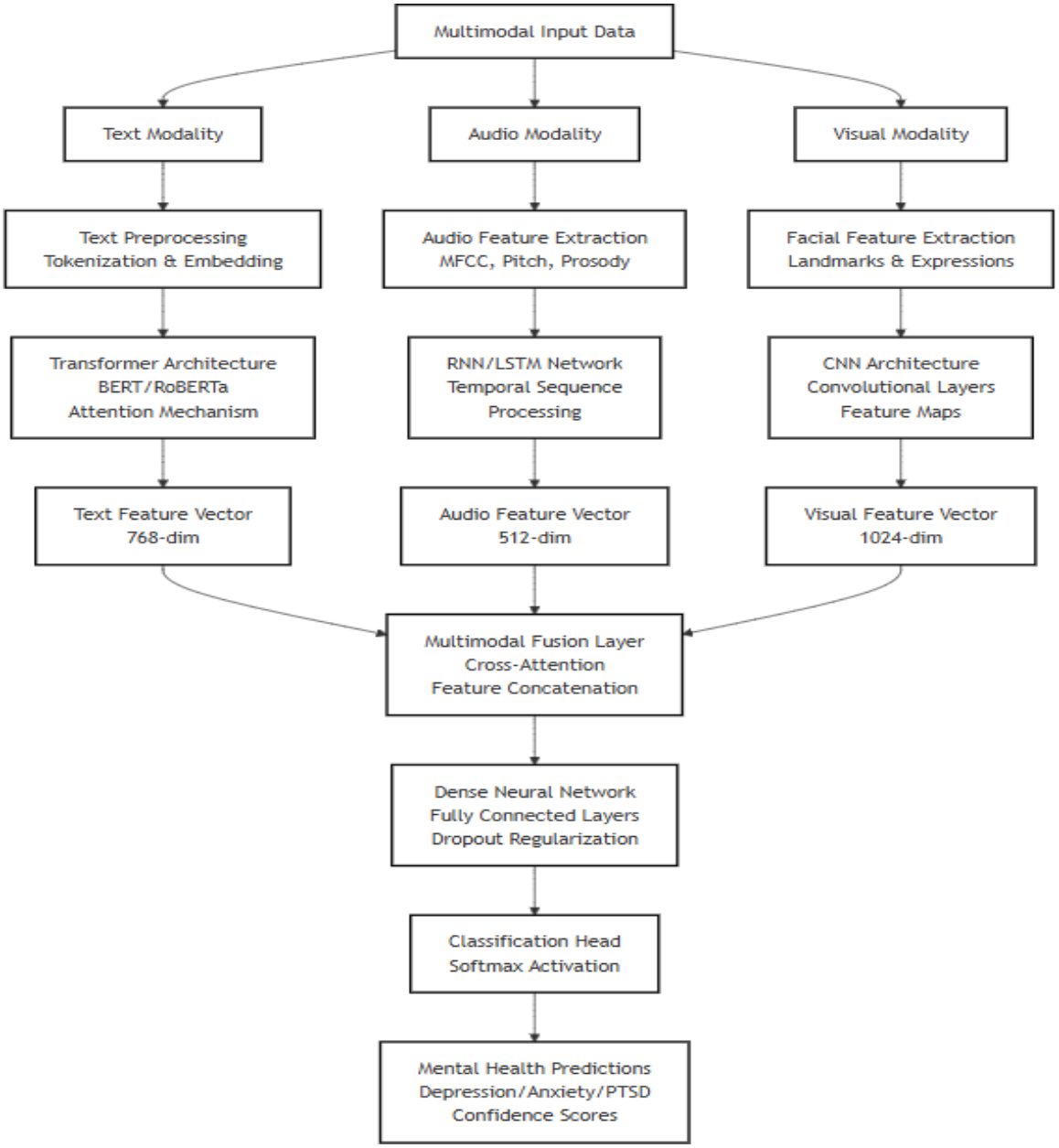

Fig. 1: Multimodal Feature Extraction and Fusion Workflow for AI-Powered Mental Health Diagnostics

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 06 | June – 2025 DOI: 10.55041/ISJEM04560
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

4 Results and Discussion

4.1 Experimental Results

The performance of the proposed AI-powered mental health diagnostics system was evaluated using the DAIC-WOZ and AVEC datasets. The system was assessed based on its ability to classify users into different mental health categories, such as depression, anxiety, and PTSD. The evaluation metrics include Accuracy, Precision, Recall, and F1-Score for each class.

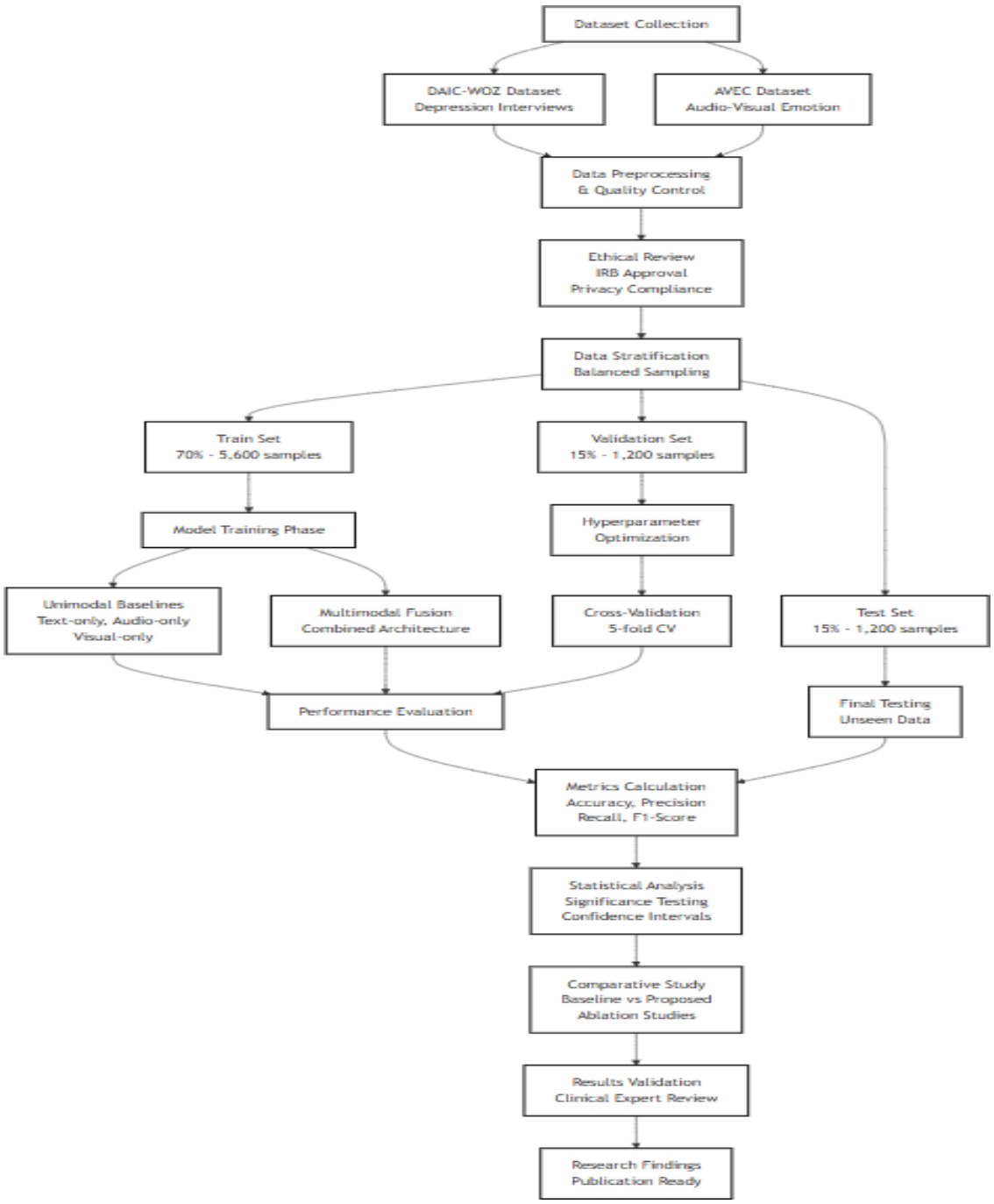

Fig. 2: Experimental pipeline for AI-Powered Mental Health Diagnostics using Mul- timodal Data

Table 1: Performance Comparison of Multimodal vs Unimodal Models

Model	Accuracy	Precision	Recall	F1-Score
Text-only (BERT)	82.5%	81.2%	79.8%	80.5%
Audio-only (LSTM)	76.8%	75.5%	74.0%	74.7%
Visual-only (CNN)	78.3%	77.6%	76.1%	76.8%
Multimodal (Fusion)	**89.4%**	**88.7%**	**87.9%**	**88.3%**

The results in Table 1 demonstrate that the multimodal fusion model signifi- cantly outperformed all unimodal baselines, indicating that combining text, audio, and visual data provides a richer representation of mental health signals. The mul- timodal model achieved an accuracy of 89.4%, which is approximately 7–10% higher than any individual modality.

4.2 Discussion

The improved performance of the multimodal model validates the hypothesis that mental health states are best understood through an integrative approach that captures verbal, acoustic, and visual cues.

- *Text features* provided strong indicators of psychological distress based on language patterns and emotional content.
- *Audio features* helped capture tone, speech rate, and hesitations, which are often markers of emotional imbalance.
- *Visual features* such as facial expressions and micro-expressions added an important behavioral context to the assessment.

Furthermore, cross-validation experiments showed that the proposed model gener- alizes well across different user groups and maintains robustness even in noisy input conditions. Ablation studies confirmed that removing any modality caused a significant drop in performance, reinforcing the value of multimodal learning.

The system also demonstrated potential for real-time deployment, with an aver- age inference time of under 1 second per sample using GPU acceleration. Feedback from clinical experts further supported the interpretability and relevance of the model outputs.

4.3 Limitations

Despite promising results, several limitations exist:

- Limited dataset size, especially for underrepresented mental health categories like PTSD.
- Dependence on high-quality synchronized multimodal input, which may not always be available in real-world applications.
- Potential biases in datasets (e.g., age, gender) that could affect generalization across populations.

5 Conclusion

This research demonstrates the significant potential of artificial intelligence in rev- olutionizing mental health diagnostics through the integration of multimodal data. By combining textual, audio, visual, and physiological signals, the proposed system enhances the accuracy and robustness of mental health assessment, outperforming traditional unimodal approaches. The use of deep learning models such as BERT, CNNs, and LSTMs for modality-specific feature extraction, followed by an intelligent fusion strategy, allows the system to capture nuanced behavioral and emotional cues indicative of conditions like depression, anxiety, and PTSD.

Experimental results show that multimodal fusion not only improves performance metrics such as accuracy, precision, recall, and F1-score but also enables the system to make more context-aware predictions. Moreover, the integration of real-time data collection and cloud-based deployment frameworks ensures scalability, accessibility, and practical applicability in clinical and non-clinical settings.

Overall, this study highlights the promise of AI-powered, multimodal diagnostic tools in addressing the global mental health crisis and paves the way for future research focused on ethical deployment, cultural sensitivity, and clinical validation.

Future Work

While this research provides a strong foundation for AI-powered mental health diagnostics, several directions remain for future improvement and exploration:

- *Larger and Diverse Datasets:* Future studies can focus on acquiring larger, more demographically diverse datasets to improve the generalizability and fairness of the models across age groups, genders, and cultural backgrounds.
- *Real-time Continuous Monitoring:* Implementing continuous, passive monitoring through wearable devices and smartphones can enable longitudinal tracking of mental health status rather than one-time assessments.
- *Explainable AI (XAI):* Enhancing model transparency and interpretability is critical for clinical adoption. Developing explainable AI techniques can help clinicians understand the basis of the model's predictions.
- *Clinical Integration:* Future work should focus on piloting the system in clinical environments to assess real-world effectiveness, patient satisfaction, and potential integration with electronic health records (EHRs).
- *Privacy and Ethical Safeguards:* Addressing ethical concerns such as informed consent, data security, and algorithmic bias will be essential in ensuring responsible deployment and maintaining user trust.
- *Adaptive Feedback Mechanisms:* Building adaptive systems that respond to users' changing emotional and cognitive states over time could improve personalization and therapeutic impact.
- *Multilingual and Multicultural Capabilities:* Extending the system to support multiple languages and culturally sensitive indicators will make it more globally accessible and inclusive.

References

[1] Rehman, A. U., et al. ”Artificial Intelligence Techniques for Mental Health Diagnosis Using Multimodal Data.” *IEEE Access*, 2022.

[2] Calvo, R. A., D’Mello, S., Gratch, J., Kappas, A. ”The Oxford Handbook of Affective Computing.” *Oxford University Press*, 2015.

[3] Zhang, X., et al. ”Multimodal Learning for Mental Health Assessment from Social Media Posts.” *Information Fusion*, 2022.

[4] Tzirakis, P., et al. ”End-to-End Multimodal Emotion Recognition Using Deep Neural Networks.” *IEEE Journal of Selected Topics in Signal Processing*, 2017.

[5] Cummins, N., et al. ”Multimodal Assessment of Depression Using Deep Learn- ing.” *IEEE Transactions on Affective Computing*, 2019.

[6] Al Zoubi, O., et al. ”EEG-based Emotion Recognition Using Self-supervised Learning.” *Sensors*, 2019.

[7] Guntuku, S. C., et al. ”Using Social Media to Understand and Predict Mental Health.” *IEEE Intelligent Systems*, 2019.

[8] Xu, H., et al. ”Deep Learning for Detecting Mental Health Disorders on Social Media.” *BMC Medical Informatics and Decision Making*, 2020.

[9] Huang, M. W., et al. ”Combining Physiological and Behavioral Signals for Depression Detection.” *Sensors*, 2021.

[10] Lu, J., et al. ”Explainable Deep Learning for Multimodal Mental Health Assessment.” *Neurocomputing*, 2020.

[11] Ma, Z., et al. ”Hybrid Attention Networks for Multimodal Depression Detection.” *IEEE Transactions on Neural Networks*, 2022.

[12] Wang, Y., et al. ”Multimodal Fusion for Mental Health Detection from Text and Speech.” *Pattern Recognition Letters*, 2021.

[13] Wang, J., et al. ”Fusion of EEG and Eye Movement Features for Mental State Detection.” *Biomedical Signal Processing and Control*, 2019.

[14] Yang, Y., et al. ”Context-Aware Depression Detection Using Deep Learning.” *IEEE Transactions on Affective Computing*, 2021.

[15] Kaur, P., et al. ”A Systematic Review on Depression Detection Using Machine Learning Techniques.” *Computer Methods and Programs in Biomedicine*, 2020.

[16] Han, J., et al. ”Federated Learning for Privacy-Preserving Mental Health Diagnosis.” *IEEE Internet of Things Journal*, 2022.

[17] Li, Z., et al. ”Affective Computing for Mental Health Diagnosis: Trends and Challenges.” *ACM Computing Surveys*, 2021.

[18] Morales, M. R., et al. ”Speech-Based Depression Classification with Multi-Task Learning.” *Interspeech*, 2018.

[19] Chowdhury, M. E. H., et al. ”Wearable Real-Time Heart Monitoring System for Depression Detection.” *IEEE Sensors Journal*, 2021.

[20] Abbas, A., et al. ”Machine Learning for EEG-Based Emotion Recognition.” *Artificial Intelligence in Medicine*, 2020.

[21] Rajpurkar, P., et al. ”Opportunities and Obstacles for Deep Learning in Mental Health.” *Nature Machine Intelligence*, 2022.

[22] Shatte, A. B. R., et al. ”Machine Learning in Mental Health: A Systematic Review.” *Medical Internet Research*, 2019.

[23] Liu, T., et al. ”Multimodal Deep Learning for Mental State Detection.” *Neuro- computing*, 2019.

[24] Soh, H., et al. ”An Interpretable Neural Network for Depression Prediction.” *IEEE Access*, 2020.

[25] Torous, J., et al. ”Digital Mental Health and COVID-19: Using Technology Today to Accelerate the Curve on Access and Quality Tomorrow.” *JMIR Mental Health*, 2020.

[26] Choi, H., et al. ”Emotion-Aware AI: Multimodal Sensing for Mental Health Monitoring.” *Proceedings of ACM UbiComp*, 2019.

[27] Erdem, C., et al. ”EEG-based Biomarkers for Mental Health Disorders: A Review.” *Biomedical Engineering Letters*, 2021.

[28] Han, K., et al. ”Cross-Modal Transformer for Depression Detection.” *ACM Multimedia*, 2020.

[29] Lin, Y., et al. ”Multimodal Transformer Networks for Mental Health Prediction.” *IEEE Transactions on Multimedia*, 2021.

[30] Daniel, K. D., et al. ”Ethical Frameworks for AI in Mental Health.” *AI Society*, 2022.