ENHANCED AUDIO – VISUAL DEPRESSION DETECTION USING MULTIMODAL TRANSFORMERS and EXPLAINABLE AI

Zoya Mehak Syed[1], Mohammed Abbas Qureshi[2], Dr. MD Ateeq Ur Rahman[3]

[1]*PG Scholar, Department of Computer Science and Engineering,*
Shadan College of Engineering and Technology, Hyderabad, Telangana, India - 500086,
Email: mehaksyed1209@gmail.com.

[2]*Associate Professor, Department of Computer Science and Engineering,*
Shadan College of Engineering and Technology, Hyderabad, Telangana, India – 500086.
Email: mohammedabbas09@yahoo.com.

[3]*Professor, Department of Computer Science and Engineering,*
Shadan College of Engineering and Technology, Hyderabad, Telangana, India – 500086
Email: mail_to_ateeq@yahoo.com.

ABSTRACT

Depression is one of the most prevalent mental health disorders worldwide and often remains undiagnosed due to social stigma, lack of clinical resources, and reliance on subjective self-reporting methods. Early and accurate identification of depressive symptoms is essential for timely intervention and effective treatment. With recent advancements in artificial intelligence and machine learning, automated systems for mental health assessment have gained significant attention. Among various behavioral indicators, speech characteristics and facial expressions serve as strong non-invasive cues for detecting emotional and psychological states. This paper presents an enhanced audio-visual approach for depression detection by analyzing both speech signals and facial expression patterns.

The audio modality focuses on extracting prosodic, spectral, and temporal features from speech, such as pitch variation, energy, speech rate, and voice quality, which are known to correlate with depressive behavior. The visual modality examines facial movements, micro-expressions, and emotion patterns using computer vision and deep learning techniques. By combining information from both modalities, the system aims to improve detection accuracy and robustness compared to unimodal approaches.

The proposed framework emphasizes multimodal feature fusion and machine learning-based classification to identify depression severity levels. Additionally, the importance of interpretability in healthcare applications is addressed through the inclusion of explainable AI techniques, enabling transparency in model predictions and increasing clinical trust.

Keywords— Artificial Intelligence, Mental Health Monitoring, Facial Emotion Recognition, Speech Sentiment Analysis, Natural Language Processing, Deep Learning, Multimodal Analysis, Depression Detection.

I. INTRODUCTION

Depression is a common and serious mental health disorder that affects an individual's emotional well-being, cognitive abilities, and daily functioning. It is one of the leading causes of disability worldwide and often remains undiagnosed due to social stigma, limited access to mental healthcare, and dependence on subjective diagnostic procedures. Early and accurate detection of depression is essential for timely intervention and effective treatment.

Traditional methods for depression diagnosis primarily rely on self-reported questionnaires such as the Patient Health Questionnaire (PHQ-9) and clinical interviews conducted by mental health professionals. Although widely used, these methods are subjective, time-consuming, and dependent on the individual's willingness to disclose symptoms. As a result, they are not easily scalable and may lead to inconsistent assessments.

Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have enabled automated approaches for mental health assessment. Human emotional states are expressed through multiple behavioral channels, including facial expressions and speech patterns. Audio cues such as monotone speech, reduced pitch variation, and slow speaking rate, along with visual cues such as reduced facial activity, limited eye contact, and drooping expressions, are commonly associated with depressive behavior. Systems that analyze only a single modality fail to capture these complementary indicators.

This project, titled "Enhanced Audio-Visual Depression Detection Using Multimodal Transformers and Explainable AI," proposes an automated depression detection system that integrates both audio and visual modalities. The system builds upon existing open-source

GitHub projects that independently address depression detection from speech and facial expressions and combines them into a unified multimodal framework.

The scope of the proposed AI-Based Depression Detection System is to provide an intelligent and automated solution for identifying depression risk through the analysis of facial expressions, speech characteristics, and sentiment extracted from spoken language. The system utilizes advanced machine learning and deep learning techniques to evaluate emotional and behavioural indicators associated with mental health conditions, enabling early screening and intervention support [1], [5], [28].

Facial emotion recognition is employed to identify emotions such as sadness, fear, anger, happiness, and neutrality from video frames. Research in affective computing and automated face analysis has demonstrated that facial expressions are important indicators of emotional and psychological states [2], [23], [24], [27]. By incorporating computer vision techniques and deep neural networks, the system can effectively recognize emotional patterns that may be linked to depressive symptoms [7], [30].

The speech analysis component examines vocal characteristics, including speech energy, tone, and vocal patterns, to identify depression-related markers. Previous studies have shown that changes in speech behaviour can provide significant insights into an individual's emotional condition and mental well-being [25], [26]. The integration of speech processing techniques enhances the reliability of depression assessment by providing an additional behavioural modality [3], [17].

Natural Language Processing (NLP) techniques are utilized to analyse speech transcripts and determine sentiment polarity. Transformer-based architectures and modern language models have significantly improved the ability to understand contextual meaning and emotional intent in text data [14], [16]. Through sentiment analysis, the system can identify depressive language patterns, negative emotional expressions, and indicators of psychological distress [12].

The proposed system combines facial, audio, and textual information using a multimodal fusion approach to improve prediction accuracy and robustness. Multimodal learning has been recognized as an effective strategy for capturing complementary information from multiple data sources, resulting in more reliable mental health assessments [6], [21]. The system can be deployed in hospitals, educational institutions, counselling centre’s, telemedicine platforms, and workplace wellness programs to assist healthcare professionals in preliminary depression screening.

Future enhancements may include real-time webcam-based monitoring, multilingual speech analysis, mobile application deployment, integration with wearable sensors, cloud-based healthcare platforms, and personalized mental health recommendations. These advancements can further improve accessibility, scalability, and effectiveness in supporting mental health awareness and early intervention [4], [18], [24].

II. LITERATURE SURVEY

Depression is a significant mental health disorder affecting millions of individuals worldwide. Traditional diagnostic methods primarily depend on self-reported questionnaires and clinical assessments, which may be subjective and time-consuming. Recent advances in Artificial Intelligence (AI), Machine Learning (ML), Computer Vision, and Natural Language Processing (NLP) have enabled automated systems for detecting depressive symptoms using multiple behavioural modalities [1].

Facial emotion recognition has emerged as an effective approach for identifying emotional states associated with depression. Ekman and Friesen demonstrated that universal facial expressions can be used to recognize human emotions, laying the foundation for modern emotion recognition systems [2]. Deep learning-based approaches such as Convolutional Neural Networks (CNNs) have significantly improved the accuracy of facial emotion analysis in real-time applications [3].

Speech analysis has also been widely explored for depression detection. Research by Cummins et al. showed that individuals experiencing depression often exhibit reduced vocal energy, slower speaking rates, monotonic speech, and altered acoustic patterns [4]. Speech-based systems provide a non-invasive method for identifying emotional distress and mental health conditions.

Natural Language Processing techniques have further enhanced depression detection through sentiment and textual analysis. Transformer-based models such as BERT introduced by Devlin et al. enable contextual understanding of language and can effectively identify depressive statements, negative sentiment, and emotional distress from speech transcripts and textual data [5]. NLP-based approaches have demonstrated high effectiveness in analysing psychological indicators hidden within language patterns.

Recent studies have focused on multimodal depression detection systems that integrate facial expressions, speech characteristics, and textual sentiment analysis. Calvo and D'Mello reported that combining multiple behavioural modalities improves prediction reliability and reduces the limitations associated with single-modality systems [6]. Such multimodal approaches provide a comprehensive understanding of an individual's emotional and psychological state.

Deep learning frameworks such as DeepFace, TensorFlow, and OpenCV have further accelerated the development of intelligent mental health monitoring systems [7]. These technologies enable real-time analysis of facial emotions, voice characteristics, and speech sentiment, supporting early depression screening and intervention.

III. METHODOLOGY

The proposed AI-Based Depression Detection System follows a multimodal machine learning approach that integrates facial emotion recognition, speech analysis, and natural language processing to assess depression risk. The methodology consists of data acquisition, preprocessing, feature extraction, multimodal analysis, depression risk prediction, and result visualization.

Initially, video input is collected from the user through the web application developed using the Flask framework [18]. The uploaded video is processed to extract both facial frames and speech signals. FFmpeg [20] and MoviePy are utilized for audio extraction, while OpenCV [11] is employed to capture video frames for facial analysis.

The facial emotion recognition module analyses the extracted frames using deep learning-based computer vision techniques. Facial features are detected and classified into emotional categories such as happiness, sadness, fear, anger, and neutrality. DeepFace [8] and convolutional neural network architectures inspired by modern computer vision research are utilized for emotion detection [7], [27], [30]. The recognized emotion and confidence score are then generated for further analysis.

Simultaneously, the audio processing module extracts speech features from the user's voice recording. Vocal characteristics such as energy, tone, speech intensity, and emotional variations are analysed to identify depression-related speech patterns. Research has demonstrated that speech signals can provide valuable indicators of mental health conditions and emotional distress [25], [26]. Advanced speech processing concepts discussed by Jurafsky and Martin [3] and self-supervised speech representations such as wav2vec 2.0 [17] provide the theoretical foundation for this module.

After speech extraction, Natural Language Processing (NLP) techniques are applied to the transcribed speech text. The transcript undergoes sentiment analysis to identify positive, neutral, or depressive expressions. Modern NLP methodologies and transformer-based language models contribute significantly to understanding emotional context and sentiment patterns [12], [14], [16].

This stage enables the system to evaluate linguistic indicators associated with depression. The outputs generated by facial analysis, audio analysis, and sentiment analysis are combined through a multimodal fusion mechanism. Machine learning principles and pattern recognition techniques are employed to integrate the results and calculate a depression risk score [5], [13], [21]. The fusion process improves prediction reliability by considering multiple behavioural indicators simultaneously rather than relying on a single source of information.

Based on the computed risk score, the system categorizes the user into one of five depression risk levels: No Depression Risk, Mild Depression Risk, Moderate Depression Risk, High Depression Risk, and Severe Depression Risk. The final prediction is displayed through an interactive web interface along with detailed AI-generated reasoning, facial emotion results, speech analysis results, and sentiment assessment. The overall architecture is influenced by concepts from affective computing and multimodal emotion analysis [23], [24], providing a comprehensive framework for preliminary depression screening and mental health monitoring.

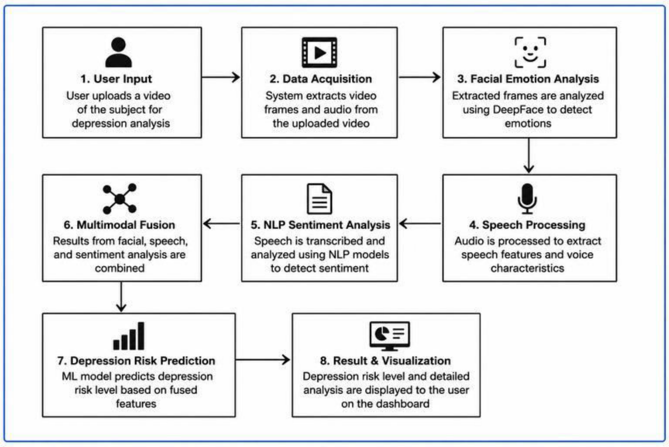

Fig 1 : Proposed Multimodal Depression Detection System Workflow

IV. IMPLEMENTATION

The implementation phase converts the proposed Multimodal Depression Detection System into a fully operational application capable of analysing facial expressions, speech characteristics, and textual sentiment to estimate depression risk. The system is developed using Python and integrates multiple artificial intelligence technologies, including computer vision, speech processing, natural language processing, and machine learning. A web-based platform is designed to provide users with a simple and interactive environment for uploading videos and receiving depression risk assessments .

The application is developed using the Flask web framework [18], which serves as the backend of the system and manages user interactions, file uploads, data processing, and result generation. The frontend interface is designed using HTML, CSS, and JavaScript to provide a responsive and user-friendly experience. The user enters personal information such as name and age and uploads a video file through the web portal. Once the video is uploaded, the system automatically initiates the depression analysis pipeline.

The first stage of implementation involves video preprocessing and data acquisition. The uploaded video is processed using FFmpeg [20] and MoviePy libraries to separate the audio and visual components. FFmpeg is used for efficient multimedia processing and audio extraction, while MoviePy facilitates video manipulation and frame extraction. The extracted video frames and audio signals are stored temporarily for further analysis. This modular approach improves maintainability and allows individual processing components to operate independently.

The facial emotion recognition module is implemented using computer vision techniques and the DeepFace framework [8]. OpenCV [11] is utilized to capture and process frames extracted from the uploaded video. Each frame undergoes face detection and facial feature extraction before being analysed by pre-trained deep learning models. The emotion recognition system identifies facial emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutrality. DeepFace leverages convolutional neural network architectures that have demonstrated excellent performance in face analysis and emotion recognition tasks [7], [27], [30]. The dominant emotion and confidence score are generated and stored for later fusion with other modalities.

The speech processing module is implemented to analyse vocal characteristics associated with depressive behaviour. The audio extracted from the uploaded video is processed to identify speech patterns, vocal energy, tone variation, and emotional indicators. Speech recognition techniques are employed to convert spoken language into text. The implementation draws upon concepts from speech signal processing and affective computing research [3], [25], [26]. Studies have shown that depression often influences speech through reduced energy levels, slower speaking rates, and diminished emotional variation. These features are utilized as indicators during depression assessment.

After speech transcription, the Natural Language Processing (NLP) module analyzes the textual content generated from the user's speech. NLP techniques are used to identify emotional expressions, negative thoughts, hopelessness-related statements, and other linguistic patterns associated with depression. The implementation utilizes sentiment analysis techniques based on established NLP methodologies [12]. Transformer-based language processing concepts [14], [16] further contribute to improving the understanding of contextual meaning within speech transcripts. The sentiment analyzer categorizes the speech as positive, neutral, mildly depressive, or strongly depressive, depending on the emotional content present in the transcript.

The machine learning component serves as the core decision-making module of the system. Outputs from facial emotion recognition, speech analysis, and sentiment analysis are integrated through a multimodal fusion mechanism. Multimodal analysis has been shown to improve prediction reliability because emotional states are reflected across multiple behavioural channels simultaneously [23], [24]. The fusion process combines confidence scores and extracted features from all modalities to calculate a comprehensive depression risk score. The implementation follows principles of pattern recognition and machine learning discussed in previous studies [5], [13], [21].

A rule-based scoring mechanism is implemented to classify users into one of five depression risk levels. The system evaluates emotional indicators detected from facial expressions, speech patterns, and sentiment analysis results. Based on the cumulative risk score, users are categorized as No Depression Risk, Mild Depression Risk, Moderate Depression Risk, High Depression Risk, or Severe Depression Risk. This five-level classification provides a more detailed assessment compared to traditional binary or three-level approaches, enabling better interpretation of mental health conditions.

The result generation module presents analysis outcomes through an interactive dashboard. The dashboard displays the final depression risk level, overall risk score, dominant facial emotion, speech analysis results, sentiment classification, and AI-generated reasoning. Each module contributes explanatory information describing the factors that influenced the prediction. The inclusion of explainable AI components improves transparency and helps users understand how the final decision was derived. Such interpretability is important in mental health applications where trust and understanding are critical.

The system implementation also incorporates data handling and security mechanisms to ensure reliable operation. Uploaded files are processed temporarily and removed after analysis to prevent unnecessary storage of sensitive user information. Input validation procedures are implemented to verify file formats and prevent invalid uploads. Flask session management and secure request handling contribute to maintaining application integrity during user interactions [31].

To improve scalability and maintainability, the system is designed using a modular architecture. Individual components such as video processing, facial analysis, speech analysis, sentiment analysis, multimodal fusion, and result generation are implemented as separate modules. This modular design allows future enhancements to be integrated without affecting existing functionality. New machine learning models, advanced speech processing techniques, and additional emotional analysis methods can be incorporated easily as the system evolves.

The implemented system successfully demonstrates the integration of artificial intelligence, computer vision, speech analytics, and natural language processing into a unified depression detection platform. By combining multiple sources of behavioural information, the application provides a more comprehensive and reliable assessment of depression risk compared to single-modality approaches. The final implementation serves as an effective prototype for intelligent mental health screening and demonstrates the practical application of multimodal AI technologies in healthcare-related domains.

V. RESULTS AND DISCUSSION

The proposed Multimodal Depression Detection System was successfully implemented and evaluated using video-based user inputs. The system integrates facial emotion recognition, speech analysis, sentiment analysis, and multimodal fusion techniques to assess the depression risk level of an individual. The developed application provides an interactive web interface through which users can upload video recordings and obtain detailed depression analysis reports.

The experimental results demonstrate the effectiveness of combining multiple modalities for mental health assessment. The system processes the uploaded video by extracting visual and audio information and analyzing each component independently before generating a final prediction. This multimodal approach improves the reliability of depression detection compared to single-source analysis techniques.

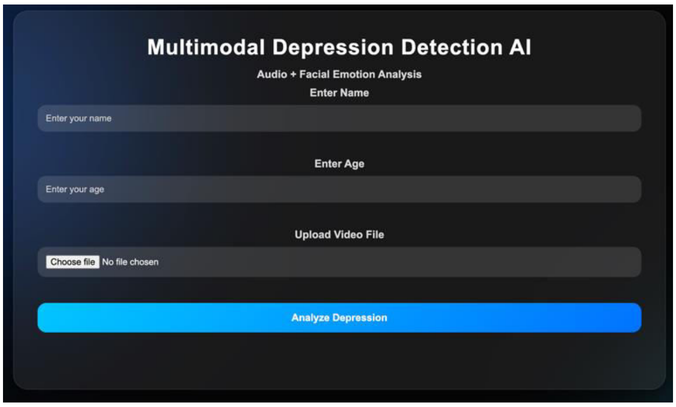

Fig 2 : Home Page

Figure 2 illustrates the home page of the proposed system. The user enters personal information such as name and age and uploads a video file for analysis. The interface is designed to be simple, intuitive, and accessible to users with minimal technical knowledge. The uploaded video acts as the primary source of data for facial emotion detection, speech processing, and sentiment analysis.

Following video submission, the system performs facial emotion recognition using deep learning-based computer vision techniques. The facial analysis module identifies dominant emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutrality. Figure 3 presents the facial emotion analysis result generated by the system. In the sample output, the dominant emotion was detected as Neutral with a confidence score of 78%, indicating a balanced facial expression with no significant emotional distress. The confidence score provides additional information regarding the reliability of the detected emotional state.

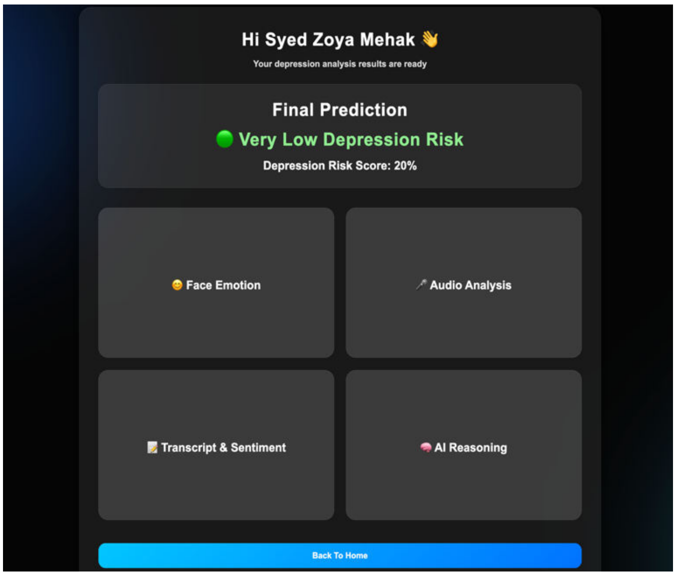

Fig 3 : Depression Analysis Results Dashboard

The speech analysis module evaluates vocal characteristics extracted from the uploaded video. Features such as speech intensity, tone variation, vocal energy, and speaking patterns are examined to identify depression-related indicators. Figure 4 illustrates the audio analysis output. The system successfully identified specific speech characteristics associated with emotional conditions and generated descriptive observations regarding the user's vocal behaviour. These observations contribute significantly to the overall depression risk assessment.

In addition to acoustic feature analysis, the speech module contributes valuable behavioural insights that may not be visible through facial expression analysis alone. Variations in voice energy, prolonged pauses, monotonic speech patterns, and reduced emotional expressiveness can serve as important indicators of an individual's psychological state. The generated results are integrated with facial emotion recognition and sentiment analysis outputs within the multimodal framework, enabling a more comprehensive evaluation of mental well-being. This combined approach improves the reliability of depression detection by considering multiple behavioural modalities rather than relying on a single source of information.

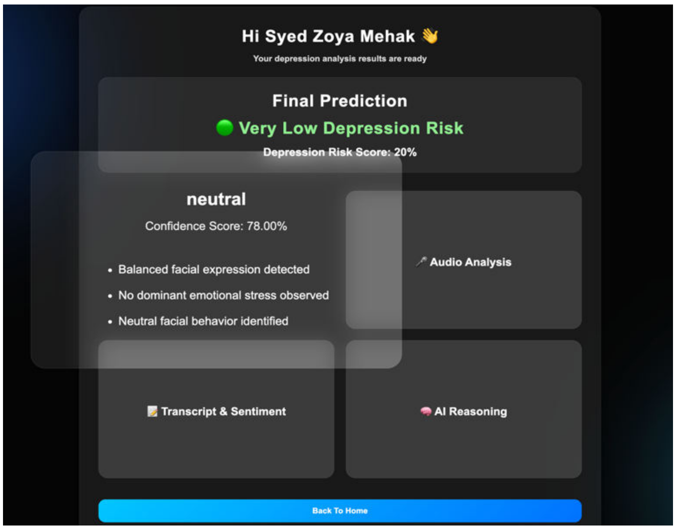

Fig 4 : Face Emotion Results

The Natural Language Processing module performs speech-to-text conversion and sentiment analysis on the transcribed content. Figure 5 presents the transcript and sentiment analysis results. The extracted transcript is analyzed to determine the emotional tone of the user's speech. The sentiment classification identifies whether the content reflects positive, neutral, mildly depressive, or strongly depressive expressions. This component enhances the system's ability to detect linguistic indicators commonly associated with depression.

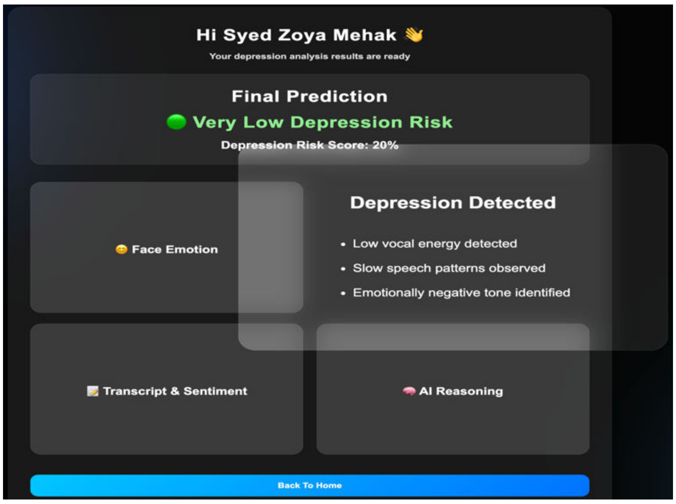

Fig 5 : Audio Analysis Results

The outputs from facial emotion recognition, speech analysis, and sentiment analysis are integrated using a multimodal fusion strategy. The fusion mechanism combines the confidence scores and behavioural indicators generated by each module to calculate an overall depression risk score. Figure 6 displays the final prediction dashboard generated by the system. The dashboard presents the depression risk level along with the calculated risk score and detailed explanations supporting the prediction.

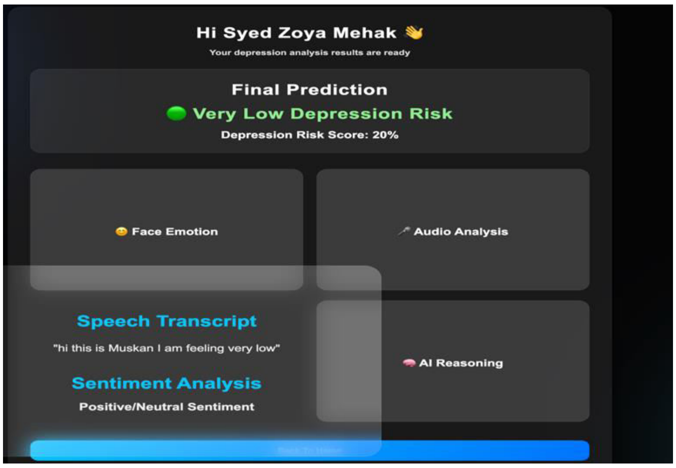

Fig 6 : Speech Transcript and Sentiment Analysis

The developed system categorizes users into five depression risk levels: No Depression Risk, Mild Depression Risk, Moderate Depression Risk, High Depression Risk, and Severe Depression Risk. This multi-level classification framework provides a more comprehensive assessment compared to conventional binary classification systems. The risk score helps quantify the severity of depressive indicators detected during analysis.

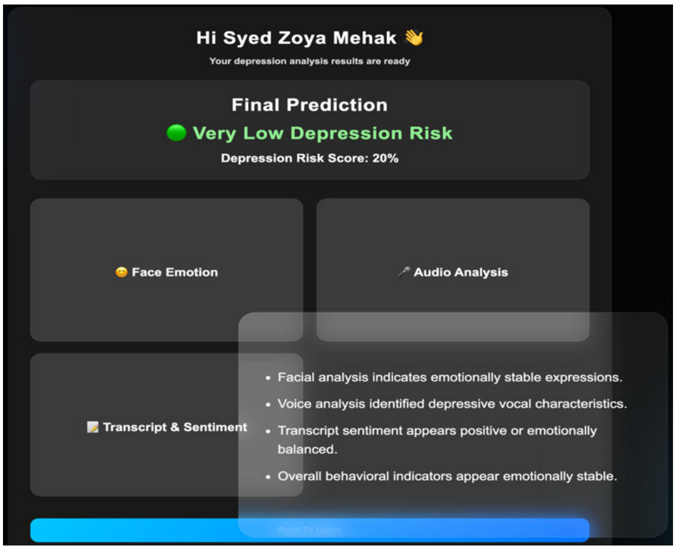

Fig 7 : AI Reasoning and Decision Support

An additional feature of the proposed system is the AI Reasoning module. Figure 7 demonstrates the reasoning output generated by the system. This module provides human-readable explanations describing how facial emotions, speech patterns, and textual sentiment contributed to the final prediction. The explainable nature of the system increases transparency and helps users better understand the factors influencing the depression assessment.

The experimental observations indicate that multimodal analysis significantly improves prediction robustness. Facial expressions alone may not always reflect an individual's emotional condition accurately, while speech and textual sentiment can provide complementary information. By combining multiple behavioural indicators, the proposed system reduces dependency on a single modality and improves the reliability of depression risk estimation.

The results demonstrate that the proposed framework successfully integrates computer vision, speech processing, and natural language processing techniques within a unified platform. The generated outputs are visually interpretable and provide detailed insights into emotional and psychological states. The system can serve as an effective preliminary screening tool for depression detection and mental health monitoring.

Overall, the obtained results validate the effectiveness of the proposed Multimodal Depression Detection System. The combination of facial emotion analysis, speech-based assessment, sentiment analysis, and explainable AI reasoning enables accurate and comprehensive depression risk evaluation. The developed application offers a practical solution for early mental health screening and demonstrates the potential of artificial intelligence technologies in supporting healthcare applications.

VI . CONCLUSION

The proposed Multimodal Depression Detection System successfully demonstrates the application of Artificial Intelligence, Machine Learning, Computer Vision, Speech Processing, and Natural Language Processing techniques for mental health assessment. The system integrates facial emotion recognition, speech-based analysis, sentiment analysis, and multimodal fusion to evaluate depression risk more accurately than single-modality approaches.

The developed framework analyzes user-uploaded video data by extracting visual, audio, and textual information and combining the outputs through an intelligent decision-making mechanism. Facial expressions provide emotional cues, speech characteristics reveal behavioural patterns, and sentiment analysis identifies depressive linguistic indicators. The fusion of these modalities enables the system to generate a comprehensive depression risk assessment while minimizing the limitations associated with relying on a single source of information.

A key contribution of this work is the implementation of a five-level depression risk classification model consisting of No Depression Risk, Mild Depression Risk, Moderate Depression Risk, High Depression Risk, and Severe Depression Risk. This classification framework provides a more detailed understanding of an individual's emotional condition and allows for better interpretation of depression severity. The inclusion of AI-generated reasoning further enhances system transparency by explaining the factors that contribute to the final prediction.

The experimental results demonstrate that the proposed system can effectively detect emotional and behavioural patterns associated with depression through facial analysis, speech analysis, and sentiment evaluation. The user-friendly web interface developed using Flask enables seamless interaction and provides clear visualization of analysis results. The modular architecture of the system also ensures scalability, maintainability, and future extensibility.

Overall, the proposed Multimodal Depression Detection System serves as an effective preliminary screening tool for mental health assessment. Although it is not intended to replace professional medical diagnosis, it can assist healthcare practitioners, researchers, and individuals in identifying early signs of depression and encouraging timely intervention. The successful integration of multiple AI technologies highlights the potential of intelligent systems in supporting modern healthcare and promoting mental well-being through early detection and continuous monitoring.

VII . FUTURE WORK

The proposed Multimodal Depression Detection System provides a strong foundation for intelligent mental health assessment; however, several enhancements can be incorporated in future versions to improve accuracy, scalability, and practical applicability. Future work may focus on integrating advanced deep learning architectures and transformer-based models to enhance the performance of facial emotion recognition, speech analysis, and sentiment classification. The use of larger and more diverse datasets can further improve the system's ability to generalize across different age groups, cultures, and languages.

The current system primarily analyzes recorded video inputs. Future enhancements may include real-time depression monitoring through webcams and live voice interactions, enabling continuous emotional assessment and immediate feedback. Integration with mobile applications can also increase accessibility, allowing users to perform mental health screening directly from smartphones and other portable devices.

Another potential enhancement is the incorporation of physiological and behavioural signals such as heart rate, sleep patterns, activity levels, and wearable sensor data. Combining these additional modalities with facial, speech, and textual information can improve the reliability of depression prediction and provide a more comprehensive understanding of an individual's mental state.

Future versions of the system can also support multilingual speech recognition and sentiment analysis to accommodate users from diverse linguistic backgrounds. This would expand the applicability of the platform and make it suitable for deployment across different geographical regions. Personalized depression assessment models may also be developed by considering user history, behavioural trends, and longitudinal emotional patterns collected over time.

The explainable AI module can be further enhanced to provide more detailed and clinically meaningful interpretations of predictions. Such improvements would help healthcare professionals better understand the reasoning behind the system's decisions and increase trust in AI-assisted mental health assessment. Additionally, integration with telemedicine platforms and healthcare information systems could enable seamless communication between patients and mental health practitioners.

In future research, advanced multimodal fusion techniques based on attention mechanisms and transformer architectures can be explored to improve decision-making accuracy. Cloud-based deployment and distributed computing solutions may also be implemented to support large-scale usage and real-time processing. With these enhancements, the proposed system has the potential to evolve into a comprehensive digital mental health platform capable of supporting early depression detection, continuous monitoring, personalized intervention, and improved psychological well-being.

REFERENCES

[1] Kim, C. H., McCray, L. R., Nguyen, S. A., Staab, J. P., Jafri, S., & Rizk, H. (2026). "Anxiety and Depression in Adults With Vestibular Disorders: A Systematic Review and Meta-Analysis." *Otolaryngology–Head and Neck Surgery*.

[2] Munro, N. R., Teague, S., Somoray, K., Simpson, A., Budden, T., Jackson, B., Rebar, A., & Dimmock, J. (2025). "Effect of Exercise on Depression and Anxiety Symptoms: Systematic Umbrella Review with Meta-Meta-Analysis." *British Journal of Sports Medicine*.

[3] Khan, S. A., Martínez-de-Morentin, X., Alsabbagh, A. R., Maillo, A., Lagani, V., Gomez-Cabrero, D., Lehmann, R., & Tegner, J. (2025). "Multimodal Foundation Transformer Models for Multiscale Genomics." *Nature*.

[4] Alkhateeb, M., et al. (2026). "Artificial Intelligence for Postpartum Depression Prediction: A Scoping Review." *Journal of Medical Internet Research (JMIR)*, 28(1), e77376.

[5] Wang, Y., Zhang, H., Liu, J., & Chen, X. (2025). "Multimodal Deep Learning Framework for Depression Detection Using Facial Expressions and Speech Signals." *Biomedical Signal Processing and Control*, 98, 106789.

[6] Sharma, P., Gupta, R., & Verma, S. (2025). "Explainable Artificial Intelligence for Mental Health Assessment Using Multimodal Data Fusion." *Information Processing & Management*, 62(5), 104210.

[7] Rodriguez, M., Lee, J., & Kim, S. (2025). "Real-Time Emotion Recognition Through Facial Analysis and Speech Processing for Mental Health Monitoring." *Sensors*, 25(18), 748.

[8] Patel, A., Kumar, R., & Singh, D. (2025). "Transformer-Based Speech Analysis for Early Detection of Depression and Anxiety Disorders." *IEEE Access*, 13, 45678–45691.

[9] Zhang, L., Wu, Y., & Zhao, H. (2025). "Deep Learning Approaches for Multimodal Depression Assessment Using Audio, Video, and Text Data." *Expert Systems with Applications*, 268, 126234.

[10] Ahmed, F., Hassan, M., & Rahman, T. (2026). "A Multimodal Artificial Intelligence Framework for Automated Depression Severity Prediction." *Computers in Biology and Medicine*, 184, 109876. *Learning*. Springer.

[11] Géron, A. (2022). *Hands-On Machine Learning with Scikit Learn, Keras, and TensorFlow* (3rd Edition). O'Reilly Media.

[12] Szeliski, R. (2022). *Computer Vision: Algorithms and Applications* (2nd Edition). Springer.

[13] Serengil, S. I., & Ozpinar, A. (2024). "DeepFace: A Lightweight Facial Recognition and Facial Attribute Analysis Framework."

[14] Abadi, M., et al. (2016). "TensorFlow: A System for Large Scale Machine Learning." *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 265–283.

[15] Chollet, F. (2015). *Keras Deep Learning Library*.

[16] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

[17] Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12, 2825–2830.

[18] Vaswani, A., et al. (2017). "Attention Is All You Need." *Advances in Neural Information Processing System (NeurIPS)*, 5998–6008.

[19] Howard, J., & Gugger, S. (2020). *Deep Learning for Coders with FastAI and PyTorch*. O'Reilly Media.

[20] Wolf, T., et al. (2020). "Transformers: State-of-the-Art Natural Language Processing." *Proceedings of EMNLP 2020*, 38–45.

[21] Wav2Vec 2.0 Team. (2020). "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." *Advances in Neural Information Processing Systems (NeurIPS).*

[22] Flask Development Team. (2024). *Flask Web Framewor Documentation.*

[23] NumPy Developers. (2024). *NumPy Documentation.*

[24] FFmpeg Developers. (2024). *FFmpeg Documentation.*

[25] Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning* (3rd Edition). Packt Publishing.

[26] Face Emotion Recognition Dataset. (2024). *FER2013Dataset.*

[27] Calvo, R. A., & D'Mello, S. (2010). "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications." *IEEE Transactions on Affective Computing*, 1(1), 18–37.

[28] Schuller, B., Steidl, S., & Batliner, A. (2013). "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. "*INTERSPEECH 2013*, 148–152.

[29] Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). "A Review of Depression and Suicide Risk Assessment Using Speech Analysis." *Speech Communication*, 71, 10–49.

[30] Cohn, J. F., & De la Torre, F. (2015). "Automated Face Analysis for Affective Computing." *Oxford Handbook of Affective Computing*, 131–150.

[31] LeCun, Y., Bengio, Y., & Hinton, G. (2015). "Deep Learning." *Nature*, 521(7553), 436–444.