IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.19, No.4, October 2025, pp. 383~394
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.111407

Optimization of Multimodal Deep Learning for Depression Detection

Aditiya Hermawan[*1], Benny Daniawan[2], Edy[3], Joese Nathaniel[4]
[1,2,3,4]Faculty of Science and Technology, Universitas Buddhi Dharma, Tangerang, Indonesia
e-mail: [*1]**aditiya.hermawan@ubd.ac.id**, [2]benny.daniawan@ubd.ac.id, [3]edy.edy@ubd.ac.id, [4]joese.nathaniel@ubd.ac.id

Abstrak

Depresi merupakan kondisi kesehatan mental yang kompleks dan sering kali tidak terdiagnosis, karena gejalanya muncul melalui isyarat verbal, akustik, dan perilaku yang bersifat halus. Sistem deteksi berbasis unimodal umumnya kurang mampu menangkap spektrum gejala depresi secara menyeluruh, sehingga menghasilkan penilaian yang tidak akurat atau tidak lengkap. Penelitian ini mengusulkan sebuah kerangka kerja deep learning multimodal yang mengintegrasikan data teks, audio, dan visual untuk meningkatkan keandalan dan ketepatan deteksi depresi secara otomatis, dengan akurasi klasifikasi keseluruhan mencapai 74%. Pendekatan ini mengutamakan aspek privasi dan interpretabilitas dengan menggunakan titik-titik kunci wajah (facial keypoints) dan arah pandangan mata (gaze direction) alih-alih citra wajah mentah, serta menerapkan mekanisme attention untuk menyelaraskan dan menggabungkan fitur antar modality. Setiap modality diproses menggunakan arsitektur neural network yang sesuai dengan karakteristik datanya, dan hasilnya digabungkan dalam model fusion yang mampu mengenali pola emosional lintas-modality. Hasil eksperimen menunjukkan bahwa sistem multimodal yang diusulkan secara signifikan mengungguli model unimodal dalam performa klasifikasi. Modality visual memberikan kontribusi paling dominan terhadap akurasi deteksi, sebagaimana dibuktikan melalui analisis ablation. Temuan ini menegaskan pentingnya integrasi multimodal dalam menangkap sinyal psikologis yang kompleks dan mendukung pengembangan alat deteksi dini yang cerdas, non-invasif, dan aplikatif di platform kesehatan mental digital.

Kata kunci*—Attention Mechanism, Deteksi Depresi, Facial Keypoints, Skrining Kesehatan Mental, Multimodal Deep Learning*

Abstract

Depression is a complex and often underdiagnosed mental health condition that manifests through subtle verbal, acoustic, and behavioral cues. Traditional unimodal detection systems struggle to capture the full spectrum of depressive symptoms, often leading to inaccurate or incomplete assessments. This study proposes a multimodal deep learning framework that integrates textual, audio, and visual modalities to improve the robustness and reliability of automatic depression detection, achieving an overall classification accuracy of 74%. The approach prioritizes privacy and interpretability by using facial keypoints and gaze direction rather than raw video frames, and applies attention mechanisms to align and fuse features across modalities. Each modality is processed through dedicated neural architectures tailored to its data type, and their outputs are combined within a fusion model that learns to capture cross-modal emotional patterns. Experimental results demonstrate that the proposed multimodal system significantly outperforms its unimodal counterparts in terms of classification performance. The visual modality was found to contribute most strongly to detection accuracy, as confirmed by ablation analysis. These findings highlight the value of multimodal integration in capturing complex psychological signals and support the development of intelligent, non-invasive screening tools for use in digital mental health applications.

Keywords— *Attention Mechanism, Depression Detection, Facial Keypoints, Mental Health Screening, Multimodal Deep Learning*

Received September 22th,2025; Revised October 30th, 2025; Accepted October 31th, 2025

1. INTRODUCTION

Depression is one of the most prevalent and severe mental health disorders worldwide, affecting over 280 million people globally, according to the World Health Organization (WHO). It significantly reduces quality of life and is associated with high disability, lost productivity, and even suicide. Traditional diagnostic methods, such as clinical interviews and self-reported questionnaires like PHQ-9 or HAMD, are inherently subjective and susceptible to bias and inaccuracies [1], [2]. These limitations make early detection difficult, delaying intervention and increasing the risk of long-term consequences. This underscores the need for more objective, reliable, and scalable diagnostic tools.

In response to these challenges, artificial intelligence (AI) and deep learning have been increasingly explored to enhance depression detection. The use of multimodal data—text, audio, and video has proven particularly promising, as it allows for a more comprehensive view of an individual's mental state. Unlike unimodal approaches, which often fail to capture the complexity of depressive symptoms, multimodal systems have demonstrated superior performance. For instance, the integration of models such as BiLSTM, CNN, and IoT-based systems has enabled real-time monitoring[3], while the fusion of models like T5 and WaveNet has achieved accuracies exceeding 92%[1].

However, several key challenges persist in the development of these systems. Multimodal fusion remains difficult due to the different structures and temporal characteristics of each modality. Many models also struggle to fully leverage spatial-temporal relationships between modalities, limiting their effectiveness in capturing nuanced symptoms [4], [5]. Furthermore, interpretability remains a major barrier, as deep learning models are often considered "black boxes" and lack transparency [6]. Real-world implementation also faces difficulties such as data imbalance, lack of diversity, and privacy concerns in clinical contexts.

Recent studies have begun to address these issues through architectural innovations. For example, models like WavFace have adopted spatial-temporal attention mechanisms to better simulate clinical observations [7], and FPT-Former has improved parallel processing of multimodal inputs [8]. Meanwhile, AVA-DepressNet has demonstrated how privacy can be preserved by using facial landmarks rather than full facial imagery [9]. Despite these advances, many models remain limited in their scalability, efficiency, and generalizability, particularly in real-world settings with diverse data.

To address these gaps, this research proposes a novel multimodal deep learning framework that integrates audio, text, and visual data using a parallel transformer architecture and attention-based cross-modal fusion [10]. By leveraging facial landmarks instead of full-face images, the model promotes privacy while preserving discriminative features [9]. The study also incorporates knowledge transfer and representation learning to improve generalizability in low-resource environments, as seen in RLKT-MDD [11]. Additional emphasis is placed on interpretability [12] and predictive uncertainty estimation [13], ensuring that the frameword is no only technically robust but also ethically sound. The ultimate goal is to deliver a scalable and explainable AI System that can be deployed in real-world mental healt screening and digital support platfoams.

2. METHODS

This study aims to develop an optimized multimodal deep learning model for the early detection of depression. The proposed approach integrates audio, video, and text data using a parallel transformer architecture, coupled with attention-based cross-modal fusion and knowledge transfer techniques. The research methodology, as depicted in Figure 1, is structured into several

key stages: data collection, preprocessing, feature extraction, model development, evaluation, and validation. Each stage is designed to address critical challenges, including data fusion, model interpretability, privacy concerns, and the enhancement of model performance in real-world settings. These stages will ensure that the developed model is both accurate and practical for clinical application.

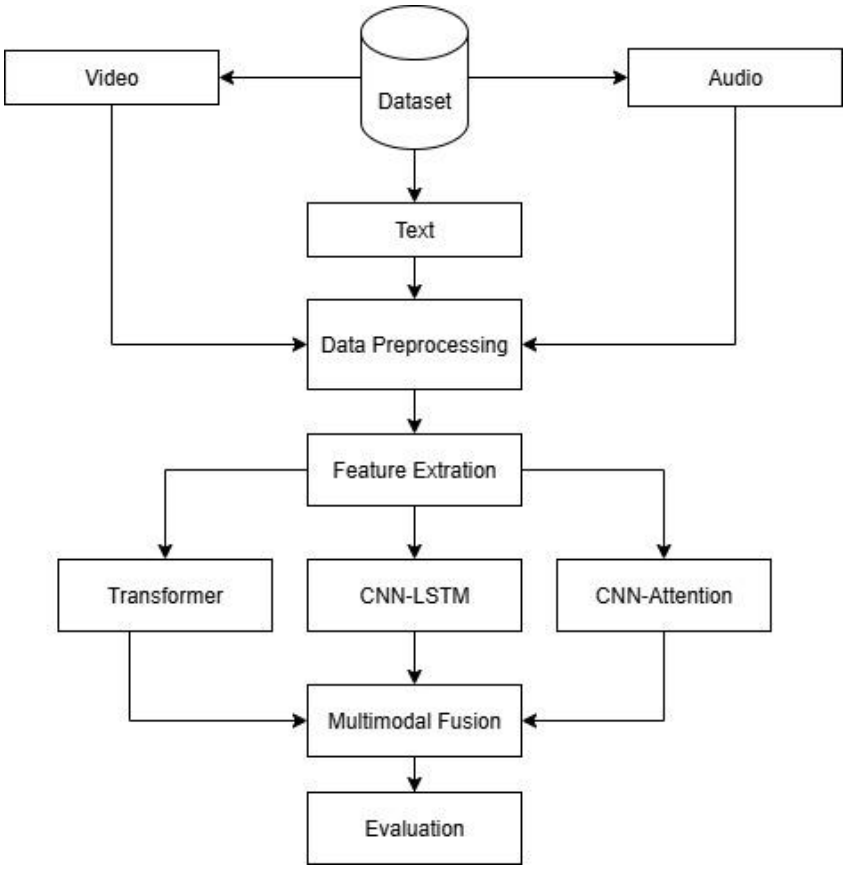

Figure 1 Reseach Method

2.1 Data Collection

The first stage of this research involves acquiring and curating multimodal datasets, primarily using the DAIC-WOZ (Distress Analysis Interview Corpus of Human and Computer Interviews), which includes audio, video, and text data. This publicly available dataset is commonly used in mental health research and provides rich, labeled data for depression detection. It includes audio recordings from clinical interviews, facial expression videos, and transcribed text from conversations, making it ideal for studying human-computer interactions and multimodal depression detection [14]. The DAIC-WOZ dataset has been extensively used in previous studies to develop and evaluate computational models for psychological distress analysis.

To ensure privacy, facial data will be processed using facial landmark detection rather than full facial images, aligning with recent advancements in privacy-preserving AI, such as AVA-DepressNet [9], which focuses on user confidentiality while enabling meaningful feature extraction. This approach ensures participant privacy is maintained, especially when handling sensitive mental health data, while still allowing the model to capture essential features for depression detection.

2.2 Preprocessing and Feature Extraction

2.2.1 Audio Preprocessing

The goal of the audio preprocessing stage is to extract prosodic features—such as pitch, tone, and rhythm that are closely linked to emotional states relevant for depression detection. This process primarily utilizes Mel-Frequency Cepstral Coefficients (MFCC), a technique that mimics human auditory perception by capturing the spectral characteristics of speech. MFCCs are effective in highlighting subtle variations in vocal tone that may indicate depressive symptoms [15]. The extraction involves segmenting the audio signal into short overlapping frames, applying the Fast Fourier Transform (FFT) to derive frequency components, and mapping these to the Mel scale using a filter bank. The final step involves the Discrete Cosine Transform (DCT) to generate compact coefficients summarizing each frame's frequency content. The general formula 1 for calculating MFCCs is shown below:

$$\mathrm{MFCC}_n = \sum_{K=1}^{K} C_k \, log(X_k) \quad (1)$$

where Ck is the Mel-scale filter, Xk is the frequency spectrum of the windowed signal, and nnn is the coefficient index.

In addition to MFCC, spectrograms are used to visualize audio signals in two dimensions time and frequency with color intensity representing amplitude. This representation is particularly valuable for identifying fluctuations in speech patterns associated with emotional states, as supported by [5]. The spectrogram is computed by applying a Fourier transform to windowed segments of the signal using a short-time analysis approach. The formula for calculating a spectrogram is expressed as:

$$S(t, f) = |\mathcal{F}\{\varkappa(t).\omega(t)\}|^2\} \quad (2)$$

where S(t,f) is the spectrogram value at time ttt and frequency fff, x(τ) is the input audio signal, and w(τ−t) is the window function centered at time ttt. These extracted features are then transformed into a vectorized format for input into the deep learning model, enabling the detection of depression-related acoustic patterns during classification.

2.2.2 Video Preprocessing

The objective of video preprocessing is to extract facial features that reflect emotional states through expressions, which are critical for identifying signs of depression. Rather than using full-face images, the method employs facial landmark detection to isolate key regions such as the eyes, mouth, and nose areas essential for capturing meaningful facial movements related to depressive cues. Tools like Dlib or OpenCV are used to detect landmark positions frame by frame, enabling efficient and privacy-preserving extraction of emotional indicators by focusing solely on landmark data, as recommended by . To process these [8] landmarks, a Convolutional Neural Network (CNN) is applied to learn spatial patterns that correlate with depressive states [16]. To further enhance temporal sensitivity, an attention mechanism is introduced to emphasize emotionally salient moments in the video, following insights from Beniwal & Saraswat (2024). This process results in a sequence of feature vectors capturing spatial-temporal facial expression dynamics, which are then fed into the model for depression prediction.

2.2.2 Text Preprocessing

The objective of text preprocessing is to extract emotional and semantic cues from conversation transcripts that are relevant to detecting depression. Text data is first tokenized using BERT's tokenizer, breaking down the conversation into smaller linguistic units that carry contextual and emotional meaning [17]. BERT (Bidirectional Encoder Representations from Transformers) is then used to capture the relationships between tokens, allowing the model to detect nuanced indicators such as negative sentiment, reduced verbal complexity, or cognitive

distortions typical signs of depression [8]. This is achieved through BERT's self-attention mechanism, which dynamically weighs each token's contribution to the sentence representation. The attention mechanism is mathematically defined as:

$$Attention = softmax\left(\frac{QK^T}{\sqrt{dk}}\right)V \quad (3)$$

where Q, K, and V are the query, key, and value matrices, respectively, and dk is the dimension of the key vectors. This formula enables the model to focus on relevant parts of the text based on their contextual importance.

Following tokenization and attention-based encoding, each token is transformed into a high-dimensional embedding vector. These embeddings are aggregated to form a comprehensive sentence-level representation that captures both syntactic and semantic features of the text. The resulting vector encapsulates emotional content and linguistic structures reflective of mental health states. This vectorized representation is then passed into downstream modeling processes, where it contributes as one modality in the multimodal depression detection system, offering deep insights into participants' verbal expressions.

2.3 Model Development

The model development in this study involves three modality-specific neural architectures, integrated into a unified multimodal framework for depression detection. For textual data, a Transformer-based architecture is utilized, leveraging the self-attention mechanism to capture long-range dependencies and contextual relationships within patient responses. Pre-trained language models such as BERT are employed to generate deep sentence representations, allowing the system to recognize subtle linguistic markers often associated with depressive states (Beniwal & Saraswat, 2024). The ability of BERT to process language bidirectionally ensures that even nuanced emotional cues embedded in text are preserved. In addition to processing language, the Transformer architecture plays a central role in aligning modalities during fusion, learning joint representations that connect semantic, acoustic, and visual patterns relevant to mental health assessment.

To process audio data, a hybrid CNN-LSTM architecture is used. CNN layers extract spatial features from MFCCs and mel spectrograms, capturing the frequency characteristics essential for identifying vocal indicators of depression, such as monotone pitch or flat prosody [15]. These spatial features are passed into LSTM layers, which model temporal dependencies like slow speech rate or prolonged pauses—common traits in depressive speech patterns [5]. The combination of convolutional and recurrent networks allows for a detailed representation of both acoustic structure and speech rhythm over time, enhancing the model's capacity to detect depression-related vocal changes. Similarly, for video data, a CNN-Attention model is employed to extract and focus on emotionally salient facial expressions. Using facial landmarks rather than raw video preserves privacy while still capturing key indicators such as gaze aversion or reduced expressivity [16] (Fang et al., 2023). The attention mechanism enhances the temporal sensitivity of the model by emphasizing moments of significant emotional variation.

Once all three modalities are independently encoded, a fusion module integrates the outputs using an attention-based strategy that learns the most informative cross-modal features. This fusion process allows the system to combine linguistic, acoustic, and visual cues into a single decision-making framework. Each modality contributes complementary information—text provides semantic content, audio offers prosodic patterns, and video encodes facial expressivity. This multimodal integration has been shown to improve performance over unimodal systems, as demonstrated in related works such as WavFace and CRADDS [3], [18]. Ultimately, the model is designed to exploit the unique strengths of each modality while mitigating their individual weaknesses, leading to a more accurate, interpretable, and privacy-aware depression detection system.

2. 4 Evaluation

The multimodal deep learning model for depression detection was evaluated using key metrics including accuracy, precision, recall, F1-score, and confusion matrix, which are widely used in mental health machine learning studies [19]. Accuracy provides a general measure of correctness, while precision and recall assess the model's ability to reduce false positives and false negatives, respectively both crucial in minimizing misdiagnosis [5]. The F1-score balances these two metrics, making it especially valuable for imbalanced datasets common in depression detection (Flores et al., 2025). Additionally, the confusion matrix offers detailed insights by displaying the distribution of correct and incorrect predictions, helping identify specific performance issues [15]. Together, these evaluation metrics ensure a comprehensive understanding of the model's diagnostic capability.

3. RESULTS AND DISCUSSION

This study utilized a benchmark dataset for depression detection containing synchronized multimodal data audio, video, and text—from clinical interviews. To preserve emotional signal integrity, only participant responses were analyzed, excluding the interviewer's speech. The data were split into training and testing sets (80:20), ensuring no participant overlap for unbiased evaluation. Each modality was independently preprocessed: textual data were embedded using a pretrained DistilBERT model, audio data were transformed into MFCC and mel-spectrograms and modeled using a CNN-LSTM architecture, while visual data consisting of 3D facial landmarks and gaze directions were normalized and processed via a CNN-Attention model. These unimodal outputs were later fused using an attention-based transformer architecture to create a final multimodal classification. The system's performance was evaluated using standard metrics such as accuracy, precision, recall, F1-score, and confusion matrix, offering comprehensive insight into the individual and combined contributions of each modality to the overall classification outcome.

3.1 Performance of Individual Modalities

3.1.1 Text Modality

The textual data in this study were derived solely from participant utterances extracted from the DAIC-WOZ dataset, with interviewer prompts removed to ensure the content reflected only the participants' mental states. After preprocessing, which involved removing repetitive phrases and irrelevant text, the responses were encoded into 512-dimensional vectors using a pretrained DistilBERT model. The classification model was then evaluated using standard metrics, achieving 69% accuracy. For the non-depressed class, it obtained a precision of 0.85, recall of 0.69, and F1-score of 0.76, while for the depressed class, the precision dropped to 0.46 with a recall of 0.69 and F1-score of 0.56. These findings indicate that the text modality performs better in identifying non-depressed individuals but tends to generate more false positives when detecting depression.

Tabel 1 Performance Metrics for Text Modality

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Non-Depressed	0.85	0.69	0.76	336
Depressed	0.46	0.69	0.56	131
Accuracy			0.69	467
Macro Avg	0.66	0.69	0.66	467
Weighted Avg	0.74	0.69	0.70	467

Figure 2 illustrates the confusion matrix for this modality. Out of 336 non-depressed instances, the model correctly classified 231 and misclassified 105 as depressed. For the depressed class, 91 out of 131 instances were correctly predicted, while 40 were misclassified as non-depressed. The distribution of misclassification suggests a tendency of the model to over-identify

depressive tendencies, possibly due to the overlapping linguistic features between emotional expressions in both classes.

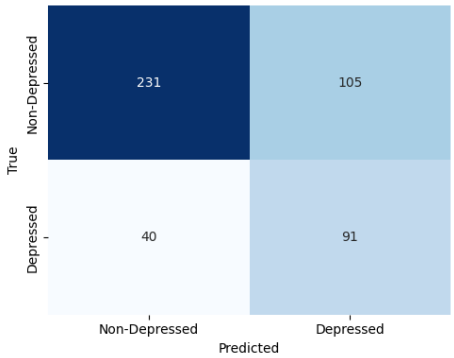

Figure 2 Confusion matrix for the text modality

3.1.2 Audio Modality

The audio modality in this study focused exclusively on participants' speech from the DAIC-WOZ dataset, deliberately excluding dialogue from the virtual interviewer to isolate vocal characteristics relevant to the speaker. Two key acoustic features Mel-Frequency Cepstral Coefficients (MFCC) and mel spectrograms were extracted, both widely used in speech emotion recognition for their effectiveness in capturing depression-related cues. The mel spectrogram (Figure 3) visualizes frequency intensity over time, revealing patterns in prosody, rhythm, and tempo, while MFCCs (Figure 4) condense spectral information to represent timbral and phonetic aspects of speech, closely mimicking human auditory processing.

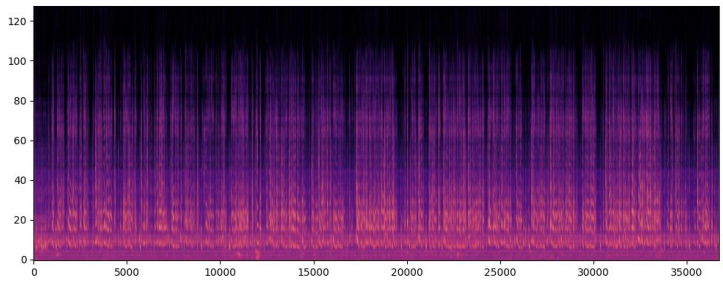

Figure 3 Mel spectrogram, representing frequency energy over time.

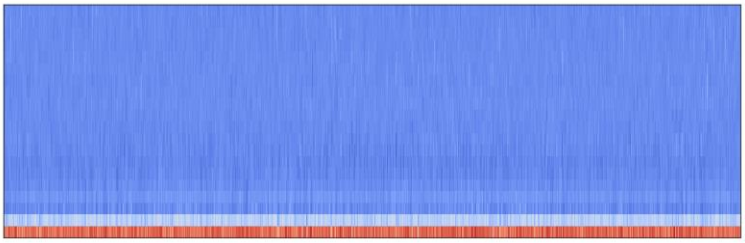

Figure 4 MFCC, capturing perceptually relevant speech features.

The audio features were processed using a CNN-LSTM model, where convolutional layers captured spatial patterns from MFCC and spectrogram inputs, and LSTM layers modeled temporal dependencies. Trained with an 80:20 train-test split, the model achieved 71% accuracy. It performed well in identifying non-depressed individuals, with a precision of 0.76, recall of 0.82, and F1-score of 0.79. However, its performance on the depressed class was notably lower (precision = 0.57, recall = 0.48, F1-score = 0.52), indicating that while the model is reliable in detecting non-depression, it has difficulty consistently capturing depression-related vocal patterns.

Tabel 2 Performance Metrics for Audio Modality

Class	Precision	Recall	F1-Score	Support
Non-Depressed	0.76	0.82	0.79	120
Depressed	0.57	0.48	0.52	60
Accuracy			0.71	180
Macro Avg	0.66	0.65	0.65	180
Weighted Avg	0.70	0.71	0.70	180

The confusion matrix in Figure 5 shows that while the audio-based model accurately classified 98 of 120 non-depressed cases, it misclassified 31 of 60 depressed samples, indicating a tendency to underpredict depression. This model effectively captures dynamic prosodic features such as tone, rhythm, and pauses which are often affected in depressive speech. However, its performance is limited by sensitivity to background noise, recording conditions, and speaker variability, as well as the inconsistent presence of vocal symptoms among depressed individuals. Compared to previous studies like [5], which reported low F1-scores for audio-only models, this study yields comparable results, aligning with findings by [17] that audio features alone are insufficient. These results reinforce the idea that while audio cues carry useful affective signals, they are best leveraged within a multimodal framework, as discussed in the subsequent sections.

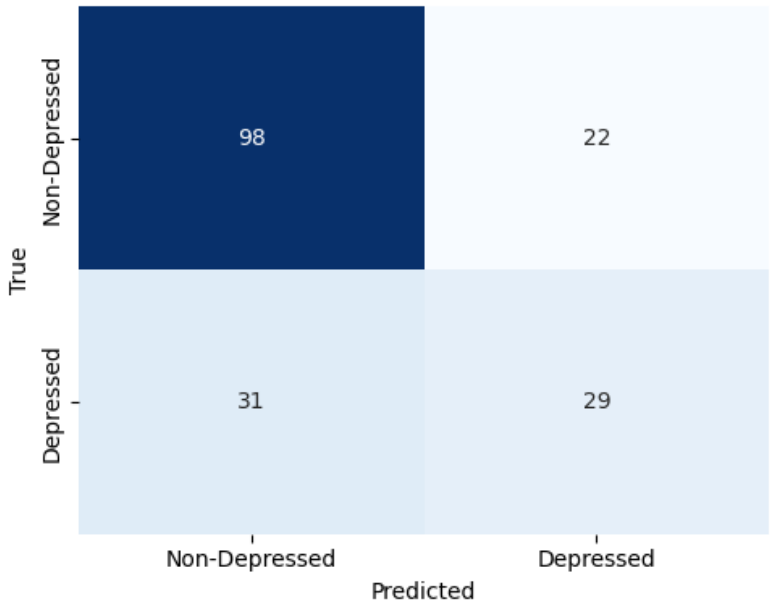

Figure 5 Confusion matrix for the audio modality

3.1.3 Visual Modality

The visual modality in this study focused on capturing micro-facial expressions and gaze behavior from the participants during their interviews. Instead of relying on full-face video frames, a privacy-preserving approach was adopted by extracting 68 facial keypoints and 3D gaze direction vectors, allowing the model to interpret emotional states without exposing raw facial imagery. This approach not only protects participant privacy but also reduces the risk of overfitting to identity-specific visual patterns.

Each video frame was processed to obtain structured data in the form of $T \times 68 \times 3$ tensors for facial landmarks and $T \times 4 \times 3$ for gaze vectors, where T denotes the number of frames. These were normalized and filtered to retain only the most emotionally relevant facial regions such as the eyes, nose, and lips. A visualization of the facial keypoints and gaze direction is shown in Gambar 7, where the blue dots represent landmark positions, and the green and red vectors indicate the left and right gaze directions respectively.

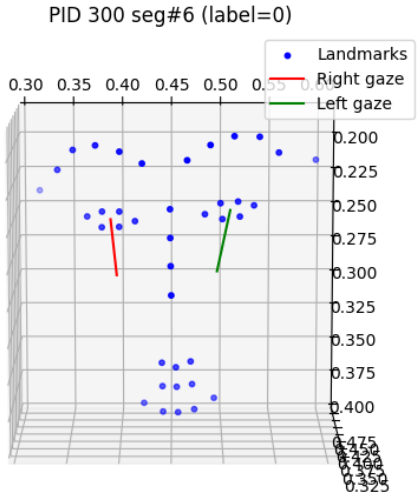

Figure 6 3D visualization of facial landmarks and gaze direction

The visual features were processed using a CNN-Attention model, where convolutional layers captured spatial patterns among facial landmarks and the attention mechanism highlighted temporally relevant facial regions linked to emotional expression. This approach enabled the detection of non-verbal depression cues such as reduced facial mobility, gaze aversion, and flattened affect signals often overlooked in text or audio. As shown in Table 3, the visual modality achieved 73% accuracy, outperforming or matching the performance of other modalities. The model demonstrated strong performance in the non-depressed class (precision = 0.83, recall = 0.75, F1-score = 0.79), and reasonable detection in the depressed class (precision = 0.58, recall = 0.68, F1-score = 0.63).

Tabel 3 Performance Metrics for Visual Modality

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Non-Depressed	0.83	0.75	0.79	120
Depressed	0.58	0.68	0.63	60
Accuracy			0.73	180
Macro Avg	0.70	0.72	0.71	180
Weighted Avg	0.74	0.73	0.73	180

Further detail is shown in Figure 7, which presents the confusion matrix of predictions for this modality. The model correctly classified 90 out of 120 non-depressed participants and 41 out of 60 depressed participants. The distribution of errors shows improved sensitivity for detecting depression, with a lower number of false negatives compared to the audio model.

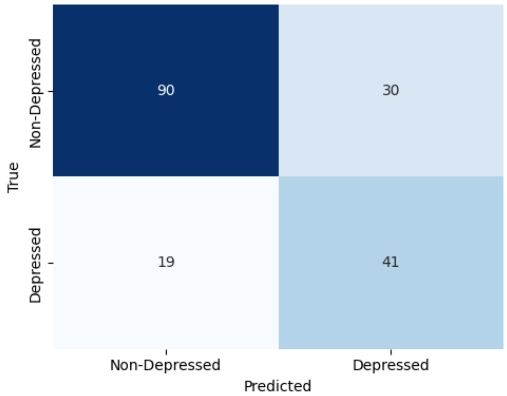

Figure 7 Confusion matrix for the visual modality

3.2 Multimodal Fusion

The multimodal fusion model combined features from text, audio, and visual modalities using an attention-based architecture to enhance depression detection. Each modality was processed independently DistilBERT for text, CNN-LSTM for audio, and CNN-Attention for visual input before being fused into a shared representation for final classification. As shown in Table 5 and the confusion matrix in Figure 8, the model achieved 74% accuracy and outperformed unimodal baselines by producing a more balanced classification between depressed and non-depressed classes. The non-depressed class reached a precision of 0.81 and recall of 0.79 (F1-score = 0.80), while the depressed class achieved 0.60 precision, 0.63 recall, and an F1-score of 0.62, demonstrating improved sensitivity to depression-related features.

Tabel 5 Evaluation Metrics for Multimodal Fusion Model

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Non-Depressed	0.81	0.79	0.80	120
Depressed	0.60	0.63	0.62	60
Accuracy			0.74	180
Macro Avg	0.71	0.71	0.71	180
Weighted Avg	0.74	0.74	0.74	180

The confusion matrix further illustrates this performance, with the model correctly identifying 95 of 120 non-depressed participants and 38 of 60 depressed individuals. The number of false positives (25) and false negatives (22) was reduced compared to unimodal systems, indicating that the fusion model achieved better generalization and class balance. These results suggest that integrating multiple modalities enables the system to capture richer emotional and behavioral cues, resulting in more reliable and robust depression detection.

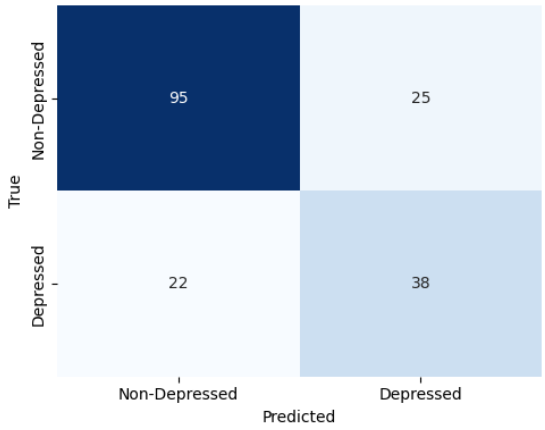

Figure 8 Confusion matrix for the visual modality

As shown in Table 6, the multimodal fusion model outperforms each unimodal model across all evaluation metrics, achieving the highest accuracy (74%) along with balanced precision, recall, and F1-score. While the text modality shows good recall, its lower precision suggests a tendency to overclassify depression. The audio model achieves moderate performance but struggles with recall, making it less sensitive to depressive cues. The visual modality offers more balanced results and the highest unimodal F1-score, yet still falls short of the fusion model. These findings highlight the benefit of integrating modalities, where the limitations of one are compensated by the strengths of others. Text contributes semantic context, audio encodes prosodic features, and visual input captures subtle non-verbal expressions such as gaze aversion and facial rigidity. Attention heatmaps support this synergy, revealing the model's ability to focus on emotionally salient time frames, which enhances temporal understanding and classification robustness.

Tabel 6 Performance Comparison of Unimodal and Multimodal Fusion Models

Modality	Accuracy	Precision	Recall	F1-Score
Text	0.69	0.74	0.69	0.70
Audio	0.71	0.70	0.71	0.70
Visual	0.73	0.74	0.73	0.73
Fusion	0.74	0.74	0.74	0.74

An ablation study further confirms the fusion model's design effectiveness. Removing the visual modality led to the greatest drop in performance, followed by audio, while excluding text had the least impact indicating the dominant role of visual cues in depression detection. Despite these promising results, the model faces notable limitations. The dataset used is relatively small and lacks demographic diversity, raising concerns about generalizability. Deep learning models remain vulnerable to overfitting, especially under data imbalance, as the depressed class is underrepresented. Furthermore, the model has not yet been validated in real-world environments such as clinical practice or digital mental health platforms. Addressing these limitations through larger datasets, real-world deployment, and techniques like explainable AI would enhance the model's utility and reliability.

4. CONCLUSIONS

This study presents a multimodal deep learning framework for automatic depression detection by combining textual, audio, and visual inputs. Through the use of facial keypoints and gaze direction instead of raw video data, the model upholds user privacy while maintaining high predictive performance. Each modality is processed through specialized neural architectures and fused using attention mechanisms that capture cross-modal emotional patterns. The results show that multimodal integration significantly improves classification outcomes over unimodal models. The visual modality proved especially impactful, with ablation analysis indicating that non-verbal cues such as facial expressions and gaze behavior offer strong indicators of depressive symptoms. These findings underscore the importance of combining multiple behavioral signals to build more comprehensive and interpretable mental health assessment tools.

However, the study is not without limitations. The relatively small and homogeneous dataset may restrict the model's generalizability across diverse populations. Additionally, the system has not been validated in real-world applications, which poses challenges for its deployment in clinical or online settings. Future work should address these issues by incorporating larger, more diverse datasets, applying explainable AI methods like SHAP or LIME to improve transparency, and testing the system in operational environments. Overall, this research contributes meaningfully to the field by demonstrating the practical potential of non-invasive, intelligent screening tools that could support early detection and intervention in digital mental health services.

REFERENCES

[1] F. Mohammad and K. M. Al Mansoor, "MDD: A Unified Multimodal Deep Learning Approach for Depression Diagnosis Based on Text and Audio Speech," *Computers, Materials & Continua*, vol. 81, no. 3, pp. 4125–4147, 2024, doi: 10.32604/cmc.2024.056666.

[2] Y. Zhao, Z. Liang, J. Du, L. Zhang, C. Liu, and L. Zhao, "Multi-Head Attention-Based Long Short-Term Memory for Depression Detection From Speech," *Front Neurorobot*, vol. 15, Aug. 2021, doi: 10.3389/fnbot.2021.684037.

[3] J. Wang *et al.*, "Automatic Diagnosis of Major Depressive Disorder Using a High- and Low-Frequency Feature Fusion Framework," *Brain Sci*, vol. 13, no. 11, Nov. 2023, doi: 10.3390/brainsci13111590.
[4] T. T. Nguyen, V. H. Q. Pham, D. T. Le, X. S. Vu, F. Deligianni, and H. D. Nguyen, "Multimodal Machine Learning for Mental Disorder Detection: A Scoping Review," in *Procedia Computer Science*, Elsevier B.V., 2023, pp. 1458–1467. doi: 10.1016/j.procs.2023.10.134.
[5] J. Ye *et al.*, "Multi-modal depression detection based on emotional audio and evaluation text," *J Affect Disord*, vol. 295, pp. 904–913, Dec. 2021, doi: 10.1016/j.jad.2021.08.090.
[6] Z. Zhang *et al.*, "Multimodal Sensing for Depression Risk Detection: Integrating Audio, Video, and Text Data," *Sensors*, vol. 24, no. 12, Jun. 2024, doi: 10.3390/s24123714.
[7] R. Flores, M. L. Tlachac, A. Shrestha, and E. A. Rundensteiner, "WavFace: A Multimodal Transformer-based Model for Depression Screening," *IEEE J Biomed Health Inform*, 2025, doi: 10.1109/JBHI.2025.3529348.
[8] Y. Li *et al.*, "FPT-Former: A Flexible Parallel Transformer of Recognizing Depression by Using Audiovisual Expert-Knowledge-Based Multimodal Measures," *International Journal of Intelligent Systems*, vol. 2024, pp. 1–13, Jan. 2024, doi: 10.1155/2024/1564574.
[9] Y. Pan, Y. Shang, Z. Shao, T. Liu, G. Guo, and H. Ding, "Integrating Deep Facial Priors into Landmarks for Privacy Preserving Multimodal Depression Recognition," *IEEE Trans Affect Comput*, 2023, doi: 10.1109/TAFFC.2023.3296318.
[10] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, in NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021.
[11] S. Yang, L. Cui, L. Wang, T. Wang, and J. You, "Enhancing multimodal depression diagnosis through representation learning and knowledge transfer," *Heliyon*, vol. 10, no. 4, Feb. 2024, doi: 10.1016/j.heliyon.2024.e25959.
[12] F. Yin, J. Du, X. Xu, and L. Zhao, "Depression Detection in Speech Using Transformer and Parallel Convolutional Neural Networks," *Electronics (Switzerland)*, vol. 12, no. 2, Jan. 2023, doi: 10.3390/electronics12020328.
[13] S. Ahmed, M. Abu Yousuf, M. M. Monowar, A. Hamid, and M. O. Alassafi, "Taking All the Factors We Need: A Multimodal Depression Classification With Uncertainty Approximation," *IEEE Access*, vol. 11, pp. 99847–99861, 2023, doi: 10.1109/ACCESS.2023.3315243.
[14] J. Gratch *et al.*, "The Distress Analysis Interview Corpus of human and computer interviews," 2014, [Online]. Available: http://www.biopac.com
[15] A. K. Das and R. Naskar, "A deep learning model for depression detection based on MFCC and CNN generated spectrogram features," *Biomed Signal Process Control*, vol. 90, Apr. 2024, doi: 10.1016/j.bspc.2023.105898.
[16] M. Fang, S. Peng, Y. Liang, C. C. Hung, and S. Liu, "A multimodal fusion model with multi-level attention mechanism for depression detection," *Biomed Signal Process Control*, vol. 82, Apr. 2023, doi: 10.1016/j.bspc.2022.104561.
[17] R. Beniwal and P. Saraswat, "A Hybrid BERT-CNN Approach for Depression Detection on Social Media Using Multimodal Data," *Comput J*, vol. 67, no. 7, pp. 2453–2472, Jul. 2024, doi: 10.1093/comjnl/bxae018.
[18] R. Flores, E. Toto, and E. Rundensteiner, "AudiFace: Multimodal Deep Learning for Depression Screening," in *Proceedings of Machine Learning Research*, 2022, pp. 1–22.
[19] R. Beniwal and P. Saraswat, "A hybrid BERT-CPSO model for multi-class depression detection using pure hindi and hinglish multimodal data on social media," *Computers and Electrical Engineering*, vol. 120, Dec. 2024, doi: 10.1016/j.compeleceng.2024.109786.