IES International Journal of Multidisciplinary Engineering Research

Mental Health Detection System Using Multimodal AI Techniques

Abhijith P M[1], Akshay K L[2], Alan Anil[3], Anaswara C V[4], Dr.Sugan J [5]

[1,2,3,4]*Student, Department of Computer Science and Engineering, IES College of Engineering, Thrissur, Kerala, India*

[5]*Assistant Professor, Department of Computer Science and Engineering, IES College of Engineering, Thrissur, Kerala, India*

Email_id: abhijithpm021@gmail.com, akshayklwork03@gmail.com, alananil212@gmail.com, anaswarakool123456@gmail.com, suganj067@gmail.com

Abstract

Mental health problems such as depression, anxiety, and stress are increasing worldwide and often remain unnoticed because people hesitate to seek help, fear judgment, or lack proper support. Many individuals and healthcare providers want accurate assessments, but they often struggle to get real-time, objective information. To solve this, we have proposed a new system using Multimodal AI technology. This system uses text, audio, and video inputs to detect emotional states in real time. Text input is analyzed using NLP and sentiment analysis, audio input is processed through speech-to-text and tone-based feature extraction using Librosa, and video input is analysed using DeepFace and OpenCV for facial emotion recognition. The system combines these results to classify the user's mental state as mild, neutral, or severe, and provides instant personalized suggestions through a secure web interface. AI is a powerful tool for mental health because it supports early detection, real-time monitoring, and personalized feedback, helping users manage their well-being more effectively.

Keywords: Mental Health Detection, Multimodal AI, Emotion Recognition, Sentiment Analysis, Real-Time Monitoring, Text Analysis, Audio-Visual Analysis

DOI: https://doi.org/10.5281/zenodo.18871580

1. Introduction

In today's fast-paced digital world, maintaining mental well-being has become a major global challenge. Traditional methods of assessing mental health such as therapist evaluations, questionnaires, or self-reports often lack objectivity and fail to deliver real-time insights. These limitations delay early detection and increase the risk of unnoticed psychological distress. The AI-powered Mental Health Detection System proposed in this project addresses these challenges by integrating advanced artificial intelligence and deep learning technologies to create an intelligent, automated, and comprehensive emotional monitoring framework. By combining Natural Language Processing (NLP), Convolutional Neural Networks (CNN), and multimodal learning, the system analyzes text, speech, and facial expressions to detect stress, anxiety, and depression levels. Through this innovative approach, individuals and healthcare professionals can receive timely feedback, enhance mental health awareness, and improve overall

emotional well-being.

1.1 Text, Audio, and Visual Analysis: Utilizes NLP for text-based emotion recognition, CNNs for facial expression detection, and audio feature extraction for voice emotion analysis. This multimodal integration ensures accurate emotional profiling across diverse inputs.

1.2 Emotion Classification Module: A machine learning-based classifier that categorizes emotional states such as happy, neutral, stressed, anxious, or depressed, enabling early detection of psychological conditions.

1.3 Data Processing and Integration Server: Collects multimodal data inputs, preprocesses them, and synchronizes results using AI-driven analytics. The server maintains structured emotional records for further analysis and visualization.

1.4 User Interface (UI): A user-friendly web dashboard that displays real-time emotion analysis, visual graphs, and personalized well-being suggestions. It also includes privacy controls to safeguard user data and identity.

1.5 Enhanced Accessibility and Support: By combining these components, the AI-powered Mental Health Detection System transforms traditional assessments into a smart, automated, and continuous monitoring solution, improving accessibility, accuracy, and emotional care across various environments.

2. Literature Review

Akram Ahmad, Vaishali Singh, and Kamal Upreti (2025) [1] proposed "Emotion Recognition Through Facial Expressions." Their study employed Convolutional Neural Networks (CNNs) to identify emotions such as happiness, sadness, anger, and fear. The model performed well with clear expressions but struggled with subtle or neutral emotions due to limited feature generalization. Despite these challenges, the work established a foundation for vision-based emotion detection using deep learning techniques. The study further highlighted the need for improved data diversity and illumination normalization to enhance real-world applicability.

Alireza Hasanzadeh and Fatemeh Nargesian (2024) [2] presented "Harnessing the Power of Hugging Face Transformers for Predicting Mental Health Disorders in Social Networks." The authors utilized pre-trained transformer models such as BERT to analyze linguistic and emotional cues from social media data. Their system achieved higher accuracy than traditional models but required significant computational resources for real-time inference. The study demonstrated the potential of transformer architectures in digital mental health monitoring and emphasized the importance of fine-tuning large language models for domain-specific psychological datasets.

Syde Muhammad Aqleem and Qaisar Abbas (2023) [3] proposed "EmotionNet-X: An Optimized CNN Architecture for Robust Facial Emotion Analysis." This work introduced a customized CNN framework incorporating dropout and batch normalization to enhance accuracy and convergence speed. While the system performed efficiently under standard lighting conditions, it showed reduced accuracy in low-light or occluded environments. The study emphasized the importance of adaptive preprocessing and lightweight architectures for real-time facial emotion recognition applications.

Rung-Huei Huang, Tsung-Han Tsai, and Yu-Chuan Su (2022) [4] presented "LSTM and Attention in Emotion Fusion." Their research combined Long Short-Term Memory (LSTM) networks with attention mechanisms to

integrate temporal and contextual dependencies from multimodal data. The model achieved strong performance in emotion recognition but remained sensitive to synchronization mismatches between modalities. This work highlighted the effectiveness of attention-based fusion in enhancing multimodal emotion understanding and suggested future work on adaptive synchronization strategies.

Vamsi Kumar and Bhanu Prasad Reddy (2020) [5] proposed "Emotion Detection in Noisy Audio." The study utilized LSTM models with noise reduction and preprocessing techniques to enhance recognition accuracy in challenging acoustic conditions. The system effectively identified emotional states such as anger and happiness even in low-quality recordings. The authors demonstrated that LSTM-based architectures are robust for real-world speech emotion detection and recommended hybrid CNN-LSTM models for improved temporal and spectral feature extraction.

Georgios Tzirakis, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou (2017) [6] presented "End-to-End Multimodal Emotion Learning." Their work integrated CNN and RNN architectures to process audio and visual data simultaneously, learning cross-modal correlations for improved emotion classification. The model achieved superior accuracy compared to single-modality systems but required extensive labelled datasets. The study provided key insights into end-to-end multimodal learning frameworks and demonstrated the potential of combining temporal and spatial features for emotion recognition.

Zixing Zhang, Fei Wening Huang, and Björn Schuller (2016) [7] proposed "Audio-Visual Emotion Recognition Using Multimodal Deep CNNs." The authors designed a deep CNN architecture that fuses acoustic and visual inputs to enhance emotion classification accuracy. The system achieved promising results on balanced datasets but suffered from performance degradation under class imbalance. This work underscored the necessity of dataset diversity and balanced multimodal representation for building reliable emotion recognition models.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Alexander Gelbukh (2016) [8] presented "Deep Learning-Based Emotion Recognition." The study employed LSTM networks to capture contextual information from textual inputs for emotion detection. Results showed improved accuracy over traditional sentiment models, although synchronization with other modalities remained challenging. The paper laid the groundwork for context-aware text-based emotion analysis and emphasized multimodal extensions combining text, audio, and visual features.

Erik Cambria, Amir Hussain, and Catherine Havasi (2014) [9] proposed "Sentic Patterns for Text-Based Sentiment." This study introduced a concept-based sentiment analysis framework that interprets deeper emotional meanings beyond surface-level word polarity. While effective for formal language, the model struggled with slang and informal expressions often present in real-world datasets. The research contributed significantly to advancing semantic-level emotion recognition in text processing and inspired the development of affective computing frameworks.

Mohammad Soleymani, Daniel Garcia, and Thierry Pun (2012) [10] presented "Multimodal Fusion for Emotion Detection." Their work applied Support Vector Machines (SVM) and Local Binary Patterns (LBP) to integrate visual, audio, and textual cues for emotion classification. The approach improved recognition accuracy

compared to unimodal systems but faced computational challenges in real-time applications. This study emphasized the benefits of multimodal fusion for achieving reliable emotion detection and established an early foundation for cross-domain affective computing research.

The reviewed studies collectively explore various deep learning and multimodal approaches for emotion recognition across text, audio, and visual data. Early works focused on facial emotion detection using Convolutional Neural Networks, which effectively identified basic emotions but struggled with subtle expressions and varying lighting conditions. Subsequent research leveraged transformer models like BERT for analyzing emotional and linguistic cues in social media, achieving high accuracy but requiring significant computational resources. Optimized CNN architectures introduced improvements in convergence speed and performance under controlled conditions, while attention-based LSTM models enhanced multimodal emotion fusion by integrating temporal and contextual features, though synchronization issues persisted.

Studies on audio emotion recognition demonstrated that LSTM and hybrid CNN-LSTM models effectively captured emotional tones even in noisy environments. Multimodal frameworks combining CNN and RNN architectures achieved superior emotion classification accuracy by learning cross-modal relationships, though they required large, balanced datasets. Overall, the literature highlights the evolution from unimodal to multimodal emotion recognition systems, emphasizing the importance of data diversity, real-time adaptability, and efficient fusion mechanisms to improve robustness and generalization in real-world applications.

3. Review of Methodology

3.1 System Design:

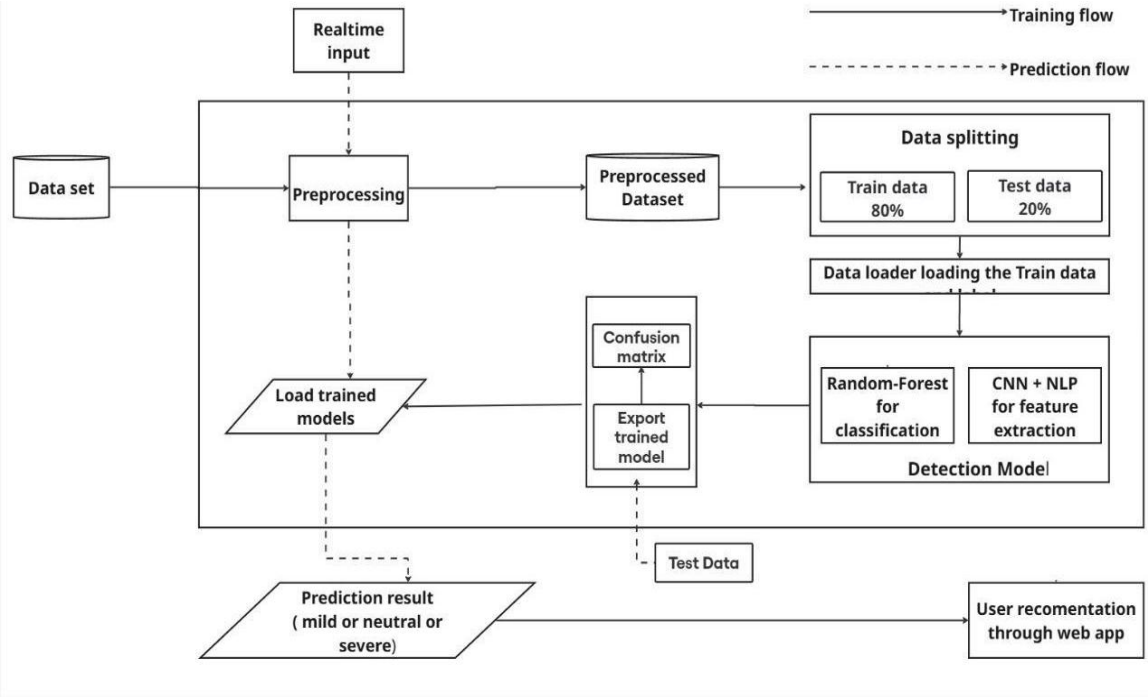

Figure 1: System Design

The Multimodal AI-based Mental Health Detection System integrates text, audio, and video inputs for real-time emotion analysis and personalized recommendations. Key stakeholders include users, AI models, and the web interface. Users provide inputs, AI models analyse emotional states, and the web interface delivers results and suggestions. The system combines preprocessing, feature extraction, and classification to predict mental states as mild, neutral, or severe, ensuring privacy, reliability, and real-time performance.

3.1 Data Collection Module

This module forms the foundation of the system by gathering input data from multiple sources:

a. Dataset Integration: Uses pre-existing multimodal datasets containing text, audio, and video samples annotated with emotional states.
b. Real-time Input: Captures user input in real time through text, speech, and facial video streams.
c. Data Storage: Organizes all collected data systematically for preprocessing and training purposes.

3.3 Data Preprocessing Module

This module transforms raw input into structured, analyzable form:

a. Text Preprocessing: Tokenizes and cleans text data, removes noise, and converts words into embeddings suitable for NLP models.
b. Audio Preprocessing: Extracts feature like Mel Frequency Cepstral Coefficients (MFCCs), pitch, and tone using tools such as Librosa.
c. Video Preprocessing: Extracts key frames, detects faces, and normalizes them using OpenCV and DeepFace for consistent analysis.
d. Unified Preprocessed Dataset: Combines processed text, audio, and video features into a single dataset for further use.

3.4 Data Splitting and Loading Module

This module prepares the data for model training and testing:

a. Dataset Division: Splits the preprocessed dataset into 80% training and 20% testing sets.
b. Data Loader: Feeds the training data batches into the detection model to optimize learning efficiency and reduce overfitting.

3.5 Detection Model Module

This is the core analytical engine of the system:

a. Feature Extraction: Employs CNNs (for visual features) and NLP models (for linguistic sentiment features).
b. Classification: Uses a Random Forest classifier to categorize emotional states into mild, neutral, or severe.
c. Model Fusion: Integrates CNN and NLP outputs to enhance multimodal accuracy.
d. Model Export: Stores the trained model for real-time predictions during deployment.

3.6 Model Training and Evaluation Module

This module ensures the system's performance and reliability:

a. Training Process: Uses the training dataset to train the CNN + NLP + Random Forest pipeline.

b. Evaluation Metrics: Tests the model using the test dataset and evaluates performance through the Confusion Matrix.
c. Model Optimization: Tunes parameters for maximum prediction accuracy.
d. Exporting Trained Model: Saves optimized model weights for deployment.

3.7 Prediction and Recommendation Module

This module handles live prediction and user interaction:

a. Loading Trained Model: Imports the saved model for use in real-time predictions.
b. Real-time Prediction: Takes new user inputs (text, audio, or video) and predicts emotional state mild, neutral, or severe.
c. User Recommendation: Provides personalized feedback or mental health support through the web interface.
d. Result Visualization: Displays results instantly with confidentiality and user-friendliness.

4. Review of Datasets

A review of datasets for the Multimodal Mental Health Detection System ensures that the data incorporated within the platform aligns with the objectives of accuracy, reliability, and ethical monitoring. Each dataset plays a crucial role in enabling the system to detect emotional states from multiple sources, providing accurate mental health assessment and personalized feedback. Proper collection, annotation, and preprocessing of these datasets ensure the effectiveness of the entire system.

4.1. Text Data

The Text Data contains all user-provided textual inputs, such as chat messages, typed responses, or social media entries (if p Each record includes a unique ID, timestamp, the text content, and the corresponding emotion/mental health label. By processing text through NLP techniques, including tokenization, sentiment analysis, and contextual embedding (BERT), the system can extract emotional features. Maintaining completeness and accuracy of this dataset is essential for reliable detection of emotional cues and mental health classification.

4.2. Audio Data

The Audio Data represents all speech inputs collected from users, including voice recordings or real-time microphone capture. Each record includes audio ID, user ID, duration, sampling rate, and labeled emotional state. Feature extraction techniques such as MFCC, pitch, and prosody analysis are applied to convert raw audio into meaningful inputs for machine learning models. Accurate and clean audio data ensures the system can detect stress, anxiety, or other emotional indicators effectively.

4.3. Facial Expression / Video Data

The Video Data captures facial expressions through webcam or recorded video clips. Each entry includes video ID, user ID, timestamp, frame sequence, and emotion label. Frames are processed to detect facial landmarks and extract visual features using CNN-based architectures. Proper labeling and quality of facial data are critical for correctly interpreting subtle emotional changes and non-verbal cues, which are key to multimodal mental health assessment.

4.4. User Profile Data

The User Profile Data contains information about all participants interacting with the system. Each record includes user ID, demographic details (age, gender), consent for data use, and history of interactions. Maintaining accurate user profiles allows the system to personalize analysis, track longitudinal emotional patterns, and provide appropriate suggestions or alerts while ensuring ethical compliance and privacy.

4.5. Interaction & Feedback Data

This dataset stores logs of user interactions, system suggestions, and feedback responses. Each entry contains interaction ID, user ID, modality used (text/audio/video), system output, and user feedback. Proper management of this dataset helps in evaluating system performance, refining models, and ensuring the system's recommendations are effective and aligned with user needs.

5. Implementation of The Multimodal AI-Based Mental Health Detection System

The implementation of the multimodal AI-based mental health detection system provides an intelligent, real-time, and non-invasive solution for identifying emotional states through text, audio, and video inputs. By integrating Natural Language Processing (NLP), Deep Learning, and Computer Vision, the system effectively analyzes multiple behavioural cues to assess mental health conditions such as depression, anxiety, and panic. The framework aims to offer early detection and personalized support, overcoming limitations of conventional, questionnaire-based diagnosis methods.

5.1 System Architecture

The system architecture of the proposed multimodal AI-based mental health detection system is designed as a layered model consisting of three main components: the User Interaction Layer, the Processing and Analysis Layer, and the Prediction Layer. The User Interaction Layer allows users to seamlessly provide inputs through multiple modes such as text via a chatbot, audio through a microphone, and video via a webcam for facial expression capture. The Processing and Analysis Layer handles data preprocessing, sentiment extraction, and feature analysis using Natural Language Processing (NLP) techniques and deep learning models. The Prediction Layer performs cross-modal fusion and applies attention mechanisms to combine multimodal features and produce the final emotion classification. The system uses SQLite as the backend database to securely store user inputs, extracted features, and prediction results, ensuring lightweight, reliable, and efficient data management. This architecture promotes modularity, scalability, and smooth interaction among all components, ensuring high accuracy and consistent performance across modalities.

5.2. Text, Audio, and Video Integration

The system integrates three synchronized processing pipelines corresponding to the input modes. In the text pipeline, user inputs from the chatbot are analysed using NLTK for sentiment and emotional context extraction. The audio pipeline functions in two ways: (i) Speech-to-Text conversion using Speech Recognition and STTX3 libraries, followed by sentiment analysis similar to text mode, and (ii) Tone analysis using Librosa, where features such as MFCCs, pitch, and energy are extracted and classified using CNN or Random Forest models. The video pipeline captures real-time facial expressions through OpenCV, with each frame analysed by a CNN (DeepFace) model to

detect micro-expressions and behavioural cues. Each modality produces an independent emotion score contributing to the final emotion prediction.

5.3. System Modules

The system consists of three main modules that work together to enable effective emotion recognition and user interaction. The User Module serves as the primary interface, allowing users to type messages, record audio, or capture video, and providing real-time feedback on detected emotional states along with visual indicators such as confidence levels and trend charts. It also offers personalized suggestions and maintains a dashboard of historical emotional patterns to support user awareness and mental well-being. The Analysis Module forms the core intelligence of the system, processing multimodal inputs by applying natural language processing and sentiment analysis to text, signal processing and feature extraction to audio, and convolutional neural network-based facial expression recognition to video. Features from all modalities are fused using Additive Cross-Modal Attention, enabling accurate and context-aware emotion prediction across Normal, Anxious, Depressed, and Panic states. The Administrator Module oversees dataset management, model retraining, and system monitoring, ensuring data privacy, enforcing role-based access control, tracking performance metrics, and maintaining audit trails for accountability. Together, these modules create a secure, responsive, and adaptive system capable of providing real-time, reliable, and personalized emotional analysis to users.

5.4. Model Training and Prediction

Each modality's model is trained separately on domain-specific datasets before multimodal fusion. The text model is trained using preprocessed chat data labelled with emotional states. The audio model uses RAVDESS and IEMOCAP datasets for speech emotion recognition, extracting MFCC and tone-related features. The video model is trained using FER2013 and CK+ datasets to recognize facial expressions. After training, feature vectors from all three modalities are merged using Additive Cross-Modal Attention for improved inter-modal correlation. The final Softmax classifier predicts emotional categories such as Normal, Anxious, Depressed, and Panic.

5.5. Database and User Interface

The system employs MySQL for maintaining supplementary data such as user credentials, chat logs, and interaction metadata. The front-end interface, developed using HTML, CSS, and JavaScript, enables an interactive and user-friendly experience. The back-end, built with Flask, connects the interface to the AI models for real-time emotion analysis. The user dashboard displays emotion results, trend graphs, and supportive recommendations based on detected states. The modular design ensures compatibility with both desktop and mobile browsers for wider accessibility.

5.6. Security and Data Privacy

To safeguard user information and maintain ethical compliance, the system integrates multiple security mechanisms. All user data are anonymized and stored in encrypted form to prevent unauthorized access. Role-based access control restricts administrative privileges, ensuring data integrity and controlled usage. The system adheres to data protection standards, and periodic audits are conducted to detect anomalies or misuse. The design ensures

confidentiality, integrity, and transparency critical for handling sensitive mental health data responsibly.

6. Requirements

6.1 Hardware Requirements:

a. Processor: Intel Core i7-10850H and above

A high-performance processor such as the Intel Core i7-10850H (or equivalent) is required to efficiently handle computational workloads, including multimodal data preprocessing and deep learning inference. It ensures smooth execution of real-time text, audio, and video analysis.

b. Primary Memory: 16GB DDR4 RAM, 3200 MHz and above

A minimum of 16GB DDR4 RAM is necessary to manage large datasets and enable parallel processing during training and testing phases. Sufficient memory ensures seamless multitasking, reduced lag, and efficient model execution.

c. Storage: 512GB Solid State Drive (SSD) and above

Fast storage such as a 512GB SSD is essential for quick access to datasets, trained models, and user session logs. SSDs improve overall system responsiveness, reduce loading times, and ensure reliable storage of sensitive mental health data.

d. GPU: NVIDIA GeForce RTX 3050, 6GB DDR6 and above

A dedicated GPU like the NVIDIA RTX 3050 with at least 6GB VRAM is required to accelerate deep learning tasks, including CNN-based facial recognition, audio feature extraction, and NLP model training. This significantly reduces computation time and supports scalability.

6.2 Software Requirements

a. Front-end: React with TypeScript

The system interface is developed using React with TypeScript to ensure a responsive, user-friendly, and interactive design. It allows smooth navigation for chatbot interactions, audio/video recording, and real-time emotion feedback.

b. Back-end: Python Django, SQLite

The back-end uses Python Django for robust server-side processing and SQLite for lightweight data storage. Django ensures secure handling of user data and real-time communication between modules, while SQLite provides efficient local storage.

c. Languages: Python, TypeScript, JavaScript

Python is employed for implementing NLP, CNN, and machine learning classifiers. TypeScript and JavaScript are used for both front-end and server logic, enabling strong type checking, flexibility, and maintainable code.

d. Tools: PyCharm, VS Code

The development utilizes PyCharm for machine learning model development, training, and debugging, while VS Code supports efficient front-end and back-end development. These IDEs streamline collaboration and improve productivity.

6.3 Functional Requirements

a. Multimodal Data Collection

The system must collect input data in three modes: text (chatbot input), audio (microphone), and video (camera).

b. Text Analysis

The system must process chatbot input using Natural Language Processing (NLP) with NLTK for sentiment and emotion analysis.

c. Audio Analysis

The system must support two parallel methods:

Speech-to-Text (STT): Convert speech into text using STTX3, followed by the same NLP pipeline as text mode.

Tone Analysis: Extract audio features (pitch, MFCCs, waveform) using Librosa and classify emotions using a CNN or Random Forest model.

d. Video Analysis

The system must process video by extracting frames using OpenCV, applying a CNN-based model (DeepFace) for facial expression recognition, and classifying emotions.

e. Model Training and Testing

The system must allow training on labelled datasets across text, audio, and video modalities. It must support testing mode, where unseen input data is processed to detect emotions accurately.

f. Emotion Output

The system must output classified emotions (e.g., happy, sad, stressed, neutral) from each modality. It should also support fusion of results from multiple modalities to improve reliability.

g. User Interaction

The system must provide an interface for users to input text, audio, and video seamlessly. Results must be displayed in an easily interpretable format (e.g., reports or real-time feedback).

6.4 Non-Functional Requirements

a. Security

The system must ensure data privacy and protect user inputs (text, voice, video) from unauthorized access or misuse.

b. Performance

The platform must provide near real-time analysis of text, audio, and video data with minimal latency.

c. Scalability

The system must be designed to handle a growing number of users and multimodal inputs without performance degradation.

d. Usability

The interface should be intuitive and user-friendly, enabling users with minimal technical expertise to interact

with the chatbot, microphone, and camera.

e. Accuracy

The system must ensure high accuracy in emotion classification through robust machine learning models and effective multimodal fusion.

7. Result And Discussion

The analysis concludes that the Multimodal AI-based Mental Health Detection System offers significant advantages over traditional mental health assessment methods, particularly in terms of accuracy, real-time feedback, and personalized recommendations. By integrating text, audio, and video inputs, the system can detect emotional states more comprehensively than conventional self-reported questionnaires or manual observation. This multimodal approach allows for real-time analysis, providing users with immediate insights into their mental well-being and actionable suggestions for coping or seeking support. The automation and AI-driven classification reduce human bias, ensuring more consistent and reliable evaluations.

However, the adoption of such AI-based systems faces notable challenges. Many users may lack awareness or trust in automated mental health tools, limiting engagement. Additionally, collecting high-quality, diverse, and labelled datasets for training AI models is resource-intensive and time-consuming. Privacy and data security concerns are also critical, as sensitive personal information including text, speech, and video data needs to be protected against misuse. Finally, interpreting AI predictions in the context of human emotions is complex, and misclassification may lead to incorrect recommendations if not carefully monitored.

Overall, while the Multimodal AI-based Mental Health Detection System has the potential to revolutionize mental health monitoring by offering rapid, personalized, and accurate insights, its adoption is constrained by technical, ethical, and awareness-related challenges. Widespread implementation requires enhanced user education, strict privacy protocols, and continuous refinement of AI models to ensure safe and effective mental health support.

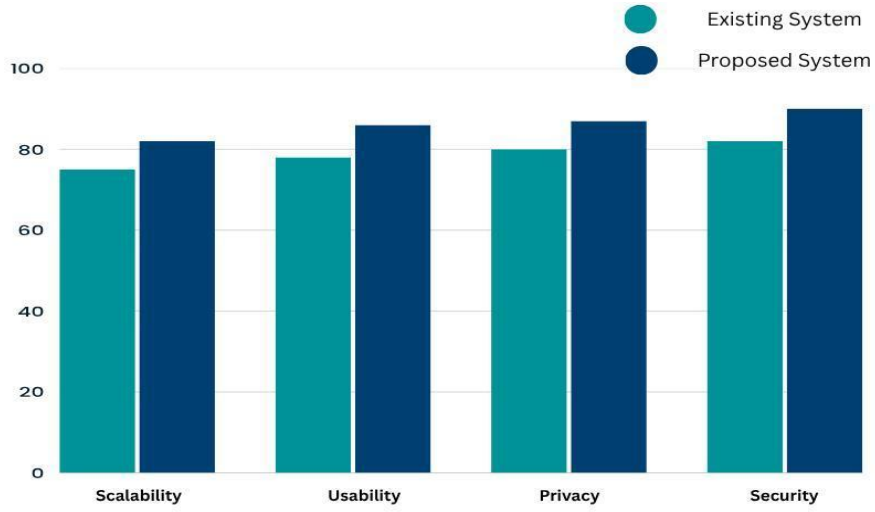

Figure 2: Comparison Chart

8.References

[1]. S. Li and W. Deng, "Facial Emotion Recognition Using DeepFace," IEEE Transactions on Image Processing, vol. 29, Jan. 2020.

[2]. T. Ghosh, S. Shubhankar, and S. Akhtar, "BERT for Mental Health Prediction on Social Media," International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), March 2020.

[3]. Z. Zhang, Z. Liu, J. Yang, and T. S. Huang, "Multimodal Emotion Recognition from Speech and Facial Expressions," Image and Vision Computing, vol. 94, Aug. 2020.

[4]. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Emotion Recognition from Multimodal Inputs: A Survey," Image and Vision Computing, vol. 81, Feb. 2019.

[5]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, June 2019.

[6]. J. Cummins, M. Scherer, and N. Cummins, "Real-Time Depression Detection from Naturalistic Speech," Proceedings of Interspeech 2018, Sept. 2018.

[7]. J. Yang, X. Yang, Y. Chen, and J. Luo, "Deep Learning for Multimodal Depression Detection," Proceedings of the ACM International Conference on Multimedia, Oct. 2017.

[8]. S. Poria, E. Cambria, and A. Gelbukh, "Multimodal Sentiment Analysis Using Deep Learning," Information Fusion, vol. 37, Sept. 2017.

[9]. M. Gjoreski, M. Luštrek, and M. Gams, "Stress Detection Using Smartphone Sensors and Speech Analysis," Proceedings of the International Conference on Mobile and Ubiquitous Multimedia, Dec. 2015.

[10]. S. Narayanan and P. Georgiou, "Speech Emotion Recognition Using MFCC and Deep Learning," IEEE Transactions on Affective Computing, vol. 4, no. 4, Dec. 2013.

[11]. M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Depression Detection on Social Media Using Text Mining," Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM), July 2013.

[12]. F. Eyben, M. Wöllmer, and B. Schuller, "Audio-Based Depression Detection Using Deep Neural Networks," Proceedings of the International Conference on Acoustic, Speech and Signal Processing (ICASSP), March 2010.