

Dual Sense AI for Mental Disorder Detection

Suchetha N V¹, Chaitra P Bhat², Manasa N Gond³, Bhavana G Gamad⁴, Chaitanya Chavan⁵

¹Associate Professor, Department of CSE, Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire and affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India

^{2,3,4,5} Student, Institute Department of CSE, Sri Dharmasthala Manjunatheshwara Institute of Technology, Ujire and affiliated to Visvesvaraya Technological University, Belagavi, Karnataka, India

DOI: <https://doi.org/10.51244/IJRSI.2025.120700202>

Received: 20 July 2025; Accepted: 26 July 2025; Published: 19 August 2025

ABSTRACT

Mental health conditions often show subtle and complex signs that are hard to detect using traditional tools. Proposed system is Dual-Sense AI for identifying mental disorder. It detects the face expression and speech to text to understand them better. It monitors the real time problems related to stress, anxiety, and depression. It takes the input of the images of face and speech to text for detection. With the both face dataset and the speech to text input the Dual-sense AI helps to monitor the disorders accurately by achieving 95% accuracy. It is more complete and adaptable way to monitor the mental disorders.

Keywords: Computer Vision, Real-Time Monitoring, Emotion Recognition, Speech Recognition.

INTRODUCTION

Mental disorders affect the people globally and are common in today's digital connected world. Many people continue with the traditional methods and receive the treatment when the situation is out of control. Some of the factors for this are self-report, ancestral practise, surveys and therapist interviews. The project dual sense AI introduces the approach which tracks the face expressions and speech to text input patterns. They carry the information of the human emotion through face and positive words like happy, negative words like depressed from the speech to text input and detects the person's mental state. These techniques help to make think more accurately and consistently in real life. Unlike manual methods, Dual-Sense AI works continuously and unobtrusively, providing mental disorders professionals with data-driven insights that are less prone to human error or bias.

It provides the user-friendly interface where the people can receive the early alerts about the changes in their normal emotional state which helps the clinicians to provide the treatment. The burden on the healthcare system is reduced but also supports for practise of early detection. It picturizes how the system is monitored. It helps the individual for early detection as they are relatedly Scalable and accessible.

LITREATURE SURVEY

The recent approach mainly focuses on analysing either through the expression and speech to text input. They suggest that expression recognition through the CNN and DL technique can consistently detect emotion accurately. Similarly for speech to text input using the NLP the model differentiates the positive words like happy, beautiful, and negative words like depressed, sad, and frustrated etc.

In the study titled "An Automated System for Depression Detection Based on Facial and Vocal Features" [1], the authors attempt to overcome these limitations by using measurable facial and vocal indicators as more objective markers of mental state. Researchers have explained about the facial recognition using ML technique, DL models like CNNs with transfer learning. The efficient NET and VGG16 net are architecture used to detect happiness, sad and fear from the face Emotional features and aiding behaviours are grouped and identified using

ML methods of the same is followed by speech to text input. Support vector machines and naïve bayes classifier are used in NLP processing dataset.

These models rely on extracting features from speech and applying text-based ML techniques. More recent advancements have incorporated NLP models to increase the performance of these classifications.

The research under discussion contributes by proposing a dual-modality framework that merges both facial and speech to text data for improved depression detection. For face recognition, the study evaluated DL models. Among these the CNN has achieved the accuracy at 83%. In vocal analysis NLP performed better than SVM achieving 79% accuracy. These were included into a web-based The videos are processed as application to assess depressive symptoms.

Although the output is promising, the authors acknowledge the limitations in dataset variety and size. To improve the consistency and generalizability of such systems, they recommend expanding the datasets and exploring additional ML methods across diverse populations. The article “An AI-based Decision Support System for Predicting Mental Health Disorders” [2] explores the limitations of conventional diagnostic tools like the SCL-90-R, a widely used psychological assessment instrument. Although the SCL-90-R and its variants (such as SCL-27 and BSI-18) are comprehensive which leads to low participation and inefficient assessments. Attempts to shorten these tools through statistical methods which performed well but they typically reduce the diagnostic range and precision. To overcome the study proposes a Decision Support System (DSS) powered by artificial intelligence which simplifies initial assessments without compromising diagnostic accuracy. The system focuses on improving accessibility and speed while maintaining diagnostic reliability. ML models, including SVM and neural networks, were previously used in similar systems, particularly for conditions such as depression and schizophrenia. However, they run “black boxes” lacking clarity and raising concerns about bias and interpretability. NEPAR algorithm is used for more transparent and ethical approach. It identifies the relevant variables enabling the original items. They also use the explainable ML algorithm such as Random Forest and logistic regression due to which 89% of accuracy is achieved in DSS across mental health disorders. The system is deployed for support in the digital platform. The ethical principles such as fairness, explainability and transparency are incorporated to avoid the issues in early AI based tools. In the future DL algorithm could be combined with demographic data for accuracy and adaptability. Overall, the research presents the future solutions for ethical standards. It aims to increase the mental disorder care by reducing the time of evolution in health care tools. The paper titled “Depression Recognition over Fusion of Visual and Vocal Expression Using Artificial Intelligence” [3] explores the face and vocal cues for identifying depression through AI-based systems. This dual approach is important given that many cases of depression remain undetected, often due to societal stigma and the limitations of traditional diagnostic tools like self-reported questionnaires. Previous studies have attempted to detect depression using various data modalities, applying ML techniques such as Convolutional Neural Networks, Long Short-Term Memory networks, and PCA-SVM hybrid models. These models were trained on well-known datasets such as DAIC-WOZ, AVEC, and SEMAINE. Notably, audio-text-based models built on the DAIC-WOZ dataset good performance than models that relied only on single modalities, while video data models trained on AVEC achieved around 70% accuracy. However, many of these approaches faced challenges like limited population diversity, dataset-specific tuning, or the constraints of using only one input type (i.e., unimodal analysis). The study proposes a multimodal framework that integrates visual and vocal data using a combination of CNNs for facial analysis and LSTMs for speech processing. The CNN models extract visual features using techniques like Histogram of Oriented Gradients (HOG), while the LSTM network analyses Mel-Frequency Cepstral Coefficients (MFCC) from speech for emotion and mood detection. The system calculates probabilities from both channels and fuses them to improve accuracy and decision-making reliability. Results from the proposed model show a significant improvement over single-modality system. Specifically, the model achieved 82% accuracy in detecting depression through audio analysis and 93% through visual inputs. This suggests that combining visual and vocal indicators provides a more complete, picture leading to better recognition of depressive states. By automating the analysis process, the system also makes early detection more accessible and scalable. This could reduce barriers to seeking help and encourage earlier intervention. It suggests that behavioural features, such as response latency and pause frequency in diagnostic performance. The study titled “Facial Expression Recognition System for Stress Detection with Deep Learning” by José Almeida and Fátima Rodrigues [4] Detect the stress through face expressions using DL techniques. The authors developed

a CNN model trained via transfer learning to classify expressions into several emotional categories. The testing dataset of CNN has 89.6% of accuracy.

To evaluate its effectiveness in detecting stress, the authors grouped certain expressions—specifically anger, disgust and fear—as indicators of stress, while the remaining expressions were considered non-stress-related. When tested on this classification task, the model has a higher accuracy of 92.1%, showing the face expression detection as a tool for stress detection. Despite these promising results, the authors acknowledged several limitations. Firstly, the dataset the model did not contain explicit stress or non-stress labels. Instead, the classification relied on inferred relationships between certain emotions will not hold true due to other things like mental fatigue etc. Secondly the inputs are precise which could impact the model's generalizability to new or varied input data.

To address these limitations, the researchers plan to enhance their system in future work. This means gathering user feedback and detecting the accuracy of stress in real-world scenarios and expanding the speed and consistency of the CNN. Overall, the study supports the growing field of emotion-based stress detection, demonstrating to offer efficient, non-invasive methods for recognizing stress through facial analysis, although further refinement and validation are needed. The research titled “Stress and Anxiety Detection through Speech Recognition Using DNN” [5] explores the use of speech analysis to identify emotional states such as stress and anxiety. The emotion condition can be manifested through the change in person’s speech. The variation in words speaks about the person’s mental state. To analyse these facial features CNN study was utilized the MEL frequency coefficient chroma feature extraction to capture the face emotion using CNN and classify the speech into joy, fear, disgust, neutral, sadness, surprise, and anger. With the accuracy of 76% by classifying the emotions. It has several drawbacks where small input detects the full emotion in large population. With expansion of dataset additionally employing more advanced DL architecture for model’s performance. That is speech to text and face input in mental healthcare. The paper titled “Real-time Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data” [6] proposes an innovative system for the real-time detect and analyse the stress through facial emotion recognition. This system integrates computer vision and ML to evaluate individuals' emotional states using live video feeds. By utilizing a pre-trained deep learning model, the system identifies a range of stress-related emotions such as "Busted," "Irritated," "Anxious," "Relaxed," "Neutral," "Brooked," and "Shocked." “To enhance user engagement and practical applicability, the system is embedded as application that gives another visualization. Emotional distribution graphs average daily stress levels and frustration over time are also included. The personalized recommendation is included in mental wellbeing. Data collection, preprocessing, emotion, recognition, evaluation, and recommended generation are involved in methodology. The dataset of facial images is labelled with corresponding emotional states with preprocessing which ensures data quality and consistency. The paper titled “Multimodal Deep Learning Framework for Mental Disorder Recognition” [7] introduces an innovative system designed to detect mental health disorders by integrating data from multiple modalities—specifically, text, audio, and visual sources. This comprehensive approach aims to construct a more accurate and holistic representation of an individual's mental state by combining signals often analysed in isolation. The proposed framework employs Convolutional Neural Networks (CNNs) for analysing facial images, Recurrent Neural Networks (RNNs) for processing textual data, and Mel-Frequency Cepstral Coefficients from speech input. These data source are processed through specialized deep learning pipelines to produce high-level feature representations. A key contribution is the multimodal fusion mechanism, which integrates these separate feature sets into a unified model, enabling more robust and context-aware predictions. Experiments were carried to recognized benchmark datasets: Daidalos and AVEC 2014. They show that this multimodal approach consistently outperforms traditional single-modality systems, underlining the synthesizing diverse behavioural and physiological cues in mental disorders assessments. Although specific performance metrics such as precision or F1-scores were not detailed, where the multimodal integration in enhancing detection capabilities. This framework represents a promising direction for future AI-assisted mental health diagnostics, offering improved accuracy and deeper insight into complex emotional and cognitive states. In essence, this work demonstrates how advanced DL techniques—when applied across multiple input types—can push the boundaries of current mental disorder recognition systems, paving the way for more nuanced and reliable tools in mental health care.

The paper titled “Mental Disorder Detection via Social Media Mining Using Deep Learning” [8] explores the application of DL in identifying disorder like depression, through the analysis of user-generated data source.

This approach addresses the drawback of traditional diagnostic methods by leveraging publicly available digital footprints to assess psychological states. In the study, the authors collected tweets from individuals who self-identified as experiencing depression. They extracted linguistic and psychological features from the posts, including sentiment polarity, emotional tone, usage of personal pronouns, and absolutist language—which behaviours in prior research. To categorize the cause of depression, the authors applied Hierarchical Clustering to organize the data into three distinct depression levels. Subsequently, a Long Short-Term Memory neural network was trained to classify users by the extracted features. They use metrics such as a confusion matrix. Results demonstrated that the integration of DL with linguistic feature analysis yielded reliable classification outcomes, affirming the feasibility of using social media data for mental health detection. While promising, the authors acknowledged certain drawbacks, such as the relatively small and potentially biased dataset. They proposed expanding the feature set, incorporating more diverse mental health categories, and experimenting with more advanced clustering techniques in future work. This study highlights the potential of social media DL to serve as an accessible and non-invasive tool for early detection of mental disorders. It also underlines the ethical considerations and data diversity to increase for these models. The paper "Application of Data Fusion for Automated Detection of Children with Developmental and Mental Disorders [9] provides recent advancements in utilizing data fusion and artificial intelligence for the early identification of mental and developmental disorders in children. The study focuses on nine major conditions, including Autism Spectrum Disorder (ASD), ADHD, Schizophrenia (SZ), anxiety, depression, dyslexia, Post-Traumatic Stress Disorder (PTSD), Tourette Syndrome (TS), and Obsessive-Compulsive Disorder (OCD). The review underscores the urgent need for early diagnosis, as approximately 14% of children and adolescents globally are affected by these disorders, which often persist into adulthood, impacting their social, educational, and emotional development. Traditional diagnostic approaches—such as clinical interviews and neuropsychological assessments—are often time-consuming, subjective, and trained professionals. In contrast, the study highlights the growing trend of using non-invasive, cost-effective physiological signals like EEG (electroencephalogram), ECG (electrocardiogram), and PPG (photoplethysmogram) for automated detection. These signals are processed using a structured pipeline that includes signal acquisition, preprocessing, feature extraction, and classification. A variety of machine learning algorithms (e.g., Support Vector Machines, k-Nearest Neighbours, Decision Trees) and DL (e.g., CNN and Long Short-Term Memory networks) are employed in these systems. Despite notable progress, the review identifies persistent challenges, such as the lack of large, publicly available datasets, susceptibility of physiological signals to noise, limited model interpretability, and inconsistent validation practices. The authors emphasize that **data fusion**—combining multiple physiological signals—can enhance diagnostic accuracy and better account for **comorbidities** (i.e., the co-occurrence of multiple disorders). For future research, the paper advocates for the development of portable, real-time analysis tools, the establishment of standardized evaluation protocols, and the design of transparent, explainable AI models to increase the trust of clinicians and caregivers. This review reinforces the potential of AI-driven, multimodal diagnostic systems to transform the landscape of paediatric mental health care by offering early, scalable, and accurate detection methods. The paper "Dual Layer Cognitive Insight Deep-Mood Encoder (DLCDME)" [10] introduces a novel two-tiered feature encoding framework specifically designed for text-based depression detection. Recognizing that depression affects approximately many people globally with artificial intelligence—particularly NLP—in accurately detecting depressive states from written or spoken language. Previous efforts in this domain have leveraged transformer-based models such as BERT, Mental BERT, and ClinicBERT, often combined with architectures like LSTM or CNN to capture linguistic and contextual cues. However, these earlier approaches have struggled with limited feature integration and inadequate contextual embedding, reducing the consistency of depression detection. To configure the proposed DLCDME framework fuses transformer encoders with LSTM networks and integrates a multi-head method. The dual-layered design enables more effective contextual representation and fusion, leading to a knowledge for user text and improved classification performance. It uses DAIC-WOZ dataset, where it achieved a test accuracy of 95.46%, an F1-score of 0.9546, and low error metrics (Mean Absolute Error of 3.40 and Root Mean Square Error of 4.38). These results mark a significant improvement over existing methods. Ultimately about NLP architectures and feature-rich encoding to revolutionize AI-driven mental health diagnostics, making them more accurate, explainable, and scalable.

METHODOLOGY

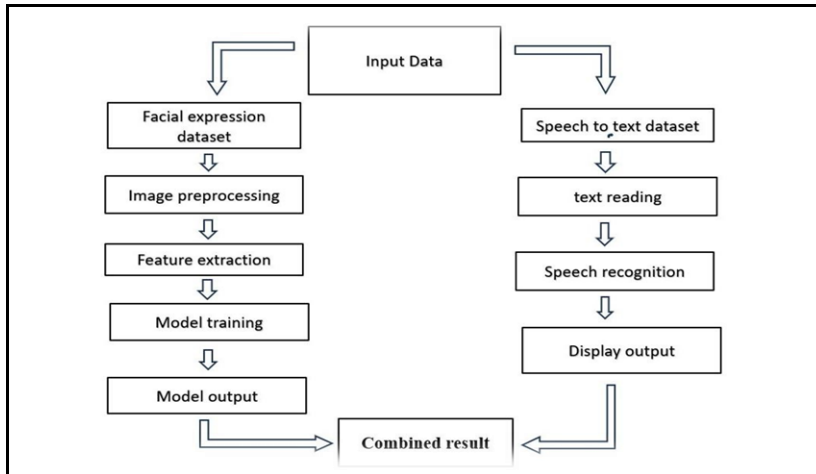


Fig. 1 System Architecture

The Fig. 1 shows proposed system architecture. The proposed methodology involves collecting multi-modal data, which includes both speech-to-text input and facial expression data. The collected data undergoes preprocessing to convert it into suitable formats for deep learning models, where facial images and speech-transcribed text are handled accordingly. Feature selection and extraction processes are applied, where facial features are detected using techniques like Haar cascades, and textual features are extracted from voice inputs using speech recognition followed by natural language processing (NLP). The model is trained on labeled datasets annotated with various mental health conditions, such as stress, no stress, depression, no depression, and anxiety, ensuring that the dataset is balanced and diverse. The system reads text data from CSV files and recognizes user speech inputs, converting them to text. The approach involves two separate input streams: a facial expression dataset consisting of images or video frames capturing emotional expressions, and a speech-to-text dataset composed of audio recordings or their transcriptions. Facial images undergo preprocessing steps like normalization, resizing, and enhancement, after which key features such as geometric landmarks and appearance-based characteristics are extracted. These features are then passed through a Convolutional Neural Network (CNN) to predict the facial expression classes. Concurrently, speech input is processed through speech recognition systems to convert spoken words into text, which is then further refined using NLP techniques. The final output is derived by combining the results from facial expression analysis and speech-to-text conversion, enabling a comprehensive assessment of mental health indicators by integrating both facial and speech cues.

RESULT AND ANALYSIS

Class	Training	Testing
Anxiety		
depressed		
Not depressed		
stress		
Nostress		

Fig. 2 Training and Testing datasets

The Fig. 2 shows sample test and training dataset. A rich and diverse dataset, comprising both transcribed speech and videos of facial expressions eroded from the AI-driven system for find mental health conditions. DL and ML are used to label dataset with particular mental health condition such as depression,anxiety,stress and

notdepressed. The face data was analyzed using CNNs and speech to text with NLP. It includes the minute details like facial landmarks, microexpression, words differentiation during result analysis. The extracted features are mapped with labels to assist the model in learning distinct behaviour. The outcome has obtained the accuracy of 89% in identifying mental disorders with both speech and face expression. This level of precision cannot be observed in the traditional methods.

Combining visual and auditory data enhanced the model's overall reliability and usefulness, significantly outperforming systems that rely solely on one type of input. This dual-modality approach offers a powerful application for mental health issues by utilizing both speech and facial cues. It provides medical professionals and mental health experts with objective, current information that can aid in more rapid, accurate diagnosis and treatment. Additionally, the broader area of facial and speech-based disorder detection is becoming most popular for identifying signs of psychological conditions like anxiety and depression as well as neurological and developmental disorders like Parkinson's disease, autism spectrum disorder, and post-stroke impairments. These systems analyze features like facial symmetry, articulation, speech pauses, voice modulation, and micro-expressions using sophisticated computer vision and audio processing algorithms. Once significant patterns have been identified, classification models can determine the likelihood or severity of a particular condition. This method stands out for being non-invasive, scalable, and efficient. It has the ability to increase the disease and cure of neurological and mental illnesses by offering the ability to monitor continuously and making mental health assessments accessible, even outside of traditional clinical settings.

The system also generates prospects for further advancements in personalized mental health care. With continued improvements in machine learning algorithms and access to larger, more diverse datasets, the model can grow to detect a greater variety of neurological and psychological disorders with even greater accuracy. Integration with wearable technology, smartphone apps, or telemedicine platforms could enable real-time mental health monitoring in everyday settings, enhancing the proactive and accessible nature of support. This is a significant step in democratizing the system's proactive psychological, emotional, and mental health monitoring.

The system used for important developments in individualized mental health treatment by building on this foundation. The model's capacity to recognize a greater variety of neurological and psychological disorders will significantly increase as machine learning methods advance and larger, more varied datasets become accessible. Furthermore, continuous and real-time mental health monitoring may become a feasible aspect of everyday life by combining this technology with wearable sensors, smartphone apps, and telehealth services. Support for mental health would become more timely, individualized, and accessible as a result. These developments have improved mental health care from reactive to proactive management, guaranteeing that everyone's emotional health is consistently and successfully preserved.

The Fig. 3 shows sample output by considering both face and voice input. By considering both the output it gives matched disorder.

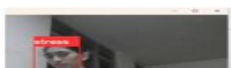




Test case no	Input 1	Input 2	Output
1.		Speech-Based Analysis Process speech patterns to identify psychological and emotional states. Recognized: I am so stressed. Detected from Voice: stress	Matched Disorder: stress
2.		Speech-Based Analysis Process speech patterns to identify psychological and emotional states. Recognized: I am not stressed. Detected from Voice: nostress	Matched Disorder: nostress
3.		Speech-Based Analysis Process speech patterns to identify psychological and emotional states. Recognized: I am depressed. Detected from Voice: depressed	Matched Disorder: depressed
4.		Speech-Based Analysis Process speech patterns to identify psychological and emotional states. Recognized: I am so happy. Detected from Voice: not_depressed	Matched Disorder: notdepressed
5.		Speech-Based Analysis Process speech patterns to identify psychological and emotional states. Recognized: I am worried. Detected from Voice: anxiety	Matched Disorder: anxiety

Fig. 3 Sample Output

CONCLUSION

The proposed system demonstrate how effective it is to combine speech and facial analysis to identify mental health issues. The dual sense approach is used in Realtime to improve the capacity to identify the mental disorder. They help the doctors to diagnose mental disorder accurately in medical field. There are also some of limitation unevenness of dataset which contains images and words for face and speech to text recognition. They must find the solution for increasing the accuracy of system. The detection of disorder using dual sense can be developed for monitoring in future with varied data source which includes behaviour data etc. The feedback loops are enabled between the doctors and patients for improving the dual sense system. It develops to initiate strategy to detect early symptoms of mental disorder and prevent them as early as possible.

REFERENCES

1. Leone, A. C., and Syed, S. H. (Year). An Investigation into Deep Learning Approaches for Facial Expression Analysis in Healthcare Contexts.
2. V. Savchenko & A. Savchenko, Emotion and Engagement Detection in E-Learning Environments Using a Unified Neural Network for Facial Expression Recognition.
3. Y. Yushchenko, Analyzing Stress and Anxiety Through Blog Content Using Data-Driven Techniques, Institute of Computer Science, University of Tartu, 2018.
4. Almeida, José, and Fátima Rodrigues. "Facial Expression Recognition System for Stress Detection with Deep Learning." ICEIS (1). 2021.
5. Almeida, J., & Rodrigues, F. (2021). Utilizing Deep Neural Networks for Stress Detection via Facial Expression Analysis. In Proceedings of the International Conference on Enterprise Information Systems (ICEIS).
6. A. W. Blap, M. Vadillo & V. C. Rin, Utilizing Social Media to Detect and Address Mental Health Issues, IEEE Int. Conf. I: 10.1109/ICHI.2017.24
7. Liu and Ling investigate the effectiveness of deep convolutional neural networks in interpreting and classifying human facial expressions.
8. T. V. Das & R. Y. Mehra, Early-Stage Depression Detection through Deep Learning on Social Platforms.
9. Lalit, M. J., & Narayan, N. K. (2022). An Overview of Emotion-Sensitive Artificial Intelligence for Mental Health Assessment Using Machine Learning Techniques. Materials Today: Proceedings, 58, 217–226.
10. Dimitriou, A., & Nikolaou, P. S. (2019). A Detailed Survey on Depression Assessment Using Visual Indicators. IEEE Transactions on Affective Computing, 10(4), 445–470.