Detection and Prediction of Mental Health Disorders like Schizophrenia using AI/ML Models

1st Muhammed Sinan T V
dept. Computer Science
(of Affiliation)
TKM Collage of Engineering
(of Affiliation)
Kollam, India
mhdsinan092@gmail.com

2nd Sreya Maxvel
dept. Computer Science
(of Affiliation)
TKM Collage of Engineering
(of Affiliation)
Kollam, India
sreyamaxwel@gmail.com

Rasal Musthafa
dept. Computer Science
(of Affiliation)
TKM Collage of Engineering
(of Affiliation)
Kollam, India
rasalkpk66@gmail.com

***Abstract*— Schizophrenia is a serious mental illness that affects how people perceive, think, and manage their emotions. Diagnosing it usually relies on subjective clinical assessments, which can delay detection and cause inconsistent treatment. This study introduces an AI/ML system that combines text analysis, facial emotion recognition, and voice audio classification to detect schizophrenia early. It uses datasets like Reddit Mental Health Posts, FER-2013, and CREMA-D. The models, including Random Forest and Convolutional Neural Networks (CNN), were tested for each type of data. The text-based model performed best with an accuracy of 82.89%, while the facial and voice models scored 54.14% and 52.48%, respectively. However, when these methods were combined, the accuracy dropped to 32.30%, showing the difficulties in merging different data types in AI systems. This work highlights the promise of AI in mental health detection and points to the need for better ways to combine multiple data sources in the future.**

Keywords*— *Schizophrenia Detection, Machine Learning, Multimodal Fusion, Mental Health Prediction, Facial Emotion Recognition, Audio Analysis.

I. Introduction

Schizophrenia is a long-term mental health condition that affects millions of people around the world. Catching it early is important for better treatment, but right now, diagnosis mostly depends on personal judgment. Thanks to progress in Artificial Intelligence (AI) and Machine Learning (ML), it's now possible to diagnose more objectively using data. This study looks at a method that combines text, facial expressions, and voice signals to identify mental health issues, especially schizophrenia.

This research investigates a multimodal AI/ML-based approach for detecting early symptoms of schizophrenia by integrating data from textual analysis, facial emotion recognition, and voice analysis. The study aims to explore whether combining these modalities can enhance detection accuracy and offer a foundation for real-time mental health monitoring systems.

II. Methodology

A. Data Collection

Text: Reddit Mental Health Dataset containing posts classified into five categories (Stress, Depression, Bipolar Disorder, Personality Disorder, Anxiety).

Facial Images: FER-2013 Dataset with grayscale images of facial expressions.

Voice: CREMA-D Dataset consisting of audio recordings labelled by expressed emotions.

B. Data Preprocessing

Text: Cleaned and tokenized, TF-IDF features (top 2000 terms) extracted.

Facial Images: Resized to 48×48 pixels, normalized before feeding into CNN.

Audio: MFCC features extracted and standardized for input into CNN.

C. Model Architectures

Text Model: Random Forest (RF) classifier trained on TF-IDF vectors.

Facial Model: Convolutional Neural Network (CNN) with Conv2D, MaxPooling, and Dense layers.

Audio Model: 1D CNN trained on MFCC feature vectors.

D. Multimodal Fusion Strategy

Late fusion using weighted averaging of predicted probabilities:

$$\text{Fused\_P} = 0.6 \times \text{P}_{\text{text}} + 0.2 \times \text{P}_{\text{facial}} + 0.2 \times \text{P}_{\text{audio}}$$

Final prediction obtained using argmax over fused probabilities.

E. Evaluation Matrix

Accuracy

Precision, Recall, and F1-Score

Confusion Matrix

Receiver Operating Characteristic (ROC) and Area Under Curve (AUC)

III. Results

Modality	***Model***	***Accuracy***
Text	Random Forest	82.89%
Facial	CNN	54.14%
Voice	1D CNN	52.48%
Fused	Weighted Fusion	32.30%

A. *ROC-AUC of Fused Model*

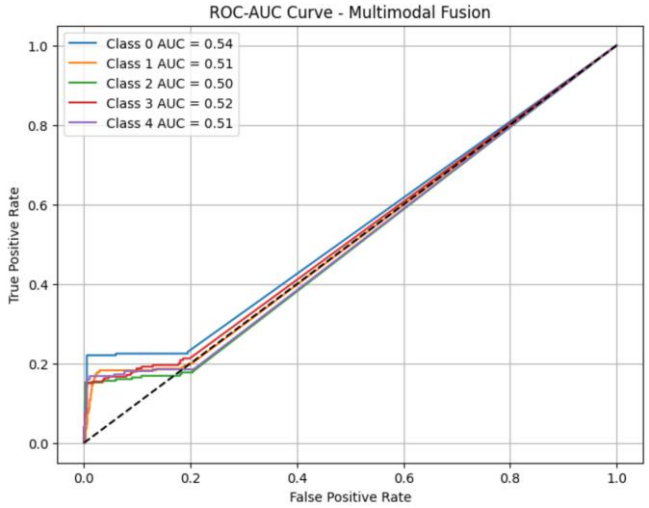

Figure 1: ROC-AUC curves for each class in the fused model.

B. *Confusion Matrix of Fused Model*

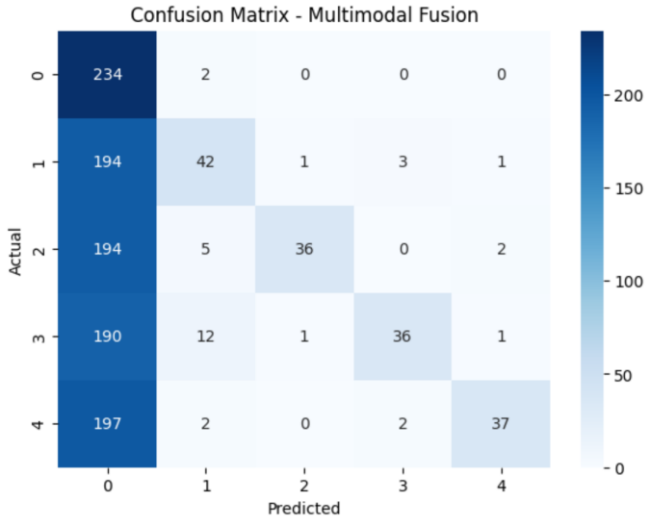

Figure 2: Confusion Matrix of Multimodal Fusion Model.

C. *Accuracy Comparison of Modalities*

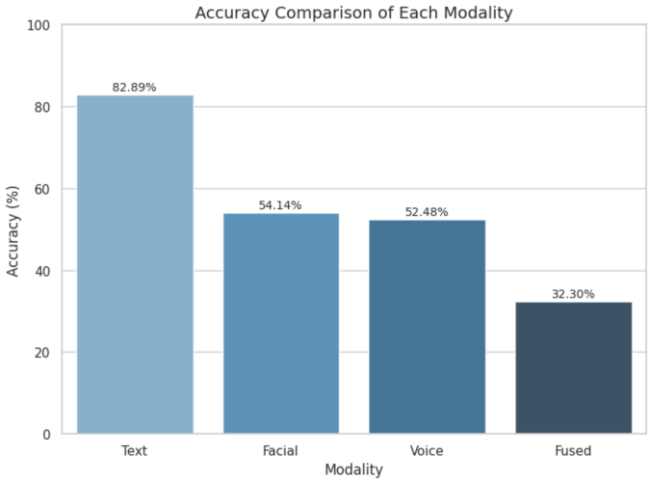

Figure 3: Accuracy Comparison of Text, Facial, Voice, and Fused Models.

IV. Discussion

Our findings show that text is the most reliable way to predict mental health, probably because it carries a lot of useful language clues. Although facial expressions and voice have some potential, they didn't perform well on their own, likely because of differences in the data and limits of the models. Combining all methods, while it sounds like a good idea, actually weakened the strong results we got from text alone because the other inputs were less reliable.

Advanced fusion methods, such as attention-based networks or using confidence scores specific to each modality, could help tackle these challenges. Also, training the facial and audio models on larger datasets and applying transfer learning can boost the system's overall performance.

V. Conclusion and Future Work

This study demonstrates the viability and challenges of AI-driven multimodal mental health detection systems. While text-based models show significant promise, the current fusion strategy does not improve performance, highlighting a key area for future exploration. Planned future work includes:

Implementing advanced deep learning models like ResNet and wav2vec.

Exploring adaptive fusion techniques based on modality confidence.

Real-time deployment through wearable devices and mobile applications.

Building personalized baseline detection models for continuous monitoring.

Acknowledgment

The authors would like to express their sincere gratitude to the IEEE Engineering in Medicine and Biology Society (EMBS) for organizing the IEEE EMBS Student Internship Program 2025, which provided a unique platform for research and innovation in the field of biomedical engineering and artificial intelligence.

We extend our heartfelt thanks to our internship coordinators and mentors for their continuous support, valuable guidance, and constructive feedback throughout the project.

We are also grateful to *TKM College of Engineering, Kerala, India*, for the academic support and resources provided during the course of this work.

Finally, we thank our peers and family members for their constant encouragement and motivation.

References

[1] Reddit Mental Health Dataset, Kaggle..

[2] FER-2013 Facial Expression Recognition Dataset.

[3] CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset.

[4] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[5] Lundberg, S.M., & Lee, S.-I., "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, 2017.

[6] Schneider et al., "wav2vec: Unsupervised Pretraining for Speech Recognition," 2019.