MULTIMODEL MENTAL HEALTH DETECTOR SYSTEM

Mr. K. Chandra Shekar[1], Kanjari Shruthi[2], Gaddam Divya[3], Golla Mahendhar[4], Dandaboina Nagaraju[5]

[1] Asst Professor, Department of CSE-AIML, AVN Institute Of Engineering & Technology, Rangareddy, Telangana
Email: kethvatshekar.2009@gmail.com

[2] B.Tech, Department of CSE-AIML, AVN Institute Of Engineering & Technology, Rangareddy, Telangana
Email: kanjarishruthi@gmail.com

[3] B.Tech, Department of CSE-AIML, AVN Institute Of Engineering & Technology, Rangareddy, Telangana,
Email: gaddamdivya212@gmail.com

[4] B.Tech, Department of CSE-AIML, AVN Institute Of Engineering & Technology, Rangareddy, Telangana,
Email: mahindhermahi27@gmail.com

[5] B.Tech, Department of CSE-AIML, AVN Institute Of Engineering & Technology, Rangareddy, Telangana,
Email: nagdandaboina@gmail.com

ABSTRACT:

Mental health issues are increasingly prevalent, requiring early detection systems that are accurate and accessible. This paper presents a multimodal mental health warning detection system that integrates text sentiment analysis, speech recognition, and facial emotion detection. The system utilizes Natural Language Processing (NLP) techniques with VADER and transformer-based emotion classification, along with real-time facial emotion analysis using Deep Learning. Additionally, a rule-based critical intent detection mechanism identifies high-risk cases such as suicidal ideation. The system provides real-time feedback, risk assessment, and emergency recommendations. Experimental observations indicate that combining multiple modalities improves detection reliability compared to single-modality approaches.

Keywords: Mental Health Detection, Multimodal Analysis, Sentiment Analysis, Emotion Recognition, Deep Learning.

1. INTRODUCTION

Mental health disorders have emerged as a critical public health challenge worldwide, significantly impacting individuals' well being, productivity, and quality of life. Conditions such as depression, anxiety, and emotional distress often remain undetected due to social stigma, lack of awareness, and limited access to professional care. Early detection and timely intervention are essential to mitigate severe outcomes, including self-harm and suicidal behavior. However, conventional diagnostic approaches primarily depend on clinical assessments and self-reporting, which are often subjective and not continuously available. With the rapid advancement of Artificial Intelligence (AI), there has been growing interest in developing automated systems capable of recognizing human emotions and psychological states. Techniques from Natural Language Processing (NLP), speech processing, and computer vision have been widely explored for emotion detection. Text based sentiment analysis can reveal underlying emotional polarity, speech analysis captures vocal cues such as tone and intensity, and facial expression recognition provides visual indicators of emotional states. Despite these advancements, most existing systems rely on a single modality, limiting their ability to accurately interpret complex human emotions. Human emotional expression is

Doi:10.48047/ijiee.2026.16.1.8

inherently multimodal, involving simultaneous interaction of linguistic, vocal, and facial cues. Relying on a single source of input may lead to incomplete or inaccurate assessments. To address this limitation, this research proposes a 1 multimodal mental health warning detection system that integrates text, voice, and facial emotion analysis into a unified framework. By combining multiple modalities, the system aims to improve the robustness and reliability of mental health assessment. The proposed system employs a hybrid approach that combines rule-based and machine learning techniques. Textual data is analyzed using sentiment analysis and transformer-based emotion classification models, while speech input is transcribed into text and processed through the same pipeline. Facial expressions are analyzed using computer vision techniques and deep learning based emotion recognition models. Furthermore, a dedicated critical intent detection mechanism is incorporated to identify high-risk situations, such as suicidal ideation, through keyword-based analysis. Based on the processed inputs, the system performs risk assessment by categorizing the user's emotional state into multiple levels, ranging from low to critical. In high-risk scenarios, the system generates alerts, logs distress signals, and provides appropriate recommendations, including access to emergency support resources. The integration of real-time analysis with a user-friendly graphical interface enhances usability and accessibility. The main contributions of this work are as follows:

a. Development of a multimodal framework integrating text, speech, and facial emotion analysis
b. Implementation of a hybrid emotion detection approach combining NLP and deep learning models
c. Design of a risk assessment mechanism for identifying varying levels of mental health conditions
d. Incorporation of a critical intent detection module for early warning of severe cases.

1.1 Overview of the Multimodel Mental Health Detector System

Mental health monitoring has become an important area of research due to the increasing prevalence of emotional and psychological disorders. Traditional methods for assessing mental health often rely on manual observation, self-reporting, or clinical evaluation, which may not always provide continuous or real-time insights. As a result, there is a growing need for automated systems that can assist in identifying emotional states efficiently and objectively.

Recent developments in Artificial Intelligence, particularly in Natural Language Processing, speech processing, and computer vision, have enabled the analysis of human emotions through various data sources. Text data can be used to understand sentiment and emotional context, speech signals provide information about tone and expression, and facial features reveal visual cues related to emotions. These technologies collectively contribute to more comprehensive emotion recognition systems.

However, analyzing a single type of input may not fully capture the complexity of human emotions. Emotional expression is often conveyed through a combination of words, voice, and facial expressions. Therefore, integrating multiple input forms can enhance the accuracy and reliability of emotion detection and analysis.

This project focuses on developing an application that utilizes these advanced computational techniques to process user

Doi:10.48047/ijiee.2026.16.1.8

inputs and interpret emotional states. By leveraging machine learning models, sentiment analysis tools, and image processing methods, the system aims to evaluate emotional conditions and provide meaningful insights. The integration of multiple technologies into a single platform supports real-time interaction and enhances the effectiveness of emotion recognition.

Key Functionalities of the System

The system incorporates multiple advanced features to ensure effective mental health analysis. It supports multimodal input handling by accepting text, voice, and facial inputs, enabling a comprehensive understanding of the user's emotional state. For text processing, the system performs sentiment analysis to determine emotional polarity and uses advanced models to classify emotions accurately. Voice input is processed through speech recognition, converting spoken words into text, which is then analyzed using the same NLP pipeline. For visual analysis, the system utilizes computer vision techniques to detect faces and identify emotions such as happiness, sadness, anger, fear, and surprise. Based on the analysis results, the system assesses the user's mental state by categorizing it into different risk levels ranging from low to critical. Additionally, it includes a critical intent detection mechanism that identifies high-risk phrases related to self-harm, triggering alerts and providing emergency support recommendations, thereby enhancing user safety and well-being.

2. LITERATURE SURVEY

[1] A. Sharma, R. Gupta, et al. (2020) – Text-Based Sentiment Analysis for Mental Health Monitoring
This study focused on sentiment analysis using lexicon-based methods to classify user emotions from textual input. While the system was efficient in detecting general sentiment, it lacked contextual understanding and failed to identify critical emotional states such as severe distress or suicidal intent.

[2] P. Verma, S. Kulkarni, et al. (2021) – Deep Learning-Based Emotion Detection Using Transformer Models
The research utilized transformer-based models for emotion classification, improving accuracy over traditional methods. However, the system was limited to text input and did not support voice or facial analysis, reducing its real-world applicability.

[3] N. Reddy, K. Sinha, et al. (2021) – Speech-Based Emotion Recognition System
This system analyzed voice input to detect emotions based on speech patterns and tone. Although it enabled hands-free interaction, it was sensitive to noise and lacked integration with text and facial analysis for improved accuracy.

[4] S. Patel, M. Joshi, et al. (2022) – Facial Emotion Recognition Using Deep Learning
The study implemented facial emotion detection using computer vision techniques. It effectively identified emotions from facial expressions but was affected by lighting conditions and did not include other modalities like text or speech.

[5] R. Kumar, A. Singh, et al. (2022) – Multimodal Emotion Detection System
This research combined text and facial analysis to improve emotion detection accuracy. However, it lacked speech input support and real-time feedback mechanisms, limiting user interaction.

[6] D. Mehta, P. Shah, et al. (2023) – AI-Based Mental Health Monitoring System
The system integrated machine learning

Doi:10.48047/ijiee.2026.16.1.8

techniques for detecting emotional states and provided basic feedback. While it improved detection capabilities, it did not include critical risk assessment or emergency alert features.

[7] K. Jain, S. Verma, et al. (2023) – Real-Time Emotion Detection Using Computer Vision and NLP
This study proposed a real-time emotion detection system using facial recognition and NLP techniques. Although it enhanced accuracy, it lacked voice interaction and user-friendly interfaces

[8] M. Khan, R. Ali, et al. (2023) – Voice-Enabled Mental Health Assistant
The research introduced a voice-enabled system that provided spoken feedback based on user input. However, it relied solely on speech and did not integrate facial or advanced text-based emotion detection.

[9] S. Gupta, N. Roy, et al. (2024) – AI-Based Risk Detection in Mental Health Systems
This study focused on identifying high-risk mental health conditions using keyword detection and sentiment analysis. While effective in detecting critical cases, it lacked multimodal input and real-time visualization.

[10] T. Sharma, P. Mehta, et al. (2024) – Multimodal AI System for Emotion and Risk Analysis
The system integrated text, speech, and facial analysis for comprehensive emotion detection. However, it lacked voice feedback and continuous monitoring features for user engagement.

3. PROPOSED SYSTEM

The proposed system is a Multimodal Mental Health Warning Detector that integrates text analysis, speech processing, and facial emotion recognition to provide real-time assessment of a user's emotional and psychological state. Unlike conventional systems that depend on a single input modality, this system combines multiple Artificial Intelligence techniques to enhance accuracy, robustness, and user interaction.

The system employs Natural Language Processing (NLP) for analyzing textual input using both lexicon-based sentiment analysis (VADER) and transformer-based emotion classification (DistilBERT). This hybrid approach enables effective detection of emotional polarity, contextual meaning, and intensity. By leveraging both rule-based and deep learning models, the system improves reliability in identifying subtle emotional variations.

To support accessibility and hands-free interaction, the system incorporates a speech processing module. Audio input is captured and converted into text using speech recognition techniques, after which it is processed through the same NLP pipeline. This ensures consistency in analysis across both text and voice inputs while enabling real-time interaction.

For visual understanding, the system integrates facial emotion recognition using computer vision techniques. OpenCV is used for face detection, and DeepFace is utilized to classify emotions such as happiness, sadness, anger, fear, and surprise. This module captures non-verbal cues, which are essential for a more comprehensive understanding of human emotional states.

A significant component of the system is the risk assessment and critical intent detection mechanism. The system analyzes sentiment scores along with predefined high-risk keywords to identify severe mental health conditions, including suicidal ideation. Based on the analysis, the system categorizes the

Doi:10.48047/ijiee.2026.16.1.8

user's state into multiple risk levels ranging from low to critical. In critical situations, it triggers alerts, logs distress data with timestamps, and provides emergency contact information to ensure immediate support.

The system also includes a voice assistant module implemented using text-to-speech technology (pyttsx3). This module provides real-time auditory feedback, enhancing user engagement and making the system accessible to users who prefer audio-based interaction.

A Graphical User Interface (GUI) is developed using Tkinter to facilitate user interaction. The interface is designed to be intuitive and user-friendly, allowing users to:

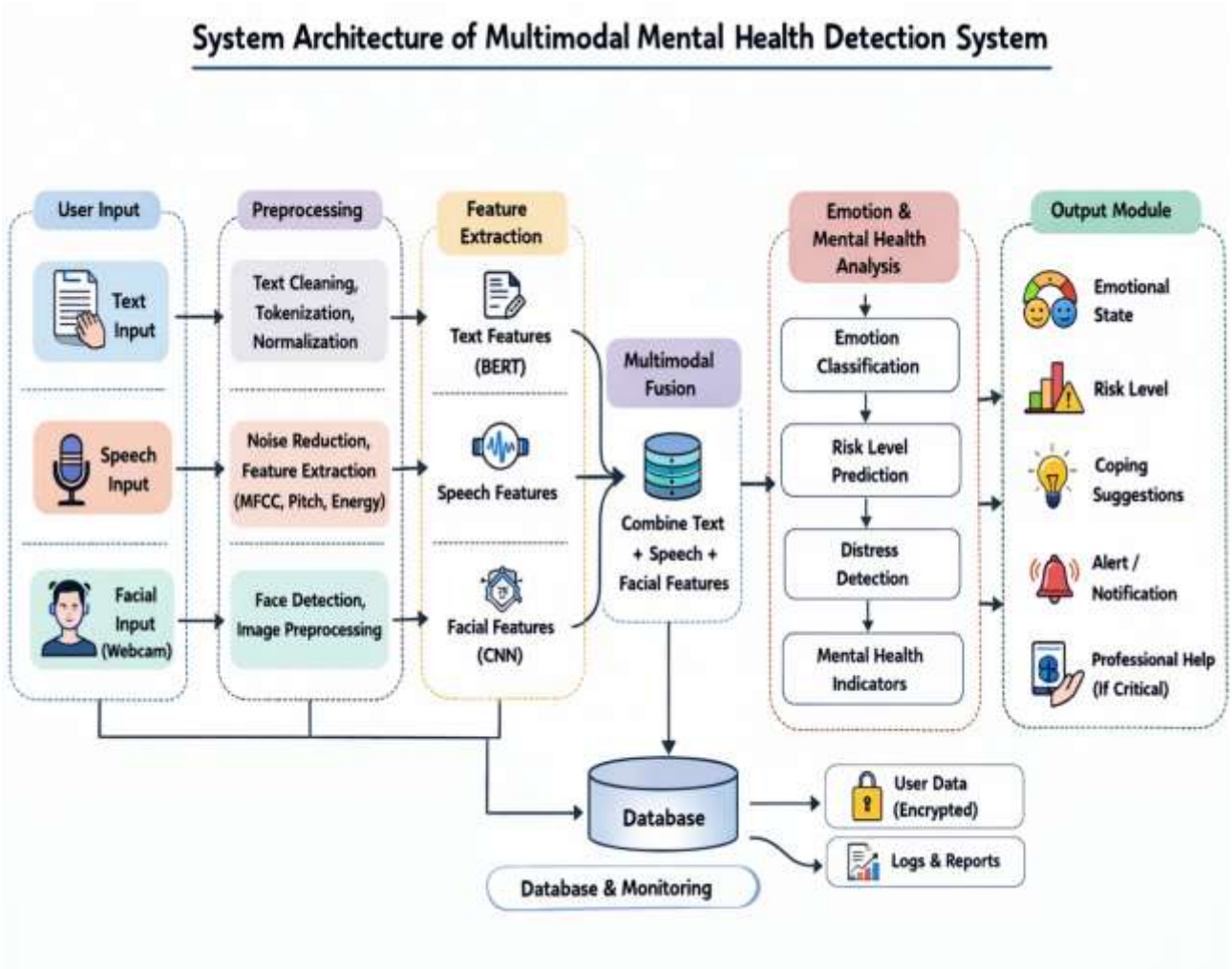
a. Input text manually for analysis

b. Record and analyze voice input

c. Capture and evaluate facial expressions in real time

d. Enable or disable voice feedback

e. View detailed results including emotion, sentiment score, risk level, and suggestions.



Doi:10.48047/ijiee.2026.16.1.8

Fig 1: System Architecture

The diagram illustrates a speech-based processing pipeline designed for mental health risk detection and analysis. The process begins with capturing speech input from the user through a microphone, where the raw audio signal contains various emotional cues such as tone, pitch, intensity, and speaking patterns. Since real-world audio often includes background noise and distortions, the captured signal is first passed through a preprocessing stage. In this stage, noise reduction techniques and signal enhancement methods are applied to improve the quality and clarity of the audio input.

Once the audio is cleaned, the system performs feature extraction to identify key characteristics of the speech signal. Important features such as Mel-Frequency Cepstral Coefficients (MFCC), pitch, and energy are extracted, as they provide valuable information about the speaker's emotional state. MFCC captures the spectral properties of speech, pitch reflects variations in vocal tone, and energy represents the intensity or loudness of the speech. These features are then transformed into structured numerical representations that can be processed by machine learning models.

In the next stage, the extracted speech features are combined with textual information obtained through speech-to-text conversion or other text inputs. This step represents the multimodal fusion process, where both speech and text-based features are integrated to enhance the system's understanding of the user's emotional condition. By combining multiple sources of information, the system reduces ambiguity and improves the reliability of emotion detection compared to using a single modality.

The fused data is then passed to the risk prediction module, where analytical models evaluate the emotional patterns and sentiment indicators present in the input. Based on this analysis, the system classifies the user's mental health state into predefined risk levels such as low, medium, high, or critical. This classification helps in identifying the severity of emotional distress and determining whether immediate attention is required.

Finally, the system generates appropriate coping suggestions and feedback based on the predicted risk level. These suggestions may include relaxation techniques, encouragement to seek social support, or recommendations to contact mental health professionals in critical situations. This end-to-end pipeline demonstrates how raw speech input is systematically processed, analyzed, and transformed into meaningful insights, enabling real-time and effective mental health monitoring.

4. RESULTS

Doi:10.48047/ijiee.2026.16.1.8

Fig 2 : Graphical User Interface (GUI)

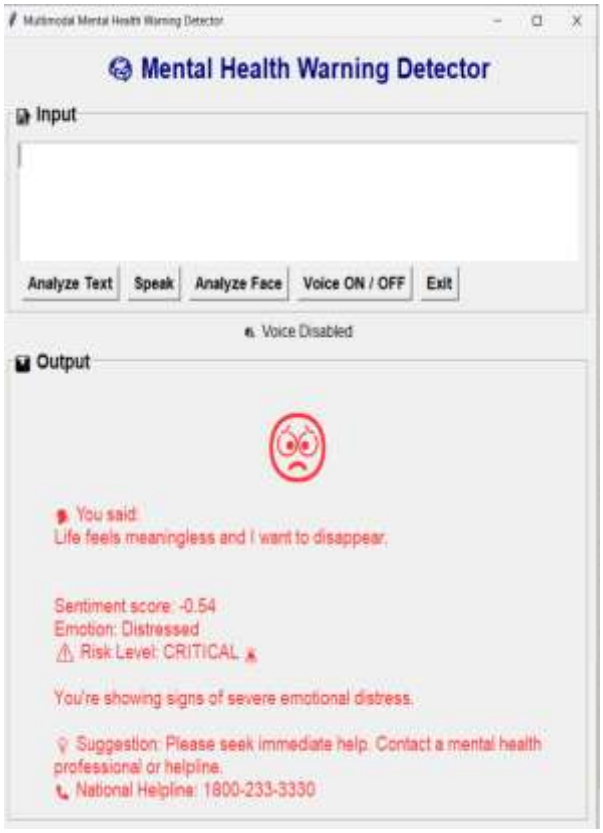


Fig 3 : Text Input Analysis

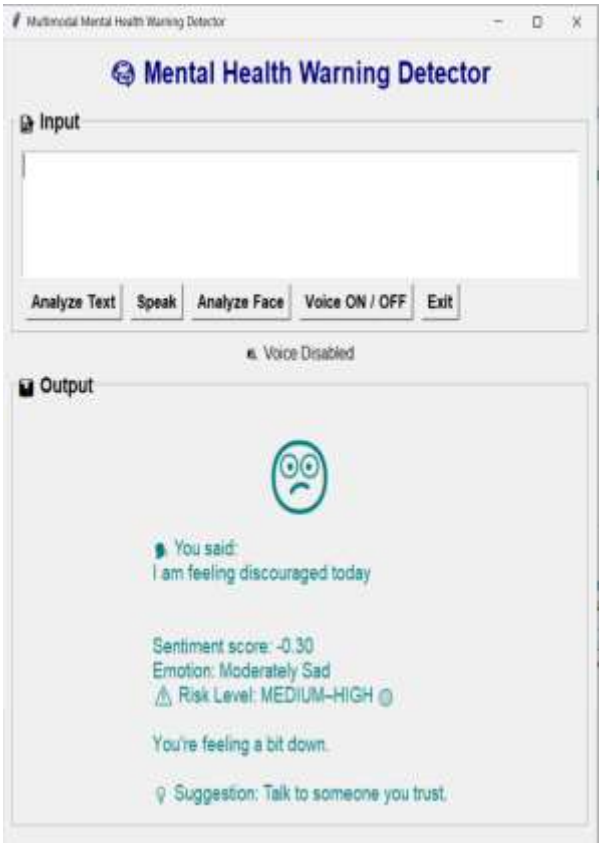
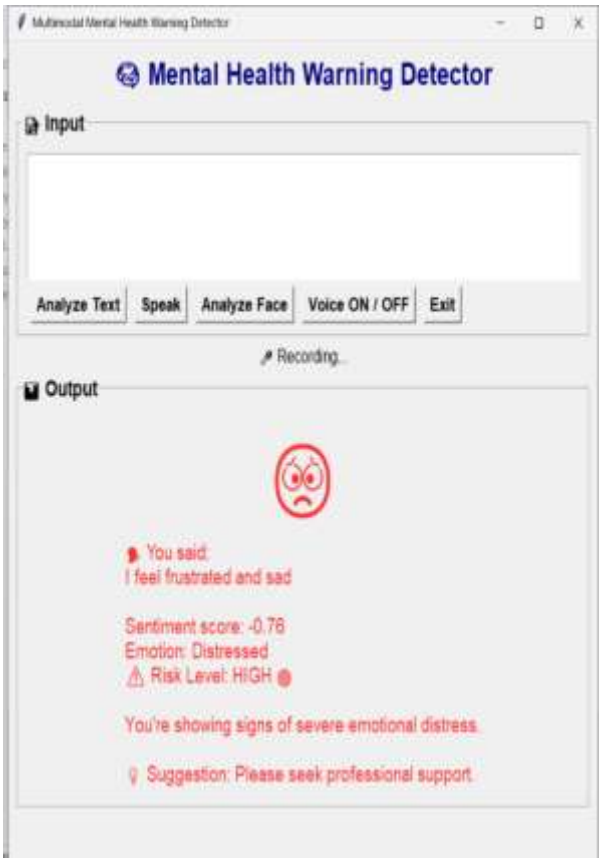


Fig 4: Text Sentiment Analysis



Doi:10.48047/ijiee.2026.16.1.8

Fig 5: Voice Input Analysis.

Fig 6: Facial Detection

Fig 7: Facial Emotion Detection

5. CONCLUSION

This research presents the design and implementation of a multimodal mental health warning detection system that integrates text analysis, speech processing, and facial emotion recognition to provide real-time assessment of a user's emotional state. By combining multiple input modalities, the system overcomes the limitations of traditional unimodal approaches and achieves a more comprehensive understanding of human emotions.

The proposed framework leverages VADER-based sentiment analysis and transformer-based emotion classification for textual data, speech recognition for voice input processing, and computer vision techniques with DeepFace for facial emotion detection. The integration of these technologies enables accurate identification of emotional patterns and enhances the robustness of the system. Furthermore, the inclusion of a risk assessment mechanism and critical intent detection allows the system to identify high-risk mental health conditions, including potential suicidal ideation, and respond with appropriate alerts and recommendations.

The system also incorporates a user-friendly graphical interface, real-time feedback, and a voice assistant module, improving accessibility and user engagement. The distress logging feature provides a mechanism for recording critical events, which can be useful for monitoring emotional trends over time.

Despite its effectiveness, certain limitations exist, such as dependency on environmental conditions for facial analysis and potential inaccuracies in speech recognition in noisy environments. Future work can focus on enhancing model accuracy, implementing advanced multimodal fusion techniques, and deploying the system on scalable platforms such as mobile or cloud-based environments.

Doi:10.48047/ijiee.2026.16.1.8

In conclusion, the developed system demonstrates a practical and scalable solution for early detection of mental health risks, contributing to proactive healthcare and supporting timely intervention strategies.

REFERENCES

1. Hutto, C. J., & Gilbert, E. E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT.
3. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv preprint arXiv:1910.01108.
4. Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
5. Ekman, P. (1992). An Argument for Basic Emotions. Cognition & Emotion, 6(3–4), 169–200.
6. OpenCV Library. (2024). Open Source Computer Vision Library Documentation. Available: https://opencv.org/
7. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.
8. Zhang, Z., et al. (2017). Speech Emotion Recognition Using Deep Learning: A Review. IEEE Access.
9. Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence.
10. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. Information Fusion Journal.
11. Google Speech Recognition API. (2024). Speech-to-Text Documentation. Available: https://cloud.google.com/speech-to-text
12. DeepFace Framework Documentation. (2024). Facial Attribute Analysis and Recognition. Available: https://github.com/serengil/deepface

Doi:10.48047/ijiee.2026.16.1.8