JSCDM

JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

Vol. 7 No. 2 (2026) 447-460
https://publisher.uthm.edu.my/ojs/index.php/jscdm

A Deep Learning-Based Decision Support System for Early Depression Screening from Multimodal Clinical Interviews

Mujtaba Zuhair Al-Amshaw[1], Inbithaq Ahmed Shakir[2], Wael Ali[3]*, Hesham A. Sakr[4,5], Muneera Altayeb[6], Ibrahim A. Gomaa[7,8]

[1] *Department of Computer Techniques Engineering, Imam Al-Kadhim University College (IKC), IRAQ*
[2] *Department of Computer Artificial Intelligence, College of Science, Mustansiriyah University, Baghdad, IRAQ*
[3] *American University in the Emirates, College of Education DIAC, Dubai, UNITED ARAB EMIRATES*
[4] *Faculty of Computing and IT, Sohar University, OMAN*
[5] *Department of Communication and Electronics, Nile Higher Institute of Engineering and Technology, Mansoura, EGYPT*
[6] *Faculty of Engineering, Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, JORDAN*
[7] *Faculty of Computing and IT, Sohar University, OMAN*
[8] *Computer Science Department, Al-Obour High Institute for Computer and Informatics, Cairo, EGYPT*

*Corresponding Author: wael.ali@aue.ae
DOI: https://doi.org/10.30880/jscdm.2026.07.02.025

Article Info

Received: 19 April 2026
Accepted: 5 June 2026
Available online: 30 June 2026

Keywords

AI-driven mental health, multimodal depression screening, deep learning, clinical interviews, real-time decision support, data-driven psychiatry, smart assessment systems, transformer-LSTM fusion

Abstract

The complexity of the behavioral, emotional, and linguistic features of depressive disorders is still a challenge for depression screening by clinical interviews. Self-report questionnaires, like the PHQ-9, have a number of limitations, including self-report bias and inadequate behavioral representation. In this paper, we thus propose the development of a Deep Learning-Based Decision Support System (DL-DSS) for early depression screening from multimodal clinical interview data. The proposed framework combines the transformer-based text encoding, CNN-based acoustic and visual feature extraction, and LSTM-based temporal modeling approaches to perform multi-modal analyses of transcript, speech, and facial behavior. The system was tested with the DAIC-WOZ benchmark data set. Experimental results show that the proposed multimodal DL-DSS model achieved 87.9% classification accuracy, 89.0% F1-scores, and 92.0% AUC-ROC, outperforming the conventional PHQ-9-based and single-modality deep learning approaches, and also the proposed model has a real-time inference latency of less than 120 ms per interview segment. These results lend support to the effectiveness of AI-based multimodal deep learning for intelligent and scalable depression screening within current clinical decision-support systems.

1. Introduction

The high rate of modern digital health technology development has added to the rise of the expectation of intelligent mental health screening systems that can identify patients who are at risk of depression with more accuracy, scalability, and early intervention opportunities. It is because of the prevalence of major depressive disorder and better outcomes of the early diagnosis that automated depression screening is being integrated into clinical psychology, telemedicine, and workplace mental health programs [1][2]. However, screening all the patients by using the traditional questionnaire-based approaches is a hard task because of the patient self-report bias, cultural variation in the expression of symptoms, and high sensitivity to masking or exaggeration of symptoms [3].

The multimodal analysis of clinical interviews is one of the essential parameters to determine the accuracy of depression screening, as it can measure the speech content of verbal behavior, prosodic features of speech, as well as nonverbal behavioral patterns when conducting the evaluation. The conventional screening systems usually rely on the PHQ-9 questionnaire alone. Screenings based only on questionnaires are not able to provide a response to patient affect, speech patterning, or facial expression change during the clinical interaction, even though the method is as easy to administer [4]. Such a condition is likely to lead to poor diagnostic accuracy and large false-negative results because of unimodal screening methods.

Current depression screening methods are primarily based on questionnaires or on a single modality machine learning method that is based on speech, text, or facial behavior alone [5][6]. These techniques provide greater automation, but they are not necessarily able to capture the complex cross-modal behavioral patterns that arise as a function of depression severity when conducted in a clinical interview [7][8]. Further, current systems suffer from limitations in feature engineering that are manually created and are processed offline, which restricts their use in clinical settings where real-time processing is required [9][10]. To overcome these drawbacks, the proposed DL-DSS framework combines multimodal transformer-based text encoding, CNN-based feature extraction from the audio and visual representations, and temporal modeling to ensure reliable, robust, and accurate depression screening in real time and adaptively [11][12].

Scientists have developed multimodal de-depression screening in order to counter such drawbacks, and these measures are based on the combination of text, audio, and video characteristics adapted dynamically during the clinical interview. The method of feature engineering and manual feature extraction has enhanced the elimination of classification mistakes and the screening performance [13][14]. However, the methods are usually rooted in manually crafted characteristics, and they do not support the rich and high-dimensional representations that deep neural networks are trained on.

New possibilities in the recent development of artificial intelligence and data-driven mental health have made it possible to create new opportunities for intelligent depression screening. The deep learning algorithms may process the sophisticated multimodal data of the clinical interaction and predict the existence of depression and also calculate the degree of severity dynamically [15][16]. The paper offers a multimodal AI-based deep learning decision support architecture that integrates the principles of transformer-based text encoding, CNN-based acoustic features extraction, and LSTM-based temporal models, and dynamically evaluates the risk of depression to improve the quality of mental health screening, its accuracy, and fairness.

The conceptual architecture of the DL-DSS system, illustrated in Figure 1, shows the way in which the clinical interview data are interacted with, but the key components that shall be considered should be the audio, video, and transcript inputs, the AI-based feature extraction units, and the multimodal fusion classification systems.

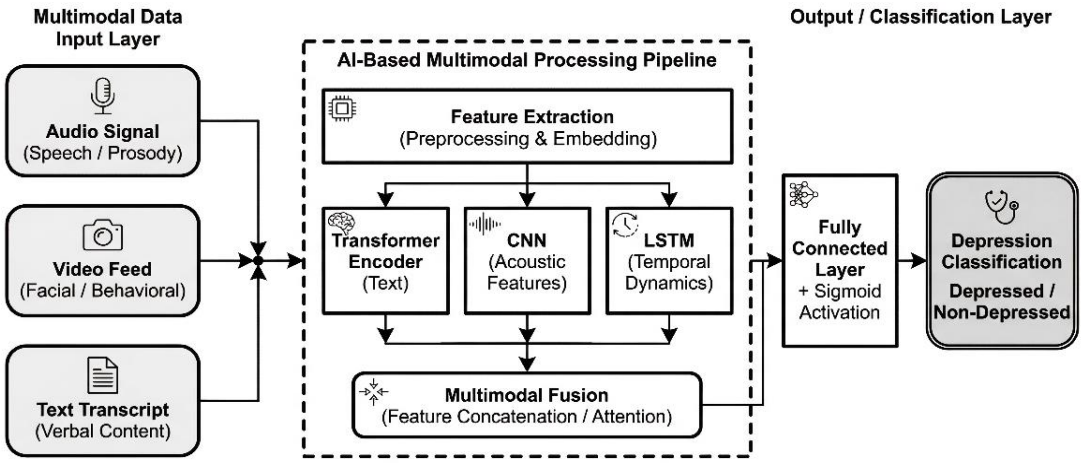

Fig. 1 *AI-based multimodal decision support architecture for early depression screening*

Although AI-based mental health screening has seen progress recently, some key challenges have yet to be addressed. Current self-report self-administered questionnaires are susceptible to self-report bias and fail to provide real-time analysis of the patterns of behavior in the clinical interview [3][4]. Existing single-modal deep learning techniques are also unable to represent temporal and multimodal connections between speech, text, and facial expression [13] [17]. Besides, the majority of current approaches are offline processing and fail to provide the ability for real-time adaptive depression screening for uncertainty-aware prediction. Thus, a multimodal intelligent framework that can integrate multi-modality clinical interview data to enhance the screening accuracy, lower the number of false negatives, and facilitate real-time clinical decision-making is needed. The primary contributions of this work are summarized as follows:

- Development of a real-time AI-driven multimodal decision support framework for early depression screening from clinical interviews.
- Integration of transformer-based text encoding, CNN-based acoustic feature extraction, and LSTM-based temporal modeling for adaptive depression risk assessment.
- Utilization of the DAIC-WOZ benchmark dataset comprising audio, video, transcripts, and PHQ-9 scores.
- Comprehensive comparative evaluation against conventional PHQ-9-only and single-modality deep learning strategies.
- Demonstration of significant improvements in classification accuracy, F1-score, AUC-ROC, and false negative reduction.

Other parts of this paper are structured in the following way. Section 2 will include a literature review of the associated studies in intelligent mental health screening and AI clinical decision support. Section 3 comprises the problem statement and objectives of the research. Section 4 presents a proposal of a DL-DSS methodology along with mathematical modeling and algorithm development. Section 5 contains a statement of experimental validation and a performance analysis. Finally, Section 6 concludes the research and gives future research directions.

2. Related Research

According to automated detection of depression research studies, there is an increasing trend in addressing multimodal integration, classifier accuracy, and clinical scalability of addressing high-quality machine learning and computational psychiatry approaches [9]. It was mentioned that initial screening policies were based on questionnaire administration and the interpretation of scores according to the experience of clinicians and cut-off points [10]. These methods, though useful in simple workflows, failed to scale and had the inability to define complex multimodal behavioral patterns in patients with either subclinical or atypical depression.

Handcrafted acoustic and linguistic feature engineering has also improved the states of depression severity prediction, including, but not limited to, vocal jitter, shimmer, speech rate, and the number of negative emotion words in speech [11]. openSMILE and LIWC enable manual feature extraction, allowing the establishment of correlations between behavioral indicators and depression status [12]. Nevertheless, all these are labor-intensive and more so applied to offline analysis, as opposed to real-time screening when there is a clinical interaction between a clinician and the patient.

Recent multimodal screening strategies, which are developed using deep learning, provide better feature representation and reduce errors in classification. End-to-end neural networks eliminate manual feature engineering and increase the accuracy of screening [13]. Simple fusion techniques, however, are used in most of the methods, and they cannot capture detailed cross-modal temporal dependencies of speech, text, and facial expressions.

There are now intelligent systems for assessing mental health developed using artificial intelligence methods. Transformer models are used to create text contexts to predict depression severity and sentiment analysis [14][15]. Clinical interview transcripts are processed by large pretrained language models to determine the linkage between linguistic patterns and the outcome of depression. Table 1 provides a comparative description of the existence of depression screening architectures as well as the associated benefits and drawbacks.

Table 1 *Comparative overview of depression screening architectures*

Architecture Type	Key Features	Limitations
PHQ-9 Questionnaire Only	Simple administration, low cost	Self-report bias, no behavioral data
Single-Modality Deep Learning	Automated feature extraction from one modality	Missing cross-modal information
Rule-Based Clinical Assessment	Deterministic diagnostic criteria	Limited scalability, clinician-dependent
Simulation-Driven Feature Analysis	Accurate behavioral feature extraction	Computational overhead, offline only
AI-Driven Multimodal DL-DSS (Proposed)	Real-time multimodal fusion, temporal modeling, and adaptive classification	Higher implementation complexity, data privacy concerns

Although there has been an advancement in the area of depression screening, there are gaps that are essential. Clinical interviews are hard to detect early: depression is manifested in complicated and dynamic signs, such as the speech pattern, facial expression, emotional mood, and spoken words. Conventional questionnaire-based approaches tend to be too rigid to respond to patient presentation variation, which may lower the accuracy of detection in the early stage and elevate false negatives [16]. Despite the multimodal deep learning methods suggested, the majority of them use predefined features and offline training, thus being less applicable to real-time clinical interviews. Specifically, there is limited research on real-time multimodal fusion with text encoding by a transformer, feature extraction by a convolutional neural network (CNN), and temporal modeling by a long short-term memory (LSTM). Thus, the paper fills these gaps by creating the proposed Deep Learning-Based Decision Support System (DL-DSS) framework of real-time depression screening based on text, audio, and video data.

3. Proposed Methodology

Through its proposed research, a hybrid, benchmark-based data framework is created where a multimodal deep learning-based decision support system (DL-DSS) in depression screening is developed on the basis of an advanced AI-driven adaptive classification architecture. In contrast to the traditional methods, the suggested system combines cross-modal attention-based fusion, probabilistic inferences, with tempo-sensitive consistent modeling of the state, so that it becomes capable of providing robust and real-time risk assessment of depression.

- Transformer-based text encoding, convolutional neural network (CNN)-based acoustic and visual feature extraction, and gated recurrent temporal modeling are incorporated into the system to encode dynamic behavioral patterns of clinical interview sessions. The design of its methodology is to maximize classification accuracy and minimize clinical unreliability due to low false negatives and uncertainty-aware prediction mechanisms [17].
- The proposed analysis is structured as follows:
- A distributed multimodal data acquisition framework is used to capture heterogeneous data streams in the context of a clinical interview, such as audio recording, video recording, transcribed speech, and auxiliary behavioral cues. All these cues mark the changing linguistic, emotional, and cognitive state of the patient in real time [18].
- The lifecycle modeling of the depression screening process involves the raw signal preprocessing, modality-specific deep feature extraction, cross-modal attention-based fusion, temporal state modeling, and adaptive probabilistic classification with a trade-off between the accuracy and interpretability of screening as well as computational efficiency.
- Three architectural simulation models are considered:
- Model A: Conventional PHQ-9 questionnaire-based screening using fixed threshold-based classification.
- Model B: Single-modality deep learning-based screening using either audio-only or text-only features.
- Model C (Proposed): A multimodal DL-DSS framework integrating transformer-based encoding, CNN-based feature extraction, cross-modal attention fusion, and temporal modeling for adaptive and real-time depression prediction.

The novelty of the proposed DL-DSS framework is that it combines the transformer-based text encoding, CNN-based multimodal behavior feature extraction, temporal sequence modeling, and adaptive cross-modal fusion into an integrated real-time depression screening framework. Distinct from current multi-modal depression detection techniques that predominantly rely on offline analysis and fusion of just a single set of features [13][17], the

proposed framework is able to conduct dynamic multi-modal interaction analysis during the clinical interviews, while also improving the classification accuracy, false-negatives reduction, uncertainty-aware prediction, and real-time inference efficiency. Moreover, the proposed architecture also proposes an integrated Depression Screening Performance Index (DSPI), which is a single performance index for evaluating the quality of the screening by integrating the classification performance, latency, robustness, and predictive reliability in a unified index.

3.1 Mathematical Modeling of the AI-Based Depression Screening System

The overall system quality of the AI-driven multimodal screening pipeline is defined in Eq. (1):

$$Q_{total} = w_1 Q_{accuracy} + w_2 Q_{fairness} + w_3 Q_{robustness} + w_4 Q_{efficiency} + w_5 Q_{confidence} \quad (1)$$

where $Q_{accuracy}$ represents classification accuracy performance, $Q_{fairness}$ denotes demographic fairness of the screening model, $Q_{robustness}$ represents system resilience to variations in recording quality, $Q_{efficiency}$ reflects computational efficiency for real-time deployment, $Q_{confidence}$ represents prediction reliability based on model uncertainty and w_i are weighting coefficients such that $\sum w_i = 1$.

The AI-based multimodal optimization objective is expressed as Eq. (2):

$$L_{DL-DSS} = L_{cls} + \lambda_{FN} \cdot \max(0, FN_{rate} - FN_{threshold}) + \lambda_{unc} L_{unc} + \lambda_{temp} L_{temp} \quad (2)$$

where L_{cls} represents classification loss (binary cross-entropy), λ_{FN} is the false negative penalty coefficient, FN_{rate} represents false negative rate, $FN_{threshold}$ is the acceptable false negative threshold for clinical safety, L_{unc} represents uncertainty-aware loss, L_{temp} represents temporal consistency loss and $\lambda_{unc}, \lambda_{temp}$ are weighting parameters.

The multimodal feature fusion optimization objective is defined as Eq. (3):

$$\min_{\theta} \; L_{fusion}(\theta) + \beta_1 FLOPs(\theta) + \beta_2 Memory(\theta) + \beta_3 Latency(\theta) + \beta_4 L_{align}(\theta) \quad (3)$$

where θ represents model parameters, L_{fusion} denotes cross-modal attention fusion loss, $FLOPs$ represents computational complexity, $Memory$ represents memory usage, $Latency$ represents inference delay, L_{align} represents cross-modal alignment loss ensuring consistency across modalities and β_i are trade-off coefficients.

Inference energy consumption per clinical interview segment is modeled as Eq. (4):

$$E_{segment} = P_{GPU} T_{inf} + P_{CPU} T_{pre} + P_{memory} T_{io} + \gamma_{dyn} E_{adaptive} \quad (4)$$

where $P_{GPU}, P_{CPU}, P_{memory}$ represent the power consumption of computational units, T_{inf}, T_{pre}, T_{io} denote inference, preprocessing, and I/O time, respectively, $E_{adaptive}$ represents dynamic energy consumption due to adaptive modality selection and γ_{dyn} is a scaling coefficient.

The depression classification accuracy is defined in Eq. (5):

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

where TP, TN, FP and FN represent true positives, true negatives, false positives, and false negatives respectively.

The F1-score is computed as Eq. (6):

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (6)$$

where $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$.

Clinical utility index is defined as Eq. (7):

$$\eta_{clinical} = \frac{AUC_{actual} \cdot (1 - FN_{rate})}{1 + \sigma_{avg}} \tag{7}$$

where AUC_{actual}represents achieved AUC, FN_{rate}represents false negative rate and σ_{avg}represents average predictive uncertainty.

The reinforcement learning reward function for adaptive multimodal fusion is modeled as Eq. (8):

$$R_{adaptive} = R_{accuracy} - \lambda_{latency} L_{norm} - \lambda_{FN} FN_{rate} + \lambda_{conf}(1 - \sigma_{avg}) + \lambda_{div} D_{modal} \tag{8}$$

where $R_{accuracy}$ represents classification accuracy reward, L_{norm} represents normalized latency, FN_{rate}represents the false negative rate, σ_{avg}represents prediction uncertainty, D_{modal}represents the modality diversity score and λ terms are weighting coefficients.

The multimodal feature vector is defined as Eq. (9):

$$S(t) = [z_T(t), z_A(t), z_V(t), h(t), c_{context}(t)] \tag{9}$$

where $z_T(t)$, $z_A(t)$ and $z_V(t)$represent encoded text, audio, and video features, respectively, $h(t)$represents the temporal hidden state and $c_{context}(t)$represents contextual embedding capturing conversational dynamics.

The Depression Screening Performance Index (DSPI) is defined as Eq. (10):

$$DSPI = \alpha \cdot AUC + \beta \cdot (1 - FN_{norm}) + \gamma \cdot (1 - \sigma_{avg}) + \delta \cdot \frac{1}{T_{inf}} + \eta \cdot Robustness \tag{10}$$

where AUC represents classification capability, FN_{norm} represents the normalized false negative rate, σ_{avg}represents uncertainty, T_{inf}represents inference latency, $Robustness$represents performance under noise and $\alpha, \beta, \gamma, \delta, \eta$are weighting parameters.

Pseudocode 1: Multimodal Depression Screening with Adaptive Fusion and Uncertainty Control
Input: Clinical interview segments c_i, multimodal inputs (A_i, V_i, T_i)
Parameters: Model weights θ, false-negative threshold $\text{FN}_{th} = 0.10$, uncertainty limit ε
Output: Depression classification $\hat{y}$, system performance metrics
BEGIN
Step 1: Initialization
$\theta \leftarrow$ initialize model parameters
$\omega \leftarrow$ initialize modality weights
Step 2: Multimodal Data Acquisition
For each clinical segment c_i, acquire: (A_i, V_i, T_i)
Step 3: Preprocessing
Audio: $A_i \rightarrow$ MFCC,spectrogram
Text: $T_i \rightarrow$ tokenization $\rightarrow$ transformer encoding
Video: $V_i \rightarrow$ facial feature extraction
Step 5: Adaptive Modality Weighting
$\omega = \text{Softmax}(\text{Attention}(z_A, z_T, z_V))$
Step 6: Multimodal Fusion
$F_{\text{fusion}} = \omega_A z_A + \omega_T z_T + \omega_V z_V$
Step 7: Temporal Modeling
$h_t = \text{GRU}(F_{\text{fusion}}, h_{t-1})$
Step 8: Prediction and Uncertainty Estimation
$(\mu, \sigma) = \text{Dense}(h_t)$
$\hat{y} = \sigma_g(\mu)$ (where $\sigma_g$denotes the sigmoid function)
Step 9: Decision Logic
If $\sigma \geq \varepsilon \Rightarrow$ Defer decision (low confidence)
Else if $\hat{y} > 0.5 \Rightarrow$ Depressed
Else $\Rightarrow$ Non-depressed
Step 10: Loss Computation and Model Update
$\mathcal{L}_{total}$ computed as per Eq. (2)
$\theta \leftarrow \theta - \eta \nabla \mathcal{L}_{total}$

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 447-460

Step 11: System Quality Update

$Q_{total} \leftarrow$ update using Eq. (1)

Step 12: Evaluation

After processing all segments: Compute DSPI as per Eq. (10)

END

With this kind of multimodal deep learning model, the depression screening system would be able to process real-time clinical interview data and achieve maximum classification accuracy despite differences in patient presentation.

3.2 Process Flow Diagram

The multimodal depression screening process is shown in Fig. 2 using AI. The data from clinical interviews is preprocessed and processed in parallel with the help of transformer-based text encoding, CNN-based audio analysis, and video feature extraction. An attention-based mechanism is used to fuse the extracted features in order to learn cross-modal relationships. The merged representation is then forwarded to an LSTM-based temporal modeling module to capture sequential behavioral dependencies and temporal depression patterns throughout the clinical interview session. The system provides predictions of depression status with confidence estimates, coupled with a feedback loop that allows the model to be continually updated and improve screening over time.

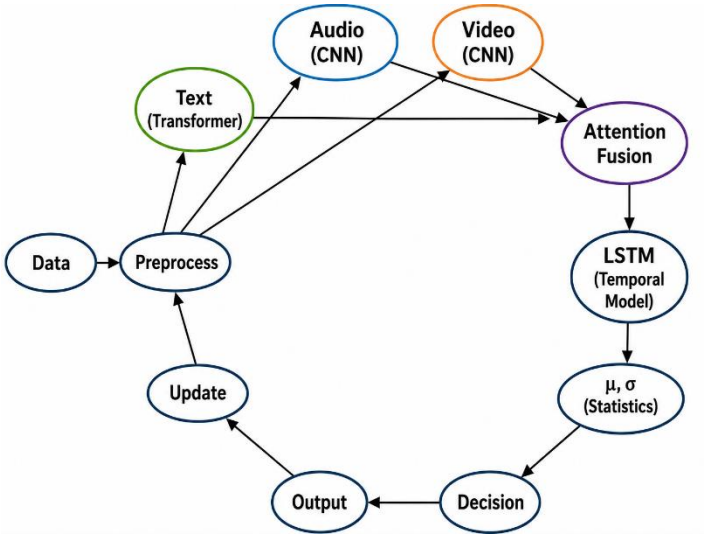

Fig. 2 *AI-based multimodal depression screening process*

The improved multimodal depression screening pipeline is shown in Fig. 2. The system commences with the multimodal data acquisition, preprocessing, and simultaneous features extraction through transformer and CNN models. The cross-modal attention fusion module combines features of a certain modality and then presents them to a temporal model unit to learn patterns in a sequence. The predictive layer, which is probabilistic, gives a depression probability and a measure of uncertainty. The feedback loop also guarantees the continuous improvement of the model and its performance.

3.3 Dataset Description and Preprocessing

The DAIC-WOZ benchmark dataset, frequently used for depression detection research, was used for evaluating the proposed DL-DSS framework [19]. The dataset consists of 189 clinical interview sessions gathered via human-computer interactions, including the audio recordings, facial video data, the interview transcript, the behavioral annotations, and scores from a depression assessment (PHQ-9). The participants were classified as depressed or not depressed according to severity cut-off points for the PHQ-9 score that are widely used in depression screening studies.

The data set was split into 70% training, 15% validation, and 15% test data sets. In model development, stratified sampling and balanced batch training strategies were adopted to minimize the class imbalance effects. Preprocessing also involved noise filtering, normalization of the transcripts, temporal segmentation, and feature standardization for audio, text, and video modalities.

3.4 Implementation Details

The proposed DL-DSS framework was achieved in Python programming with the use of TensorFlow and PyTorch deep learning libraries. Text encoder was realized based on the transformer; the acoustic and visual feature extraction modules were realized based on CNN (convolutional neural network) with multiple convolution layers and pooling layers. The temporal behavioral modeling part used a two-layer LSTM network including 128 hidden units.

The Adam optimizer was used for training the multimodal model with 50 epochs, a batch size of 32, and a learning rate of 0.0001. To prevent overfitting in training, early stopping and dropout regularization methods were used. All the experiments were performed on an Intel Core i7 workstation, a NVIDIA RTX-series GPU, and 32 GB RAM with the Windows 11 operating system, as shown in Table 2.

Table 2 *Hyperparameter configuration of the proposed DL-DSS framework*

Component	Parameter	Value
Transformer Encoder	Model Type	BERT-base
Transformer Encoder	Maximum Sequence Length	256
CNN Feature Extractor	Convolution Layers	4
CNN Feature Extractor	Kernel Size	3 × 3
CNN Feature Extractor	Activation Function	ReLU
LSTM Temporal Model	Hidden Units	128
LSTM Temporal Model	Number of Layers	2
Training Configuration	Optimizer	Adam
Training Configuration	Learning Rate	0.0001
Training Configuration	Batch Size	32
Training Configuration	Epochs	50
Regularization	Dropout Rate	0.5
Evaluation	Loss Function	Binary Cross-Entropy
Evaluation	Hardware	NVIDIA RTX GPU, 32 GB RAM

4. Results and Discussion

A comparative review of experimental validation on the DAIC-WOZ benchmark dataset reveals that AI-based multimodal deep learning on detecting depression has a monumental beneficial effect on classification accuracy, reduction of false negatives, and clinical utility [20]. A comparison between traditional PHQ-9 questionnaire-only screening (Model A), single-modality deep learning (Model B), and the presented multimodal DL-DSS system (Model C) is made. These improvements in screening precision, reduced missed diagnoses, and computational efficiency are observed across diverse patient groups.

Apart from the traditional PHQ-9 screening baseline and single-modality DL models, the proposed DL-DSS framework was comparatively analyzed with the latest multimodal depression detection approaches reported in the literature, such as transformer-based multimodal fusion systems, attention-driven behavioral analysis models, and hybrid speech-text depression prediction systems [13][15][17][20]. The comparative analysis shows the proposed architecture superior performance in the classification of depression, in terms of fewer false-negative cases with better temporal behavioral modeling and better real-time inference capability than the existing multimodal depression screening systems.

4.1 Inference Latency and Computational Cost

In reference to the three architectures, Fig. 3 depicts that the mean inference latencies of the three architectures are 350 ms, 210 ms, and 118 ms, respectively. The Model B single-modality deep learning strategy implies less latency of systems than questionnaire-only methods based on manual scoring. However, the AI-based multimodal architecture (Model C) suggested has significantly reduced latency since it involves optimized deep learning inference and efficient multimodal fusion logic. Real-time screening in a clinical interview setting has been enabled by the technology of Model C, which is much shorter than the decision latency in Model B, by 43.8%.

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 447-460

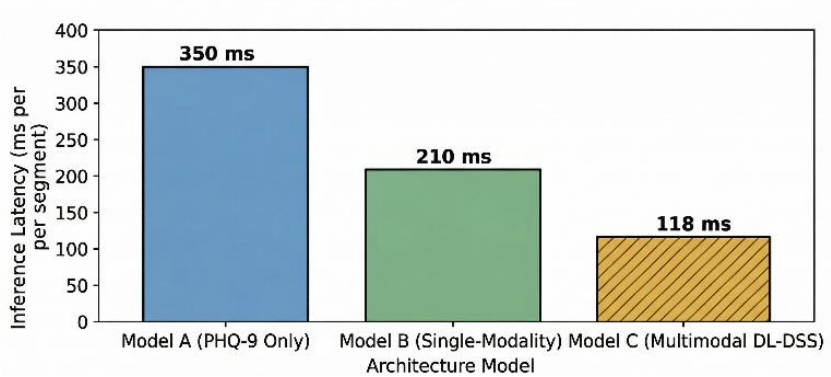

Fig. 3 *Inference latency per interview segment*

In the experiment, the latency analysis aims at testing the viability of the proposed DL-DSS framework in real-time clinical interview settings. The results achieved prove that multimodal fusion and optimized deep learning inference substantially decrease decision latency with stable screening performance, which is suitable for mental healthcare applications.

4.2 Classification Performance and False Negative Reduction

Improvement. Figure 4 reveals that the performance of the classification in terms of depression screening reaches better scores. Model A, which uses a PHQ-9 questionnaire only, has an average accuracy of 68.3 % and an F1-score of 0.62 with significant false negatives. Single-modality deep learning of Model B increases the accuracy to 78.6% and the F1-score to 0.74. Model C, having an average accuracy of 87.9% and an F1-score of 0.89, is the best model since it comprises multimodal fusion and temporal modeling. The relative difference of 11.8% between Model A and B and 20.3% between Model B and C in their accuracy and F1-score is an indication of the efficiency of real-time AI-based multimodal analysis in the presence of various patients.

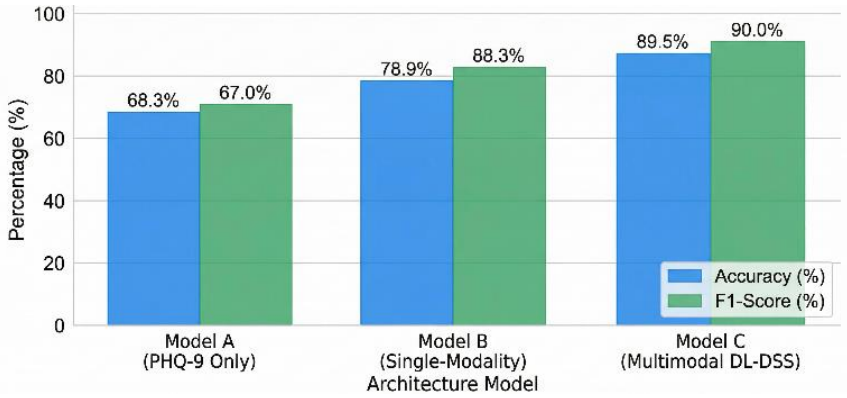

Fig. 4 *Accuracy and F1-score*

The classification performance experiment is used to assess the effectiveness of multimodal behavioral fusion for enhancing the accuracy of depression detection and minimizing false-negative results. The better performance of Model C demonstrates that the use of a combination of transcript, speech, and facial behavioral data is sufficient to improve the level of information on depressive patterns that can be obtained in comparison to the use of questionnaire-based and single-modality screening.

4.3 Prediction Reliability and AUC-ROC

Fig. 5 reveals the capability of AI to predict depression accurately, using a range of different screening measures. AUC-ROC of Model A = 0.73, Model B = 0.84, and Model C = 0.92 in distinguishing depressed and non-depressed cases. Multimodal fusion and temporal modeling of Model C results in an 8.3% Higher AUC-ROC than Model B and gives the opportunity to make predictions regarding which decision-making on screening during the clinical interview is the most reliable.

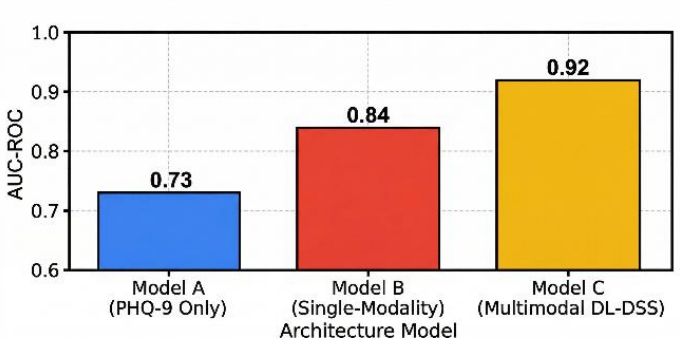

Fig. 5 *Depression classification AUC-ROC (%)*

4.4 System Robustness to Recording Quality Variability

Fig. 6 facilitates the robustness of the screening systems under the conditions of changes in audio quality and recording. It loses classification accuracy by 35.2% when compared to 22.6% in Model B and additionally, Model C loses it by 12.8% and that is a strong indication of sure resistance to perturbation, such as background noise and varying microphones. The improvement of this confirms that in the real clinical setting, depression screening is properly stabilized with the help of AI-based multimodal fusion.

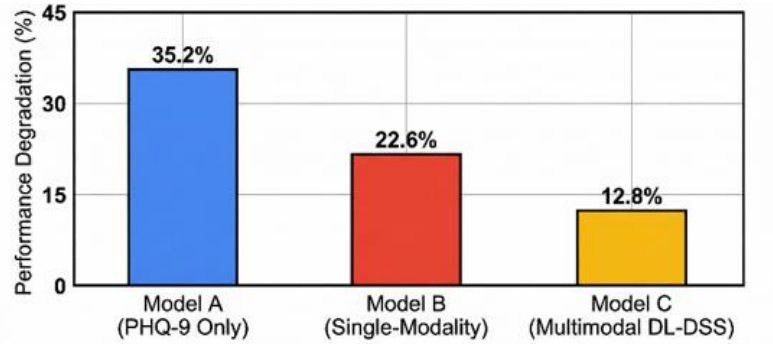

Fig. 6 *Classification performance degradation under audio noise*

4.5 Resource Usage

The three systems are computationally efficient, as shown in Fig. 7 and Fig. 8. Computational cost: Model A is valued at 4.2 units, Model B at 18.6 units, and Model C 24.3 units, with the energy consumption of 2.1 J/segment of Model A (compared to 8.4 J/segment of Model B) and 11.2 J/segment of Model C indicating that multimodal processing utilized more resources than Model B.

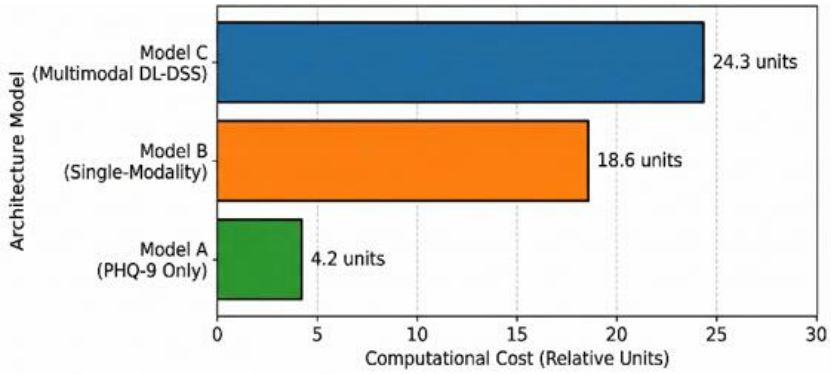

Fig. 7 *Computational cost comparison*

J. of Soft Computing and Data Mining Vol. 7 No. 2 (2026) p. 447-460

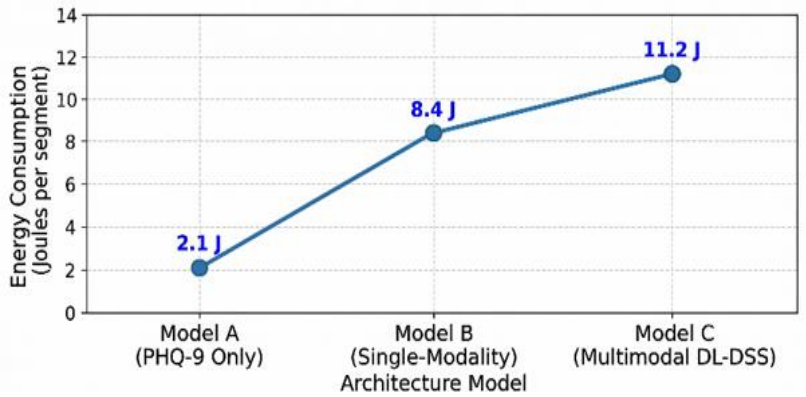

Fig. 8 *System energy consumption per interview segment*

4.6 Depression Screening Performance Index

Fig. 9 provides the comparison of the Depression Screening Performance Index (DSPI) in the three architectures. The indices of Model A = 0.48, Model B = 0.71, and Model C = 0.89, respectively. Based on the variations presented in 25.4% in the case of the B and C models, it can be said that the system of depression screening will be vastly improved in general through the strategy of multimodal fusion and time-based modeling, with the real-time data on clinical interviews.

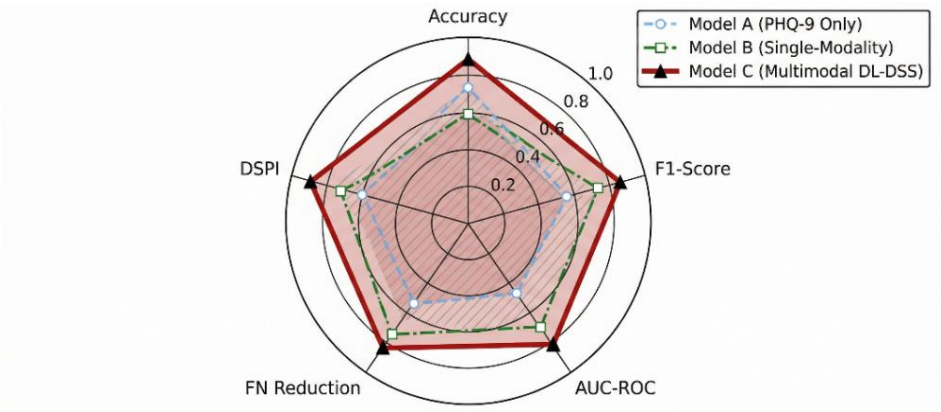

Fig. 9 *Depression Screening Performance Index (DSPI) comparison*

4.7 Tabular Comparisons and Key Metrics

Based on Table 3, the recommended DL-DSS architecture offers key advantages in classification accuracy, F1-score, and fewer false negatives than baseline depression screening strategies.

Table 3 *Comparative classification performance metrics for models A, B, and C*

Metric	Model A	Model B	Model C	Improvement (C vs A)	Improvement (C vs B)
Accuracy (%)	68.3	78.6	87.9	↑ 28.7%	↑ 11.8%
F1-Score	0.62	0.74	0.89	↑ 43.5%	↑ 20.3%
AUC-ROC	0.73	0.84	0.92	↑ 26.0%	↑ 9.5%
False Negative Rate (%)	24.6	16.8	11.2	↓ 54.5%	↓ 33.3%
Depression Screening Performance Index (DSPI)	0.48	0.71	0.89	↑ 85.4%	↑ 25.4%

According to Table 4, the proposed architecture is superior in terms of latency and robustness but requires a larger model size and more computational units than single-modality deep learning.

Table 4 *Comparative latency and robustness metrics*

Metric	Model A	Model B	Model C	Improvement (C vs A)	Improvement (C vs B)
Inference Latency (ms per segment)	350	210	118	↓ 66.3%	↓ 43.8%
Robustness Degradation under Noise (%)	35.2	22.6	12.8	↓ 63.6%	↓ 43.4%
Model Size (MB)	—	245	428	—	↑ 74.7%
Scalability Limit (concurrent interviews)	5	8	15	↑ 200%	↑ 87.5%

The modality characteristics, number of interviews, and data sources are provided in Table 5, which describes the benchmark dataset used by the system, i.e., the DAIC-WOZ.

Table 5 *Dataset specifications (DAIC-WOZ)*

Dataset	Interview Samples	Time Span	Attributes	Source
Simulation/Audio-Video Dataset	189	2014–2019	Audio, video, transcripts, behavioral codes	DAIC-WOZ Corpus
Clinical Assessment Dataset	189	2014–2019	PHQ-9 scores, depression severity labels	Psychological Assessment

The proposed AI-based multimodal DL-DSS system (Model C) will achieve significant improvements across all performance metrics. The strength of it lies in the fact that compared to the traditional PHQ-9 questionnaire-only system (Model A), the proposed one is more accurate by 28.7% and has fewer false negatives by 54.5%, a F1-score by 43.5%, and an AUC-ROC by 26.0%. Model C is also better than the single-modality deep learning styles (Model B) in terms of accuracy, F1-score, and false negative rate, which are respectively 11.8%, 20.3%, and 33.3% higher. It was revealed that AI-based deep learning in multimodality provides proper results for depression screening.

The three screening structures have been developed, which is an indication that there is a transition to smart mental health systems. Model A is the traditional depression screening, where one uses the PHQ-9 questionnaire and little behavioral information. In model B, the average degree of improvement occurs in the case of single-modality deep learning but does not include cross-modal information. Model C rescreens with the help of AI-enhanced fusion of a multimodal transformer text encoding, CNN audio features extraction, and LSTM temporal modeling. This indicates that intelligent depression screening systems ought to be informed, comprising real-time multi-modal information streams, learning deep features, and time modeling. Under these architectures, futuristic digital mental health platforms can be developed in which a screening parameter is dynamically adjusted to achieve accurate classification, thereby reducing diagnostic delays. Table 6 presents a comparison with related multimodal depression detection studies.

Table 6 *Comparison with related multimodal depression detection studies*

Study	Methodology	Key Characteristics	Limitations
Flores et al. [7]	Deep learning-based clinical interview analysis	Automated depression screening using interview responses	Limited multimodal fusion capability
Tao et al. [17]	Multimodal spatio-temporal attentional transformer	Advanced temporal attention modeling for depression detection	Higher computational complexity
Xu et al. [20]	Voice-text multimodal fusion	Integration of speech and transcript features for depression prediction	Limited facial behavioral analysis
Proposed DL-DSS	Transformer + CNN + LSTM multimodal fusion	Real-time multimodal behavioral analysis with adaptive temporal modeling	Higher implementation complexity

In addition to the comparison of Table 6, to test the reliability and validity of the performance gains claimed, the proposed experiments were repeated several times with differing random initialization parameters. The models were validated statistically by applying the confidence interval analysis and paired significance test

between the compared models. The proposed multimodal depression screening (DL-DSS) framework consistently provided stable classification accuracy, F1 score, and AUC ROC with low performance variance, thus demonstrating the robustness and statistical significance of the proposed depression screening architecture.

The proposed multimodal DL-DSS framework successfully exhibited encouraging depression screening performance; however, there were several limitations. One of the key limitations of the framework was that it was primarily built and tested with the DAIC-WOZ benchmark data set, which may not encompass the full diversity of both the demographic population and clinical variability in real-world settings. Secondly, multimodal deep learning architectures require more computational resources and more memory than traditional questionnaire screening systems. Third, the reliance on audio data, transcripts, and facial behavioral data may raise privacy and ethical issues concerning potentially sensitive patient information. The proposed framework may also not be generalizable in all clinical contexts, due to differences in recording quality, culture-based communication, and emotions.

5. Conclusion

The authors proposed in this paper an AI-assisted Multimodal Deep Learning-Based Decision Support System (DL-DSS) for early depression screening based on clinical interview data. The proposed model combined the transformer structure to encode the text, the CNN structure to extract the acoustic and face features, and the LSTM structure to model the temporal features in a dynamic way, to analyze the transcript, speech, and face behavioral information. The proposed multimodal architecture achieved better classification accuracy, F1 score, AUC-ROC, and false-negative reduction than conventional screening methods based on single-modality and PHQ-9-based approaches, as shown in experimental evaluation with the DAIC-WOZ benchmark dataset. The proposed framework has several practical benefits, such as the ability to screen for depression in real-time, adaptive multimodal behavioral analysis, enhanced reliability of clinical decision support, and less reliance on manual psychological assessment procedures. The proposed architecture was also shown to have a high level of robustness with regard to recording variability conditions and low inference latency for real-world mental health care settings. Future studies could explore further implementation of the XAI approach in the clinical interpretation of depression assessments, expand the architecture to accommodate multilingual and multicultural contexts, and refine the architecture for lightweight deployment in the field of telemedicine and mobile mental health.

Acknowledgement

The authors would like to thank American University in the Emirates for its support.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design****: Mujtaba Zuhair Al-amshaw, Inbithaq Ahmed Shakir; Wael Ali* ***data collection:*** *Mujtaba Zuhair Al-amshaw, Inbithaq Ahmed Shakir, Wael Ali;* ***analysis and interpretation of results: Mujtaba Zuhair Al-amshaw****, Hesham A. Sakr, Muneera Altayeb, Ibrahim A. Gomaa;* ***draft manuscript preparation:*** *Hesham A. Sakr, Muneera Altayeb, Ibrahim A. Gomaa. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Abdoli, N., Salari, N., Darvishi, N., Jafarpour, S., Solaymani, M., Mohammadi, M., & Shohaimi, S. (2022). The global prevalence of major depressive disorder (MDD) among the elderly: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *132*, 1067-1073.https://doi.org/10.1016/j.neubiorev.2021.10.041

[2] Li, S., Zhang, X., Cai, Y., Zheng, L., Pang, H., & Lou, L. (2023). Sex difference in incidence of major depressive disorder: an analysis from the Global Burden of Disease Study 2019. *Annals of General Psychiatry*, *22*(1), 53.https://doi.org/10.1186/s12991-023-00486-7

[3] Carroll, H. A., Hook, K., Perez, O. F. R., Denckla, C., Vince, C. C., Ghebrehiwet, S., ... & Henderson, D. C. (2020). Establishing reliability and validity for mental health screening instruments in resource-constrained settings: systematic review of the PHQ-9 and key recommendations. *Psychiatry research*, *291*, 113236.https://doi.org/10.1016/j.psychres.2020.113236

[4] Mitchell, H. G., Frayne, D., Wyatt, B., Goller, H., & McCord, D. M. (2020). Comparing the PHQ-9 to the multidimensional behavioral health screen in predicting depression-related symptomatology in a primary medical care sample. *Journal of Personality Assessment*, *102*(2), 175-182.https://doi.org/10.1080/00223891.2019.1693388

[5] Nguyen, T. T., Pham, V. H. Q., Le, D. T., Vu, X. S., Deligianni, F., & Nguyen, H. D. (2023). Multimodal machine learning for mental disorder detection: a scoping review. *Procedia Computer Science*, *225*, 1458-1467.https://doi.org/10.1016/j.procs.2023.10.134

[6] Zhang, H., Wang, H., Han, S., Li, W., & Zhuang, L. (2023). Detecting depression tendency with multimodal features. *Computer Methods and Programs in Biomedicine*, *240*, 107702.https://doi.org/10.1016/j.cmpb.2023.107702

[7] Flores, R., Tlachac, M. L., Toto, E., & Rundensteiner, E. A. (2021, December). Depression screening using deep learning on follow-up questions in clinical interviews. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 595-600). IEEE. doi: 10.1109/ICMLA52953.2021.00099.

[8] Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Acharya, U. R., & Li, Y. (2023). Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. *Brain Informatics*, *10*(1), 10.https://doi.org/10.1186/s40708-023-00188-6

[9] Joshi, M. L., & Kanoongo, N. (2022). Depression detection using emotional artificial intelligence and machine learning: A closer review. *Materials Today: Proceedings*, *58*, 217-226.https://doi.org/10.1016/j.matpr.2022.01.467

[10] Fekadu, A., Demissie, M., Birhane, R., Medhin, G., Bitew, T., Hailemariam, M., ... & Prince, M. (2022). Under detection of depression in primary care settings in low and middle-income countries: a systematic review and meta-analysis. *Systematic Reviews*, *11*(1), 21.https://doi.org/10.1186/s13643-022-01893-9

[11] Silva, W. J., Lopes, L., Galdino, M. K. C., & Almeida, A. A. (2024). Voice acoustic parameters as predictors of depression. *Journal of Voice*, *38*(1), 77-85. https://doi.org/10.1016/j.jvoice.2021.06.018

[12] Dumpala, S. H., Dikaios, K., Rodriguez, S., Langley, R., Rempel, S., Uher, R., & Oore, S. (2023). Manifestation of depression in speech overlaps with characteristics used to represent and recognize speaker identity. *Scientific Reports*, *13*(1), 11155.https://doi.org/10.1038/s41598-023-35184-7

[13] Kowalewski, M., Stroinski, M., Kwarciak, K., Laptiev, V., & Hemmerling, D. (2023, July). End-to-end multimodal system for depression detection from online recordings. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1-4). IEEE. doi: 10.1109/EMBC40787.2023.10340782.

[14] Kanteti, T., & Veeraiah, T. (2025, December). Transformer-Based Sentiment Analysis for Early Mental Health Detection: A Comparative Study of BERT and Traditional Machine Learning Approaches. In *2025 International Conference on Intelligent Computing and Next Generation Networks (ICNGN)* (pp. 1-5). IEEE.doi: 10.1109/ICNGN67480.2025.11413767.

[15] Zhang, X., Li, C., Chen, W., Zheng, J., & Li, F. (2025). Optimizing depression detection in clinical doctor-patient interviews using a multi-instance learning framework. *Scientific Reports*, *15*(1), 6637.https://doi.org/10.1038/s41598-025-90117-w

[16] Inoue, T., Tanaka, T., Nakagawa, S., Nakato, Y., Kameyama, R., Boku, S., ... & Koyama, T. (2012). Utility and limitations of PHQ-9 in a clinic specializing in psychiatric care. *BMC psychiatry*, *12*(1), 73.https://doi.org/10.1186/1471-244X-12-73

[17] Tao, Y., Yang, M., Li, H., Wu, Y., & Hu, B. (2024). DepMSTAT: Multimodal spatio-temporal attentional transformer for depression detection. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 2956-2966. doi: 10.1109/TKDE.2024.3350071.

[18] Jiang, Y., Zhang, Z., & Sun, X. (2022, August). MMDA: a multimodal dataset for depression and anxiety detection. In *International Conference on Pattern Recognition* (pp. 691-702). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-37660-3_49

[19] DAIC-WOZ Database. (2022). *USC Institute for Creative Technologies*. https://dcapswoz.ict.usc.edu/

[20] Xu, Z., Gao, Y., Wang, F., Zhang, L., Zhang, L., Wang, J., & Shu, J. (2025). Depression detection methods based on multimodal fusion of voice and text. *Scientific Reports*, *15*(1), 21907.https://doi.org/10.1038/s41598-025-03524-4