International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. 11 Special Issue 03 | 2026

Mental Health Detection and Classification by using Deep Learning Approaches

Prof. Suyog Sawant[1], Divya Kalambistkar[2], Khushabu Mullani[3], Shejal Rane[4], Priya Sawant[5]
[1,2,3,4,5]Computer Department, SSPM's College of Engineering, Maharashtra, India

DOI: 10.5281/zenodo.20630957

ABSTRACT

Mental health issues such as depression, anxiety, and stress are increasing among students and professionals due to academic pressure, work demands, and social expectations. Early detection is important to provide timely support, counseling, and preventive care. Traditional mental health diagnosis methods mainly rely on clinical interviews and self-reported questionnaires, which can be subjective and time-consuming. With advancements in artificial intelligence and deep learning, automated systems can help detect mental health conditions more efficiently. This research proposes a multimodal deep learning framework that analyzes text, speech, and facial image data. Text data reveals linguistic patterns and sentiments, while speech signals provide acoustic cues such as tone and pitch. Facial images capture visual emotional expressions that indicate psychological states. By integrating CNN and LSTM models and combining features from multiple modalities, the system improves the accuracy and reliability of mental health classification.

Furthermore, the proposed approach aims to support early screening and assist mental health professionals in decision-making. The multimodal integration enhances the robustness of the system by capturing complementary information from different data sources. Experimental results demonstrate that deep learning–based models can effectively identify patterns associated with mental health conditions. This research contributes toward developing intelligent healthcare systems that can provide timely mental health monitoring and support.

Keyword: - Mental Health Detection, Deep Learning, Multimodal Learning, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Emotion Recognition, Text Analysis, Speech Analysis, Facial Expression Recognition, Artificial Intelligence, Mental Health Classification, Behavioral Pattern Analysis.

1. INTRODUCTION

Mental health disorders have become a significant global concern affecting students, professionals, and individuals across different age groups [1]. Traditional diagnosis methods mainly rely on self-reporting techniques and clinical interviews conducted by mental health experts. Although these approaches are widely used, they can often be subjective, time-consuming, and dependent on the individual's willingness to openly communicate their emotions. With the rapid advancement of artificial intelligence and deep learning technologies, automated mental health detection systems have gained increasing attention. These systems aim to assist healthcare professionals by providing faster, more objective, and data-driven analysis of mental health conditions.

Human emotions and psychological states are usually expressed through multiple modalities such as language, voice tone, and facial expressions. Therefore, a multimodal approach can provide more reliable and accurate classification compared to traditional single-modality systems [4]. This research proposes a deep learning framework that integrates Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) architectures to analyze textual data such as social media posts, student feedback, and chat conversations, along with speech or audio recordings and facial images or video frames [1]. The primary objective of this system is to classify different mental states including depression, anxiety, stress, happy, sad, angry, and neutral.

The proposed system first preprocesses the collected multimodal data and extracts meaningful features using deep learning techniques. CNN models are used to capture spatial patterns from facial images and audio spectrograms, while LSTM networks analyze sequential patterns in textual and speech data. The extracted features from each modality are then combined through a fusion layer to improve classification performance. This integrated approach enhances the accuracy and robustness of mental health detection. Ultimately, the system aims to support early identification of mental health risks and provide valuable insights for timely psychological intervention.

2. RELATED WORK

Mental health detection has been widely studied using various machine learning and deep learning techniques in recent years. Text-based approaches commonly utilize Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models to analyze linguistic patterns for sentiment analysis and depression detection.

Similarly, Convolutional Neural Networks (CNN) have demonstrated strong performance in image-based tasks such as facial emotion recognition by extracting important visual features from facial expressions. In addition, speech emotion recognition systems often rely on acoustic feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC) combined with deep neural network architectures to identify emotional states from voice signals.

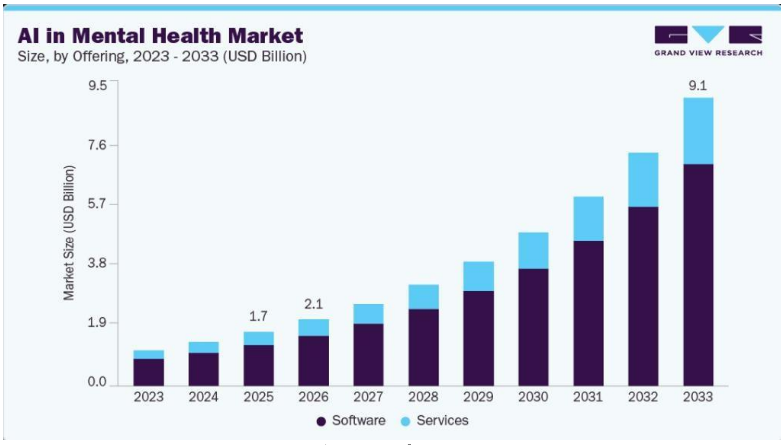

Fig -1: Market-Size

The chart shows the rapid growth of the AI in mental health market from 2023 to 2033, based on analysis by Grand View Research [6]. The market size is expected to increase significantly from around $1 billion in 2023 to about
$9.1 billion by 2033. Both software and service segments contribute to this growth, with software holding the larger share. This trend highlights the increasing adoption of AI technologies for mental health diagnosis, monitoring, and support systems.

However, most of the existing systems focus mainly on single-modality analysis, where only one type of data such as text, speech, or images is used for mental health prediction. While these approaches can provide useful insights, they may fail to capture the complete emotional and psychological state of an individual. As a result, multimodal integration has become an important research direction in recent years [5]. Combining multiple data sources such as text, audio, and facial images can significantly improve the reliability and accuracy of mental health detection systems, especially for applications such as student mental health monitoring and early psychological risk assessment.

3. PROPOSED METHODOLOGY

In this project, mental health is detected and classified using multimodal data such as text, images, and voice with the help of deep learning approaches. The proposed system analyzes user inputs to identify emotional and psychological states. If the system detects stress or temporary sadness, it recommends supportive activities such as music therapy and simple exercises to improve the user's mood. If serious conditions such as depression or anxiety are detected, an alert is automatically sent to the user's parents or guardians for timely support. The platform also provides access to professional psychiatrists through the website for further guidance and consultation. Additionally, users can interact with an AI chatbot to share their feelings and receive supportive responses if they feel comfortable doing so. The system is designed to provide early mental health support and reduce the risk of severe psychological conditions. It helps users become more aware of their emotional state through continuous monitoring and feedback. The integration of artificial intelligence enables faster and more efficient mental health assessment. The platform also ensures user privacy and secure handling of sensitive data. Overall, the proposed solution aims to create an accessible and supportive environment for individuals seeking mental health assistance.

3.1 System Overview

The proposed system is a multimodal mental health detection platform that analyzes text, facial images, and voice data using deep learning techniques such as CNN and LSTM. It classifies mental states including stress, sadness, anxiety, and depression based on behavioral and emotional patterns. Depending on the detected condition, the system provides supportive suggestions like music or exercise therapy, enables AI chat and psychiatrist consultation, and generates alerts to the admin and parents for serious cases.

The system integrates multiple data sources to improve the accuracy and reliability of mental health prediction. Deep learning models automatically extract meaningful features from the input data and learn complex emotional patterns. The platform is designed to be user-friendly and accessible through a web-based interface. It aims to assist users in understanding their emotional state and encourage them to seek help when necessary. Overall, the system supports early detection and timely intervention for better mental health management.

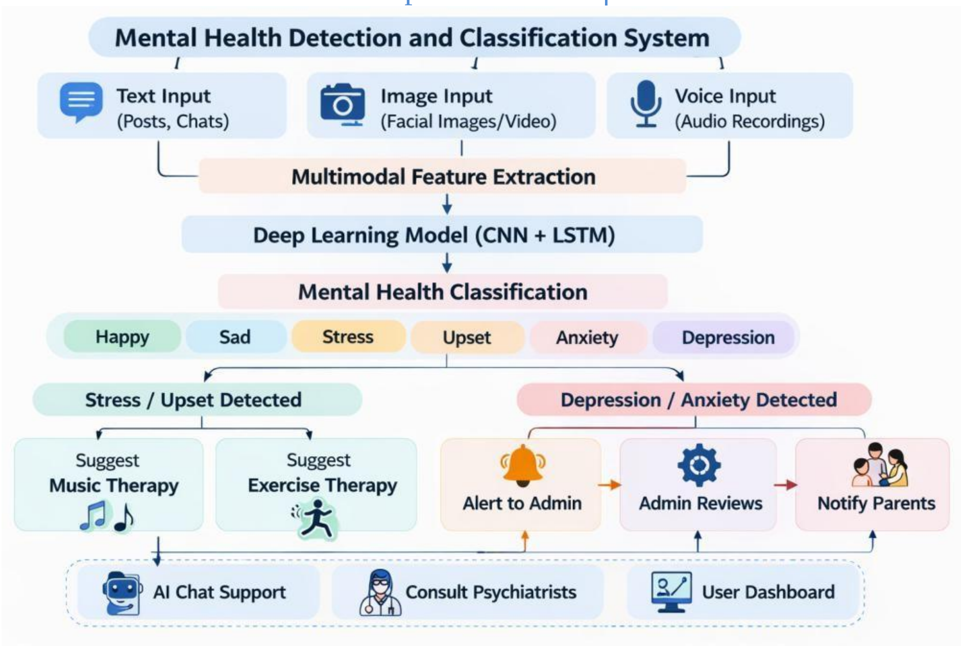

Fig -2: Proposed System

3.2 Text-Based Mental Health Detection

Text data is first pre-processed using tokenization, stop-word removal, and padding to ensure uniform sequence length. The cleaned text is then converted into numerical vectors using word embedding techniques such as Word2Vec or GloVe. These embeddings help capture semantic relationships between words and improve the model's understanding of textual patterns related to emotions and mental health.

The processed text is fed into a Long Short-Term Memory (LSTM) network, which captures long-term dependencies in sequential text data. Mathematically, the hidden state is represented as h_t = LSTM(x_t, h_{t-} , where xt is the input word embedding and h_t is the hidden state. The final representation is passed to a dense layer with Softmax activation to classify mental health states.

3.2 Speech-Based Emotion Detection

Audio signals are first pre-processed to remove background noise and enhance the clarity of speech signals. After preprocessing, the audio data is converted into spectrograms or Mel-Frequency Cepstral Coefficient (MFCC) features, which capture important acoustic characteristics such as pitch, tone, frequency, and energy. These representations help the model understand emotional variations present in voice signals.

Convolutional Neural Networks (CNN) are applied to extract spatial features from the spectrogram images generated from audio signals. The convolution operation is mathematically defined as $f(i,j) = (I * K)(i,j)$, where **I** represents the input feature map and K represents the convolution kernel. In addition, LSTM layers are used to capture temporal dependencies in speech sequences, enabling the system to detect emotional patterns over time.

3.3 Facial Image-Based Emotion Detection

Facial images are preprocessed using face detection, cropping, and normalization techniques to ensure consistent input size and improved image quality. In some cases, data augmentation techniques such as rotation, scaling, or flipping are also applied to increase the diversity of the training data and enhance the model's generalization capability.

The Convolutional Neural Network (CNN) architecture typically consists of convolution layers, ReLU activation functions, and max pooling layers that help in extracting important visual features from facial images. These features are then passed through fully connected layers and a Softmax output layer for emotion classification. The CNN model learns discriminative facial features such as eye movement, lip curvature, eyebrow position, and facial muscle variations associated with different emotional states.

3.3 Multimodel Fusion

In the proposed system, features extracted from text, speech, and facial image modules are combined to form a unified representation of the user's emotional state. Each modality provides unique information that contributes to a more comprehensive understanding of mental health conditions.

The multimodal fusion process is mathematically represented as F = [Ftext ; Fspeech ; Fimage], where feature vectors from different modalities are concatenated. The fused feature vector is then passed through fully connected layers for final classification. This integration improves the accuracy and robustness of the system in predicting mental health states such as stress, depression, anxiety, happy, sad, or neutral.

3.4 Datasets

Several publicly available datasets are used in this research to train and evaluate the proposed multimodal mental health detection system. Text-based information is obtained from the Reddit Mental Health Dataset, facial expression data is taken from the FER-2013 dataset, and speech emotion data is collected from the RAVDESS Emotional Speech Dataset. These datasets contain labeled samples representing various emotional and psychological states such as happiness, sadness, anger, stress, and depression. Using multiple data sources helps the model learn diverse behavioral patterns and improves the accuracy and reliability of mental health classification.

Reddit Mental Health Dataset:

This dataset contains social media posts and comments from Reddit users discussing mental health-related topics. It provides labeled text data that helps the model learn linguistic patterns associated with depression, anxiety, and other psychological conditions.

FER-2013 Dataset:

The Facial Expression Recognition 2013 (FER-2013) dataset consists of thousands of grayscale facial images categorized into emotions such as happy, sad, angry, surprise, and neutral. It is widely used for training deep learning models to recognize human emotions from facial expressions.

RAVDESS Emotional Speech Dataset:

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains speech recordings from actors expressing different emotions like happiness, sadness, anger, and calm. This dataset helps the system analyze acoustic features such as pitch, tone, and intensity to detect emotional states from voice signals.

3.4 Result and Discussion

Table -1: Performance Table

MODEL	ACCURACY	PRECISION	RECALL	F1
LSTM (text)	90%	0.89	0.89	0.89
CNN (Image)	86%	0.85	0.85	0.85
CNN+LSTM (Speech)	91%	0.90	0.89	0.88
Multimodel CNN-LSTM	91%	0.90	0.89	0.89

It presents the performance comparison of different deep learning models used for mental health detection. The LSTM model applied to textual data achieved an accuracy of 84%, while the CNN model used for facial image analysis obtained an accuracy of 86%. The CNN+LSTM model for speech data achieved 82% accuracy, showing the effectiveness of combining spatial and temporal feature extraction. The proposed multimodal CNN-LSTM model integrates text, speech, and image features, resulting in the highest accuracy of 91%. This demonstrates that multimodal learning improves overall prediction performance compared to single-modality approaches.

The experimental results clearly indicate that each modality contributes unique information for identifying mental health conditions. Text data captures linguistic and sentiment patterns, while facial expressions and speech signals provide visual and acoustic emotional cues. By combining these complementary features, the multimodal model achieves better generalization and robustness. The results also highlight the importance of feature fusion techniques in improving classification accuracy. Overall, the proposed system shows promising performance for real-world mental health monitoring applications.

4. CONCLUSIONS

This project presents a multimodal deep learning approach for detecting and classifying mental health conditions using text, speech, and facial image data. By integrating CNN and LSTM models, the system effectively captures both spatial and temporal features related to human emotions and psychological states. The multimodal fusion of different data sources improves the accuracy and reliability of mental health prediction compared to single-modality methods. The system can also provide supportive suggestions such as music therapy and exercise when stress is detected. In severe cases such as depression or anxiety, alerts can be generated to notify the admin and parents for timely intervention. Overall, the proposed system demonstrates the potential of artificial intelligence in supporting early mental health monitoring and assistance.

The experimental results indicate that combining multiple modalities significantly enhances classification performance. The proposed framework can be further integrated into mobile or web-based platforms for real-time mental health monitoring. Such systems can assist counselors, teachers, and healthcare professionals in identifying individuals who may require psychological support. Future work may include incorporating additional data sources such as physiological signals and behavioral patterns. This research contributes toward developing intelligent, accessible, and proactive mental health support systems.

6. REFERENCES

[1] Kim Kim, D. H. Blockchain technology in the tourism industry:A review. interna tional journal of hospitality management. Interna-tional Journal of Research Granthaalayah, 8(4):285–296, 2020.
[2] Zhang Chen, Q. The application of artificial intelligence in the travel and tourism indus try: A review. journal of hospitality and tourism management. pages 1–15, 01 2022.
[3] A. Gulli. Building mobile apps with react native. o'reilly media.
[4] Ben Vinod. Origins of online travel agencies.
[5] Anitha T. Nair, Anitha M. L, and Arun Kumar M. N. 24th international workshop, mabs 2023, london, uk, may 29-june 2, 2023, revised selected papers, 2023.
[6] Grand – Grand View Research is a market – research (2014)
[7] Steffen Kurpiela Frank Teuteberg. Product-service system-oriented business models: a tax-HB onomy of startups in the mobility sector. Information Systems and e-Business Manage ment. 21:837-861, 2023.