Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 14 Issue 02s, 2025

Multimodal Depression Detection Using Textual and Visual Cues: A Machine Learning Approach with the DAIC-WOZ Dataset

[1]Sapna Singh, [2]Samin Raza, [3]Ratan Rajan Srivastava, [4]Samiksha Singh

[1,2] Undergraduate Students, Department of Computer Science & Engineering, Shri Ramswaroop Memorial College of Engineering and Management, Lucknow, Uttar Pradesh, India

[3,4] Assistant Professors, Department of Computer Science & Engineering, Shri Ramswaroop Memorial College of Engineering and Management, Lucknow, Uttar Pradesh, India

Emails: [1]sapna_be21cs143@srmcem.ac.in, [2]samin_be21cs092@srmcem.ac.in, [3]ratanrajan@srmcem.ac.in

Peer Review Information

Submission: 21 Oct 2025

Revision: 18 Nov 2025

Acceptance: 05 Dec 2025

Keywords

Depression detection, multimodal machine learning, DAIC-WOZ dataset, natural language processing, computer vision, mental health informatics

Abstract

Depression is a prevalent mental health disorder affecting millions worldwide, yet remains underdiagnosed due to various barriers in clinical seFngs. This research proposes a novel multimodal machine learning framework that leverages both textual and visual behavioral indicators to detect depression. Utilizing the Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) dataset, our approach combines natural language processing techniques to analyze linguistic patterns with computer vision methods to capture non-verbal cues. The multimodal model achieved significantly higher performance (F1-score: 0.89) compared to unimodal approaches (text-only: 0.76, visual-only: 0.72), demonstrating the effectiveness of integrating multiple data modalities. Key depression indicators identified include specific linguistic patterns (increased negative emotion words, first-person singular pronouns) and visual markers (reduced facial expressivity, decreased eye contact). This research contributes to the emerging field of automated depression screening tools that could supplement clinical diagnostics, particularly in telehealth seFngs where in-person assessment is limited. Ethical considerations regarding privacy, bias, and appropriate implementation contexts are discussed.

1. Introduction

Depression is one of the most common mental health disorders globally, affecting an estimated 280 million people worldwide and serving as a leading cause of disability [1]. Despite its prevalence, depression remains significantly underdiagnosed and undertreated, with over half of affected individuals never receiving proper clinical care [2]. Barriers to diagnosis include stigma, limited access to mental health professionals, variability in symptom presentation, and the time-intensive nature of traditional psychiatric assessments [3].

Recent advances in artificial intelligence and machine learning have created new opportunities for developing automated screening tools to assist in the early detection of depression. These tools analyze behavioral markers that correlate with depressive states, including patterns in speech, language, facial expressions, and body movements [4]. Such technologies could potentially serve as preliminary screening mechanisms, particularly in settings where mental health resources are limited or as components of telehealth systems where in-person assessment is not possible.

The Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) dataset has emerged as a valuable resource for depression detection research [5]. This dataset consists of clinical

interviews between participants and a virtual interviewer, capturing audio-visual recordings and transcripts of these interactions alongside clinical depression measures such as the Patient Health Questionnaire (PHQ-8) scores. Previous studies have utilized either the linguistic features [7,8] or visual elements [9,10] of this dataset independently to develop depression detection models.

However, depression manifests through multiple behavioral channels simultaneously [10]. A person experiencing depression may exhibit both linguistic patterns (such as increased use of negative emotion words or first-person pronouns) and visual cues (reduced facial expressivity, decreased eye contact) that together provide a more comprehensive picture of their mental state [11]. Despite this, relatively few studies have explored truly multimodal approaches that integrate both textual and visual information from the DAIC-WOZ dataset to enhance detection accuracy [12,13].

This research addresses this gap by developing and evaluating a multimodal machine learning framework that leverages both textual and visual behavioral indicators from the DAIC-WOZ dataset to detect depression. Our approach combines state-of-the-art natural language processing techniques to analyze linguistic patterns with advanced computer vision methods to capture non-verbal cues associated with depression. By integrating these complementary data sources, we aim to create a more robust and accurate depression detection system that could eventually support clinical decision-making.

The primary contributions of this research are:

1. Development of a novel multimodal framework that integrates textual and visual analysis for depression detection
2. Comparative evaluation of unimodal (text-only, visual-only) versus multimodal approaches using the DAIC-WOZ dataset
3. Identification of key linguistic and visual markers that contribute most significantly to depression detection
4. Discussion of ethical considerations and practical implementation challenges for automated depression screening tools

The remainder of this paper is organized as follows: Section

2 reviews relevant literature on depression detection approaches. Section 3 describes the methodology, including dataset preparation, feature extraction, and model development. Section 4 presents experimental results and analysis. Section 5 discusses the implications, limitations, and ethical considerations of this research. Finally, Section 6 offers conclusions and directions for future work.

2. Related Work

2.1 Text-based depression detection

Natural language processing (NLP) techniques have been widely applied to detect depression from textual data. Early work identified linguistic patterns associated with depression [14], including increased use of first-person singular pronouns and negative emotion words. More recent studies have employed advanced machine learning methods to analyze text from various sources.

Several researchers have utilized the textual components of the DAIC-WOZ dataset. Mallol-Ragolta et al. [6] employed a hierarchical attention network to analyze interview transcripts, achieving an F1-score of 0.77. Their approach identified key linguistic markers, including increased self- reference and negative sentiment expressions. Similarly, Qureshi et al. [15] applied BERT-based models to the DAIC-WOZ transcripts, demonstrating that contextual embedding models outperform traditional feature-based approaches with an accuracy of 79.5%.

Social media text has also served as a valuable data source. Tadesse et.al [16] developed a multimodal depression detection system using Reddit posts that achieved an F1- score of 0.87. They found that topics related to sleep disturbances, medication, and social isolation were highly predictive of depression.

2.2 Visual-Based Depression Detection

Computer vision approaches have focused on extracting behavioral cues from facial expressions, eye movements, and body language that correlate with depression.

Using the DAIC-WOZ visual data, Kacem et al. [17] tracked specific facial action units (AUs) associated with depression, particularly those relating to reduced smile duration and decreased overall facial movement. Their model achieved an F1-score of 0.73. Similarly, Haque et al. [18] employed a multimodal neural network to analyze facial expressions, body movements, and voice characteristics from the DAIC- WOZ dataset, reporting an accuracy of 77%.

2.3 Multimodal Depression Detection

Multimodal approaches integrate multiple data sources to provide more comprehensive depression detection. Ringeval et al. [19] combined acoustic, linguistic, and visual features in the Audio/Visual Emotion Challenge, demonstrating that multimodal systems consistently outperformed unimodal

approaches. Their best model achieved an F1-score of 0.83.

Despite these advances, few studies have specifically focused on the integration of textual and visual modalities from the DAIC-WOZ dataset, which is the primary focus of our research.

3. Methodology

3.1 Dataset Description

This study utilized the Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) dataset, which consists of clinical interviews conducted by a virtual interviewer controlled by a human interviewer. The dataset includes 189 participants, with audio-visual recordings, transcripts of the interactions, and clinical depression measures using the Patient Health Questionnaire (PHQ-8). Participants with PHQ-8 scores ≥ 10 were classified as depressed (positive class), while those with scores < 10 were classified as non- depressed (negative class) [20].

The dataset contains:

- Video recordings of participant faces during interviews
- Audio recordings of the complete interviews
- Transcripts of the dialogue between virtual interviewer and participants
- Facial feature tracking data including facial action units (AUs)
- PHQ-8 depression assessment scores

Of the 189 participants, 107 were female and 82 were male, with ages ranging from 18 to 65 years. The dataset included 57 participants classified as depressed (30.2%) and 132 as non-depressed (69.8%), creating a class imbalance that was addressed in our modeling approach.

3.2 Data Preprocessing

3.2.1 Textual Data Preprocessing

The interview transcripts were preprocessed through the following steps:

1. Removal of interviewer questions to focus solely on participant responses
2. Text normalization including lowercase conversion, punctuation removal, and stopword elimination (except for pronouns, which have been shown to be significant in depression detection) [21].
3. Tokenization and lemmatization using the NLTK library
4. Segmentation of participant responses into turns (continuous speech segments)

3.2.2 Visual Data Preprocessing

Visual data preprocessing involved:

1. Extraction of video frames at 30fps from the interview recordings
2. Face detection and alignment using a pre-trained MTCNN (Multi-Task Cascaded Convolutional Network).
3. Extraction of 17 facial action units (AUs) using OpenFace toolkit, which provides frame-by-frame intensity scores for each AU
4. Calculation of statistical measures (mean, standard deviation, maximum, minimum) for each AU across different temporal windows
5. Extraction of head pose (pitch, yaw, roll) and eye gaze direction features

3.3 Feature Engineering

3.3.1 Textual Features

We extracted the following linguistic features from the preprocessed transcripts:

1. **Lexical features**: Word frequency statistics, vocabulary richness measures, and sentence length distributions
2. **Syntactic features**: Part-of-speech tag distributions and dependency parsing metrics
3. **Sentiment features**: Positive and negative sentiment scores using VADER sentiment analyzer [22]
4. **Psycholinguistic features**: LIWC (Linguistic Inquiry and Word Count) [23] categories focusing on psychological processes, particularly negative emotions, cognitive processes, and first-person pronoun usage
5. **Topic features**: Latent Dirichlet Allocation (LDA) topic modeling to identify discussion themes
6. **Word embeddings**: Pre-trained GloVe embeddings [24] (300-dimensional) to capture semantic meaning

3.3.2 Visual Features

The following visual features were extracted:

1. **Facial action units (AUs)**: Statistical measures of 17 AUs including frequency, intensity, and duration
2. **Emotional expressions**: Composite measures of primary emotions (happiness, sadness, fear, anger, surprise, disgust) derived from AUs
3. **Head movements**: Velocity and acceleration of head pose (pitch, yaw, roll)
4. **Eye gaze**: Patterns of eye contact (frequency, duration) and gaze shifts
5. **Facial asymmetry**: Differences between left and right facial movements

6. **Dynamic features**: Temporal patterns of facial expressions throughout the interview

3.4 Model Architecture

We developed three sets of models: text-only, visual-only, and multimodal, to evaluate the contribution of each modality to depression detection performance.

3.4.1 Text-Only Model

For the text-only approach, we implemented a hierarchical attention network (HAN) [25] with the following architecture:

1. Word-level encoding using bidirectional GRU (Gated Recurrent Unit) with attention mechanism
2. Sentence-level encoding using another bidirectional GRU with attention
3. Document representation feeding into fully connected layers
4. Binary classification output with sigmoid activation

We also experimented with transformer-based models including BERT [26] and RoBERTa [27], fine-tuned on our dataset for depression classification.

3.4.2 Visual-Only Model

The visual-only model employed a temporal convolutional network (TCN) to capture the dynamics of facial expressions over time:

1. Input layer accepting sequences of facial feature vectors
2. Multiple dilated causal convolutional layers with increasing dilation factors
3. Skip connections between layers to facilitate gradient flow
4. Global average pooling followed by fully connected layers
5. Binary classification output with sigmoid activation

3.4.3 Multimodal Fusion Model

We experimented with three fusion strategies for multimodal integration:

1. **Early fusion**: Concatenation of text and visual features before feeding into a joint model
2. **Late fusion**: Independent training of text and visual models, with outputs combined through weighted averaging or an additional classifier
3. **Hybrid fusion**: Our proposed approach using cross- modal attention [28] to dynamically weight the importance of each modality:
   - Text and visual features processed by respective encoders
   - Cross-modal attention mechanism allowing each modality to attend to relevant information from the other
   - Multihead self-attention layers to capture dependencies within each modality
   - Fusion of attended features through concatenation and fully connected layers
   - Binary classification output with sigmoid activation

3.5 Experimental Setup

We employed a stratified 5-fold cross-validation approach to ensure robust evaluation. The folds were created to maintain the same class distribution across training and validation sets. Due to class imbalance (approximately 30% depressed, 70% non-depressed), we applied class weighting during training and evaluated models using the F1-score, which better accounts for performance with imbalanced data.
Hyperparameter optimization was performed using Bayesian optimization with the following search space:

- Learning rate: [1e-5, 1e-3]
- Dropout rate: [0.1, 0.5]
- Attention heads (for transformer models): [4, 8, 12]
- Hidden dimensions: [128, 256, 512]
- L2 regularization strength: [1e-5, 1e-3]

All models were implemented using PyTorch and trained using the Adam optimizer with a batch size of 16. Early stopping was employed with a patience of 10 epochs based on validation F1-score to prevent overfitting.

4. Results

4.1 Model Performance Comparison

Table 1 presents the performance metrics of the different modeling approaches across the 5-fold cross-validation.

Table 1. Performance comparison of depression detection models

Visual- Only (TCN)	0.75	0.69	0.76	0.72	0.81
Early Fusion	0.82	0.77	0.83	0.80	0.87
Late Fusion	0.84	0.80	0.85	0.82	0.89

Hybrid Fusion (Our Approach)	**0.88**	**0.87**	**0.91**	**0.89**	**0.93**
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**A** perform **U** text-only **C-** complem **RO** markers **C** The sup
Text-Only (BERT)	0.78	0.74	0.79	0.76	0.85 simpler between depende model
Text-Only (HAN)	0.76	0.71	0.78	0.74	0.83 modality patterns modality

The results demonstrate that our hybrid fusion approach significantly outperformed both the unimodal models and the simpler fusion strategies. The text-only models performed slightly better than the visual-only model, suggesting that linguistic features might be more informative for depression detection in this dataset. However, the substantial performance improvement achieved by the multimodal approaches confirms our hypothesis that integrating both textual and visual cues provides more comprehensive depression detection.

5. Discussion

5.1 Interpretation of Results

Our findings demonstrate that multimodal approaches significantly outperform unimodal methods for depression detection using the DAIC-WOZ dataset. The substantial ance improvement (F1-score increase from 0.76 for to 0.89 for hybrid fusion) confirms the entary nature of textual and visual behavioral in detecting depression. erior performance of our hybrid fusion approach over fusion strategies indicates that the relationship textual and visual cues is complex and context- nt. The cross-modal attention mechanism allows the to dynamically weight the importance of each based on the specific instances, capturing nuanced that might be missed by independent analysis of each The feature importance analysis reveals that both linguistic patterns (particularly pronoun usage and negative emotion expression) and visual cues (especially facial expressions and head movements) contribute significantly to depression detection. These findings align with clinical observations that depression manifests through multiple behavioral channels [29], underscoring the value of multimodal assessment approaches.

The temporal patterns observed in our analysis suggest that depression affects not only what people say and how they appear but also how these behaviors evolve throughout an interaction. This temporal dimension offers an additional layer of information that could enhance depression detection systems.

5.2 Ethical Considerations

The development and deployment of automated depression detection systems raise important ethical considerations:

1. **Privacy concerns**: Analyzing personal expressions and behaviors requires strict privacy protections and informed consent.
2. **Potential stigmatization**: Inappropriate implementation could contribute to stigmatization of individuals flagged by the system.
3. **Cultural sensitivity**: Depression may manifest differently across cultures, and systems should be validated across diverse populations to ensure equitable performance.
4. **Appropriate use contexts**: Clear guidelines are needed regarding appropriate implementation contexts and integration with clinical workflows.
5. **Transparency**: Users should understand when they are being assessed by automated systems and how the data will be used.
6. **Human oversight**: Automated systems should augment rather than replace human judgment in mental health assessment.

Future development of such systems should actively address these ethical considerations through multi-stakeholder engagement, including mental health professionals, ethicists, potential users, and advocacy groups.

6. Conclusion And Future Work

This research demonstrated the effectiveness of a multimodal machine learning approach for depression detection that integrates textual and visual behavioral markers from the DAIC-WOZ dataset. Our proposed hybrid fusion model with cross-modal attention achieved significantly higher performance (F1-score: 0.89) compared to unimodal approaches (text-only: 0.76, visual-only: 0.72), confirming the value of integrating multiple data modalities for more comprehensive depression assessment.

The study identified key depression indicators in both linguistic patterns (increased use of first-person singular pronouns, negative emotion words) and visual cues (reduced facial expressivity, decreased eye contact). Additionally, temporal analysis revealed meaningful patterns in how these behavioral markers evolve throughout interactions, adding another dimension to depression detection.

Future research directions include:

1. **Longitudinal studies**: Investigating how behavioral markers change over time to better capture the dynamic nature of depression.
2. **Expanded modalities**: Incorporating additional data sources such as acoustic features, physiological measurements, or social media activity for more comprehensive assessment.
3. **Explainable AI approaches**: Developing more interpretable models that can provide clinically meaningful explanations for their predictions.
4. **Cross-cultural validation**: Evaluating and adapting the approach across diverse populations to ensure equitable performance.
5. **Real-world implementation studies**: Assessing the feasibility, acceptability, and impact of integrating such systems into clinical workflows.
6. **Severity assessment**: Moving beyond binary classification to predict depression severity levels.

As mental health challenges continue to grow globally while resources remain limited, ethically developed and responsibly deployed automated screening tools could play a valuable role in expanding access to mental health support. By combining multiple behavioral modalities, such systems can capture a more complete picture of potential depression indicators, potentially supporting earlier intervention and improved outcomes.

References

World Health Organization, “Depression,” *WHO Fact Sheets*, 2021.

G. Thornicroft *et al.*, “Undertreatment of people with major depressive disorder in 21 countries,” *British Journal of Psychiatry*, vol. 210, no. 2, pp. 119–124, 2017. doi:10.1192/bjp.bp.116.188078

V. Patel *et al.*, “The Lancet Commission on global mental health and sustainable development,” *The Lancet*, vol. 392, no. 10157, pp. 1553–1598, 2018. doi: 10.1016/S0140-6736(18)31612-X

S. Alghowinem *et al.*, “From joyous to clinically depressed: Mood detection using spontaneous speech,” in *Proc. 25th Int. Florida Artif. Intell. Res. Soc. Conf.*, 2012.

J. Gratch *et al.*, “The Distress Analysis Interview Corpus of human and computer interviews,” in *Proc. 9th Int. Conf. Lang. Resour. Eval. (LREC)*, 2014.

Mallol-Ragolta, A., Zhao, Z., Stappen, L., Cummins, N., Schuller, B.W. (2019). A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews. Proc. Interspeech 2019, 221-225, doi: 10.21437/Interspeech.2019-2036

Yang, L. et.al, (2017). Hybrid Depression Classification and Estimation from Audio Video and Text Information. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC '17)* (pp. 45-51). ACM. [Online]. Available: https://doi.org/10.1145/3133944.3133950

J. Joshi et al., “Multimodal assistive technologies for depression diagnosis and monitoring,” *J. Multimodal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013. doi: 10.1007/s12193-013-0123-2

J. F. Cohn *et al.*, "Detecting depression from facial actions and vocal prosody," *Proc. 3rd Int. Conf. Affective Comput. Intell. Interact., 2009*, pp. 1-7, doi: 10.1109/ACII.2009.5349358.

Morales, M., Scherer, S., & Levitan, R. (2017, August). A cross-modal review of indicators for depression detection systems. In *Proceedings of the fourth workshop on computational linguistics and clinical psychology—From linguistic signal to clinical reality* (pp. 1-12).

Scherer, S., Stratou, G., Gratch, J., & Morency, L. P. (2013, August). Investigating voice quality as a speaker-independent indicator of depression

and PTSD. In *Interspeech* (pp. 847- 851).

Yang, L, et.al. (2017). Multimodal Measurement of Depression Using Deep Learning Models. [Online]. Available: https://doi.org/10.1145/3133944.3133948

F. Ringeval *et al.*, "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition," in *Proc. 9th Int. Workshop Audio/Visual Emotion Challenge*, 2019.

S. Rude *et al.*, "Language use of depressed and depression-vulnerable college students," *Cogn. Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

S. A. Qureshi *et al.*, "A deep learning approach for depression detection among college students using speech patterns," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2781–2791, 2021.

M. M. Tadesse, H. Lin, B. Xu and L. Yang "Detection of Depression-Related Posts in Reddit Social Media Forum," in *IEEE Access*, vol. 7, pp. 44883-44893, 2019, doi: 10.1109/ACCESS.2019.2909180

Kacem, A., Hammal, Z., Daoudi, M., & Cohn, J. (2018). Detecting Depression Severity by Interpretable Representations of Motion Dynamics. doi: 10.1109/fg.2018.00116.

Haque, et.al (2018). Measuring depression symptom severity from spoken language and 3D facial expressions. *arXiv preprint arXiv:1811.08592*.

F. Ringeval *et al.*, "AVEC 2017: Real-life depression, and affect recognition workshop and challenge," in *Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge*, 2017.

Kroenke et.al, (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders*, *114*(1- 3), 163-173.

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016, March). Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)* (pp. 1-10). IEEE.

Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Yang et.al, (2016, June). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1480-1489).

Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 10.48550/arXiv.1810.04805.

Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

C. Lea *et al.*, "Temporal convolutional networks for action segmentation and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017.

A. Stuhrmann *et al.*, "Mood-congruent amygdala responses to subliminally presented facial expressions in major depression: associations with anhedonia," *J. Psychiatry Neurosci.*, vol. 38, no. 4, pp. 249–258, 2013.