Original Paper

# Comparative Efficacy of MultiModal AI Methods in Screening for Major Depressive Disorder: Machine Learning Model Development Predictive Pilot Study

Donghao Chen[1], PhD; Pengfei Wang[2,3], PhD; Xiaolong Zhang[2,3], PhD; Runqi Qiao[1], PhD; Nanxi Li[2,3], PhD; Xiaodong Zhang[2,3], BD; Honggang Zhang[1], PhD; Gang Wang[2,3], PhD,MD

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

[2]Beijing Key Laboratory of Mental Disorders, National Clinical Research Center for Mental Disorders & National Center for Mental Disorders, Beijing Anding Hospital, Capital Medical University, Beijing, China

[3]Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing, China

**Corresponding Author:**

Gang Wang, PhD,MD
Beijing Key Laboratory of Mental Disorders, National Clinical Research Center for Mental Disorders & National Center for Mental Disorders
Beijing Anding Hospital, Capital Medical University
No 5 Ankang Lane, Xicheng District
Beijing, 100088
China
Phone: 86 15210807053
Email: gangwangdoc@ccmu.edu.cn

## Abstract

**Background:** Conventional approaches for major depressive disorder (MDD) screening rely on two effective but subjective paradigms: self-rated scales and clinical interviews. Artificial intelligence (AI) can potentially contribute to psychiatry, especially through the use of objective data such as objective audiovisual signals.

**Objective:** This study aimed to evaluate the efficacy of different paradigms using AI analysis on audiovisual signals.

**Methods:** We recruited 89 participants (mean age, 37.1 years; male: 30/89, 33.7%; female: 59/89, 66.3%), including 41 patients with MDD and 48 asymptomatic participants. We developed AI models using facial movement, acoustic, and text features extracted from videos obtained via a tool, incorporating four paradigms: conventional scale (CS), question and answering (Q&A), mental imagery description (MID), and video watching (VW). Ablation experiments and 5-fold cross-validation were performed using two AI methods to ascertain the efficacy of paradigm combinations. Attention scores from the deep learning model were calculated and compared with correlation results to assess comprehensibility.

**Results:** In video clip-based analyses, Q&A outperformed MID with a mean binary sensitivity of 79.06% (95%CI 77.06%-83.35%; *P*=.03) and an effect size of 1.0. Among individuals, the combination of Q&A and MID outperformed MID alone with a mean extent accuracy of 80.00% (95%CI 65.88%-88.24%; *P*= .01), with an effect size 0.61. The mean binary accuracy exceeded 76.25% for video clip predictions and 74.12% for individual-level predictions across the two AI methods, with top individual binary accuracy of 94.12%. The features exhibiting high attention scores demonstrated a significant overlap with those that were statistically correlated, including 18 features (all *Ps*<.05), while also aligning with established nonverbal markers.

**Conclusions:** The Q&A paradigm demonstrated higher efficacy than MID, both individually and in combination. Using AI to analyze audiovisual signals across multiple paradigms has the potential to be an effective tool for MDD screening.

***JMIR Form Res 2025;9:e56057;*** doi: 10.2196/56057

**Keywords:** major depressive disorder; artificial intelligence; computational psychiatry; facial action unit; multimodal analysis; multiparadigm analysis; MDD

## Introduction

Depressive disorder is a common mental disorder affecting approximately 322 million patients in the world, with major depressive disorder (MDD) as one of its two main subcategories, which can significantly affect all aspects of life, including performance at school, productivity at work, and relationships with family and friends [1]. The primary methods to assess depression encompass mental status examinations and assessment scales. However, mental status examinations such as the Hamilton Rating Scale for Depression (HAMD) necessitate direct, in-person interviews conducted by clinicians, which can result in processes that are both time-consuming and labor-intensive [2]. Self-Report Symptom Inventories (SRSI) such as the Beck Depression Inventory [3] and the Patient Health Questionnaire-9 (PHQ-9) [4] are time-efficient but can be influenced by subjective biases, which allows for individual variability [5]. Therefore, the outcomes are susceptible to both intentional and unintentional subjective influences [6] and more approaches are needed to improve efficiency and accuracy.

In recent years, artificial intelligence (AI) has garnered attention for its application in signal analysis across various modalities. For instance, support vector machines have been used to analyze functional magnetic resonance imaging (fMRI) data [7] and a convolutional neural network (CNN) has been applied to an electroencephalogram (EEG) [8] to detect depression. While physiological signals such as fMRI and EEG are unaffected by subjective factors and directly reflect the participants' physical states, they involve complex procedures and high costs. In contrast, noncontact signals, including text, audio, visual content, and scale information are more accessible for analysis.

In the text modality, hidden Markov models and random forest models were developed to predict depression and posttraumatic stress disorder based on frequency of Twitter usage and content [9]. By aggregating weighted words using lexicons, the sentiment score derived from text messages demonstrated a positive association with the severity of depression as measured by the self-rated Patient Health Questionnaire-8 (PHQ-8) [10-12]. Among the audio modalities, speech patterns such as a narrowed pitch range and reduced phonemes within the vowel space have emerged as important objective indicators for assessing depressive states [13,14]. Along with prosodic features, mel frequency cepstral coefficients (MFCC) [15], detailed spectral features [16], and deep-learned acoustic characteristics [17] have also been used to identify the presence of depressive symptoms, achieving binary accuracy of up to 79% or $F_1$-score of 0.890.

For the visual or multimodal domain, several open datasets are available. One notable example is the Audio/Visual Emotion Challenge and Workshop [18], which focuses on the detection of depression and uses an audio-visual dataset that includes image features extracted from original images and audio recordings and transcribed text from Google Cloud, paired with the PHQ-8 scores. Facial action units (AU), as outlined in the Facial Action Coding System [19], serve as the foundation for facial expressions and constitute essential image features in the Audio/Visual Emotion Challenge and Workshop. Commonly observed AUs correspond to a range of expressions such as smiling and frowning (see Table S1 in Multimedia Appendix 1). A higher overall frowning (Action Unit 4, ie AU4) and head-down posture were identified in a study by Fiquer et al [20], while a lower overall AU12 and a markedly higher overall AU14 were identified in a study by Girard et al [21]. This indicated that the distribution of AUs differs significantly between depressed and nondepressed persons. Facial Action Coding System has also been employed in the analysis of stress [22], anxiety [23], and Parkinson's disease [24].

There are several other existing datasets, including Mundt-35 [25], BlackDog [26], and MODMA [27]. Most of these datasets contain a single paradigm, primarily relying on interviews such as HAMD, or targeting the scores of SRSIs such as the PHQ-8. Additionally, current multimodal AI methods mainly extract local features from utterances or sentences for video clip predictions [28,29]. At the same time, we believe the screening and diagnosis of MDD should include the entire process, similarly to the process of clinical practice.

For other paradigm options, mental imagery description (MID) [30] can manifest across different sensory modalities, encompassing visual [31,32], auditory [33], and textual information, and tends to evoke stronger emotional responses than verbal processing [30].

Thus, aiming to evaluate the efficacy of different paradigms, we aggregated them in a tool, namely the Electronic Tool for Depression (ETD), and used a state-of-the-art (SOTA) method using audiovisual signals to validate their efficacies. We propose the ETD to be a nonsubjective and easy-use MDD screening tool. The SOTA method generates predictions on video clips, and two of the four paradigms contain only visual signals; therefore, we implemented a voting mechanism for individual predictions and proposed a global feature method for the remaining vision-only paradigms. This pilot study underscores our primary contributions, which can be summarized as follows: (1) to validate the efficacy of the paradigms via AI on audiovisual signals and aggregate them within a tool for MDD screening and (2) to propose a global feature method and explore its efficacy and interpretability.

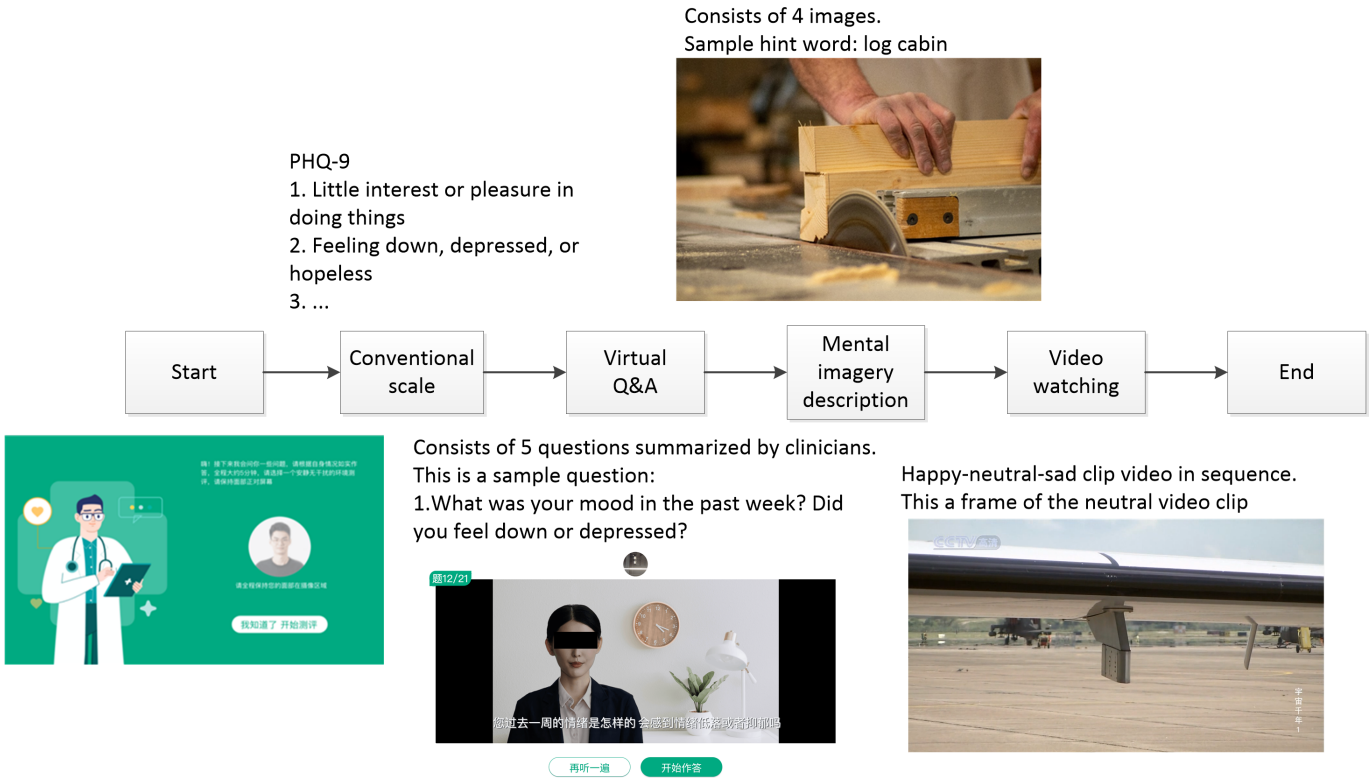
## Methods

### Design of the Task and Building the Tool

The ETD consists of four paradigms, aggregated into an application designed for an 11.5-inch tablet featuring an 8-MP front-facing camera and a 44.8 kHz sample rate microphone. Before using the ETD, clinicians adjusted the tablet to ensure that the participant's head is aligned with the device at an appropriate distance (approximately 50 centimeters) for effective face capture. The ETD structure and app design are depicted in Figure 1. Paradigm 1 uses a conventional self-rated scale, specifically the PHQ-9.

Paradigm 2 encompasses a question-and-answering (Q&A) paradigm simulating psychiatric examinations. Paradigm 3 requires participants to describe images with the hint words [30-33]. Paradigm 4 presents three video clips of varying emotional sentiment scores [34,35] in a positive, neutral, and negative sequence (41-73 seconds, average 60.33 seconds). Participants sequentially respond to these components, with recordings capturing their reactions during both the viewing and responding phases, including the PHQ-9 selections; the entire process takes approximately 5 minutes. It is essential to clarify that the scale was only used to elicit reactions, and its scores did not contribute to the predictions.

**Figure 1.** Components of the Electronic Tool for Depression (ETD). PHQ-9: Patient Health Questionnaire-9; Q&A: question-and-answer.




## Recruitment

In this study, 89 participants were recruited from April 2022 to December 2022 in Beijing. Among these, 51 were recruited from the Beijing Anding Hospital Inpatient Department and were all diagnosed with MDD by experienced psychiatrists according to the International Code of Diseases, tenth revision (*ICD-10*) [36]. All participants met the inclusion criteria, which were as follows: (1) age 18-65 years, (2) proficiency in standard Chinese, (3) educational level of primary school or above, and (4) ability to understand and cooperate with the research protocol.

Exclusion criteria included (1) diagnosis of schizophrenia, schizoaffective disorder, or other mental disorders and (2) history of organic brain disease. The remaining 38 participants were recruited openly from the general population (employees and college students) who were not experiencing depression-related symptoms.

Participants from the hospital completed two steps: the first involved using the ETD app, and the second included assessment using the HAMD-17 scale by clinicians. Community participants only completed the ETD test, and all were confirmed to have no depressive symptoms based on the PHQ-9 assessment. Finally, the asymptomatic and the healthy control groups formed the nonMDD group (48 participants), while the mild group and the moderate or severe group were collectively referred to as the MDD group (41 participants). For ease of explanation, the mild group was designated as MDD-sub1, and the moderate or severe group was designated as MDD-sub2. Sex was compared using the $X^2$ test; age and HAMD scores were compared using the Mann-Whitney $U$ test. There were no significant differences in sex ratio or age between the groups, while the MDD group had significantly higher PHQ-9 scores than the nonMDD group.

## Ethical Considerations

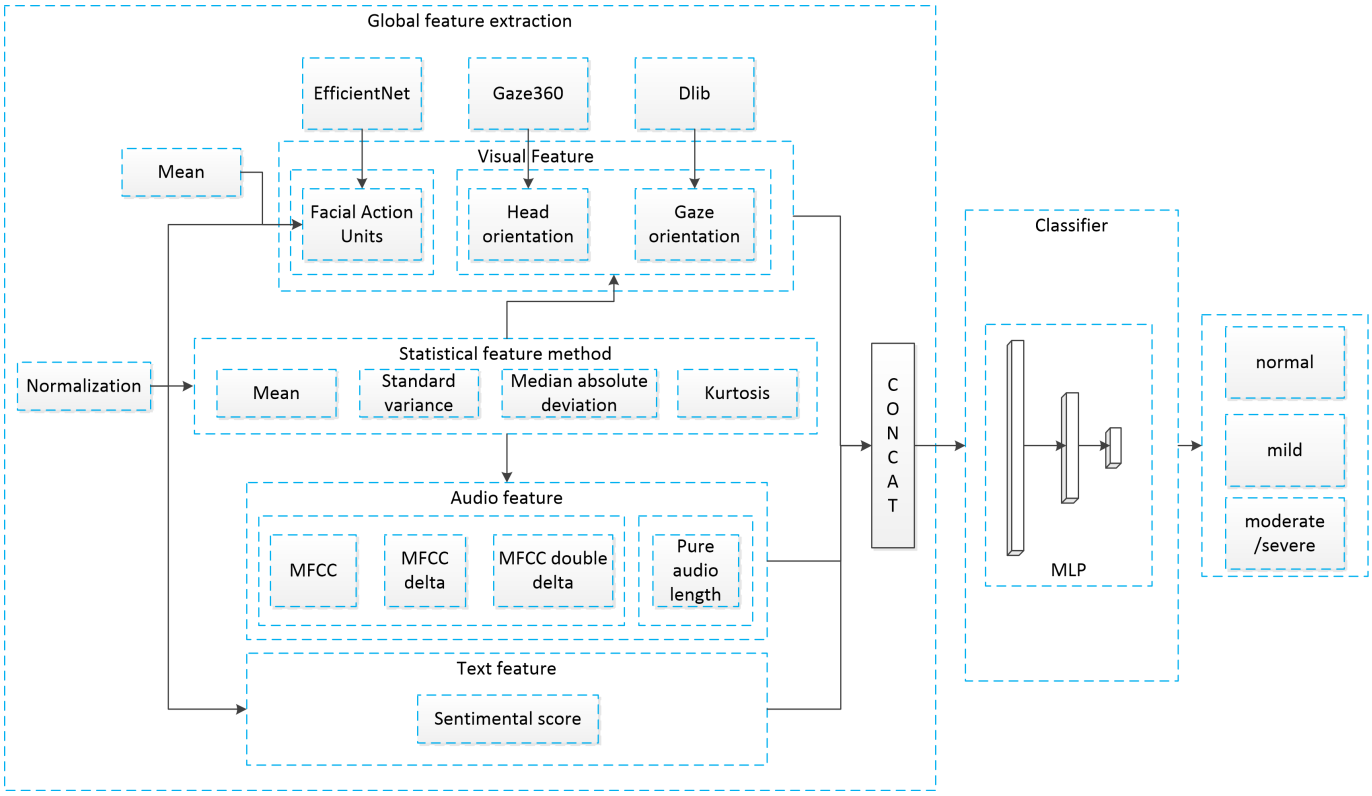
This study was approved by the Ethics Committee of Beijing Anding Hospital Capital Medical University. No compensation fee was paid to participants, with written informed consent obtained for the data usage of research analysis. Data were deidentified and all analyses followed data privacy guidelines.

## Model Training

All recorded videos underwent a manual verification process to ensure that the image ratio of a complete head, face, and eyes exceeded the empirical 95% threshold. We adopted the MFCC-based recurrent neural network (RNN) [29] as the validation model, which used a multimodal method that integrated MFCC and AU features and achieved a SOTA accuracy of 95.6% in binary classification of depression

on the DAIC-WOZ (Distress Analysis Interview Corpus) dataset [18]. We pretrained the RNN on RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song dataset) [37], aggregated the AU features, and fine-tuned the model on our dataset. We developed a CNN model for AU detection using EfficientNet [38] on BP4D [39], achieving a mean $F_1$-score of 0.76 on selected AUs. The sample size of video clips for the RNN was 11,075, comprising 6826 normal, 2933 mild, and 1316 moderate or severe instances. We established a clip-voting ratio to represent the individual results. While the RNN simultaneously processed local audio and visual data, it did not incorporate conventional scale and video watching. To address this limitation, we proposed a global feature extraction method (depicted in Figure 2) to derive global features and build AI models. For the vision modality, we used Gaze360 [40] and Dlib [41] along with AU features to estimate gaze and head orientation. For the audio modality, we extracted MFCC-based features and the pure audio duration of the human voice. For the text modality, we calculated sentiment scores using the *pyltp* package [42]. The features were concatenated in the order of visual, audio, and text features. Additionally, we incorporated statistical characteristics such as mean and variance to enhance their global representation. Normalization and bias adjustments were applied to ensure that all the features were positive for later attention computation (complete feature list is provided in Table S2 in Multimedia Appendix 1).

**Figure 2.** The global feature extraction method architecture. MFCC: mel frequency cepstral coefficients; MLP: multilayer perceptron.




We adopted a multilayer perceptron (MLP) as our classifier for global predictions, which is identical to the RNN. The MLP comprises layers with 512, 1024, 128, and 3 neurons, incorporating batch normalization and a 0.2 dropout rate to mitigate overfitting; a softmax layer was added as the prediction. The Adam optimizer was used with a base learning rate of 1e-3, $\beta_1$ of 0.9, $\beta_2$ of 0.999, and $\varepsilon$ of 1e-8. Given that deep learning methods are often considered "dark magic," we sought to enhance comprehensibility by employing Grad-Cam [43] to visualize the attention scores of the MLP's best-performing model across each feature. These results were then compared with Spearman and Kendall correlation coefficients computed using scikit-learn [44].

### Statistical Analysis

Ablation experiments were conducted on various paradigm combinations. The models predicted three levels of severity, and binary performance was assessed to distinguish between depressed or nondepressed states. Sensitivity, specificity, accuracy, and area under the curve (AUC) were measured for binary results. Accuracy was specifically calculated for severity predictions. The five-fold performances underwent the Friedman test, followed by the posthoc Nemenyi test and Cliff $\delta$ effect size. A 95% CI was computed using bootstrapping, with the exception of single fold clip prediction AUC, which used normal approximation.

## Results

Demographic characteristics are shown in Table 1 and the findings of clip prediction and the Friedman test are presented in Table 2. The difference among Q&A, MID, and QI (combination of Q&A and MID) is significant in binary sensitivity ($P$=.02), with a large effect ($\varepsilon^2$=0.47). The

differences in binary AUC and extent accuracy are close to significant ($P$=.07 and $P$=.09, respectively), with large effects ($\varepsilon^2$=0.26 and $\varepsilon^2$=0.23, respectively). The differences in binary specificity and binary accuracy were not significant and exhibited small effects. Posthoc Nemenyi test results for sensitivity are detailed in Table 3, revealing that Q&A outperformed MID ($P$= .03) with a large effect size (Cliff $\delta$=1.0). The difference between QI and MID is close to significant ($P$=.06) with a large effect (Cliff $\delta$=1.0).

The results of individual prediction and the Friedman test are presented in Table 4. In the RNN voting analysis, the differences among Q&A, MID, and QI were significant in terms of binary sensitivity ($P$<.01) with a large effect ($\varepsilon^2$=0.61). The differences in binary accuracy and binary AUC were nonsignificant ($P$ =.13 and $P$=.09, respectively) but showed large effects ($\varepsilon^2$=0.18 and $\varepsilon^2$=0.23, respectively). Posthoc Nemenyi test results on extent accuracy are presented in Table 4.

**Table 1.** Demographic characteristics.

| Factors | MDD[a] | nonMDD[b] | $P$ value |
|---|---|---|---|
| Sex, n (%) | | | .13[c] |
| Male | 12 (29.3) | 18 (37.5) | |
| Female | 29 (70.3) | 30 (62.5) | |
| Age (years), mean (SD) | 38.41 (15.12) | 35.98(12.37) | .50[d] |
| HAM-D[e], mean (SD) | 14.51 (4.66) | –[f] | – |
| PHQ-9[g], mean (SD) | 13.05 (6.00) | 4.17 (2.71) | <.001[d] |

[a]MDD: major depressive disorder.
[b]nonMDD: non-major depressive disorder.
[c]Chi-square test was used to derive the $P$ value.
[d]Mann-Whitney U test was used to derive the $P$ value.
[e]HAM-D: Hamilton rating scale for Depression.
[f]Not applicable.
[g]PHQ-9: PHQ-9: Patient Health Questionnaire-9.

**Table 2.** Clip prediction results of the MFCC-based[a] RNN[b] [29] for paradigm combinations.

| Paradigm/ statistics/ performance | Sensitivity (%), mean (95% CI) | Specificity (%), mean (95% CI) | SA[c] (%), mean (95% CI) | AUC[d] (%), mean (95% CI) | EA[e] (%), mean (95% CI) |
|---|---|---|---|---|---|
| Q&A[f] | 79.06 (77.06-83.35) | 85.71 (73.30-90.19) | 83.01 (74.43-86.10) | 78.12 (66.11-82.35) | 88.70 (83.15-91.33) |
| MID[g] | 56.99 (41.78-63.36) | 85.10 (76.36-89.81) | 76.25 (69.90-80.30) | 70.40 (65.09-73.60) | 81.43 (76.42-85.35) |
| QI[h] | 80.22 (75.88-84.72) | 85.61 (77.06-89.50) | 84.41 (81.98-86.44) | 80.36 (76.45-82.78) | 90.37 (88.81-91.64) |
| $P$ value[i] | .02 | .55 | .25 | .07 | .09 |
| Effect size ($\varepsilon^2$) | 0.47 | 0.00 | 0.07 | 0.27 | 0.23 |

[a]MFCC: mel frequency cepstral coefficients.
[b]RNN: recurrent neural network.
[c]SA: screen accuracy.
[d]AUC: area under the curve.
[e]EA: extent accuracy.
[f]Q&A: question-and-answer.
[g]MID: mental imagery description.
[h]QI: combination of Q&A and MID.
[i]Friedman test was used to calculate the $P$ value.

**Table 3.** Posthoc test results of the MFCC-based[a] RNN[b] clip prediction sensitivity between pairs of Q&A[c], MID[d], and QI[e].

| Paradigm statistic item | $P$ value[f] | Cliff $\delta$ (effect size) |
|---|---|---|
| Q&A-MID[g] | .03 | 1.0 (large) |
| Q&A-QI[h] | .90 | –0.04 (negligible) |
| QI-MID[i] | .06 | 1.0 (large) |

[a]MFCC: mel frequency cepstral coefficients.
[b]RNN: recurrent neural network.
[c]Q&A: question-and-answer.

| Paradigm statistic item | *P* value[f] | Cliff δ (effect size) |
|---|---|---|

[d]MID: mental imagery description.
[e]QI: combination of Q&A and MID.
[f]Nemenyi test.
[g]Q&A-MID: comparison between Q&A and mental imagery description.
[h]Q&A-QI: comparison between Q&A and combination of Q&A and mental imagery description.
[i]QI-MID: combination of Q&A and mental imagery description, and single paradigm mental imagery description.

**Table 4.** Individual prediction results of the MFCC-based[a] RNN[b] voting for paradigm combinations.

| Paradigm performance | Sensitivity (%), mean (95% CI) | Specificity (%), mean (95% CI) | SA[c] (%), mean (95% CI) | AUC[d] (%), mean (95% CI) | EA[e] (%), mean (95% CI) |
|---|---|---|---|---|---|
| Q&A[f] | 75.00 (57.50-87.50) | 82.22 (68.89-86.67) | 78.82 (72.94-83.53) | 90.83 (84.17-95.83) | 70.59 (57.65-77.65) |
| MID[g] | 60.00 (42.50-67.50) | 88.89 (77.78-95.56) | 75.29 (62.35-80.00) | 85.00 (81.94-90.45) | 65.88 (57.65-71.77) |
| SQI[h] | 75.00 (50.00-87.50) | 88.89 (80.00-93.33) | 82.36 (69.42-89.42) | 92.50 (86.11-96.11) | 80.00 (65.88-88.24) |
| *P* value | .29 | .26 | .13 | .09 | .009 |
| Effect size | 0.04 | 0.06 | 0.18 | 0.23 | 0.61 |

[a]MFCC: mel frequency cepstral coefficients.
[b]RNN: recurrent neural network.
[c]SA: screen accuracy.
[d]AUC: area under the curve.
[e]EA: extent accuracy.
[f]Q&A: question-and-answer.
[g]MID: mental imagery description.
[h]SQI: combination of conventional questionnaire, Q&A, and mental imagery description.

Table 5. QI outperformed MID ($P<.05$) with a substantial effect (Cliff δ=0.64). The difference between Q&A and QI was nonsignificant ($P=.14$) but indicated a large effect (Cliff δ=–0.48). In the global feature MLP analysis, differences among the paradigms were insignificant and exhibited small effect sizes, with results in Table S3 in Multimedia Appendix 1.

The best fold performance is shown in Table 6. The global feature, SQIV (combination paradigm of CS, Q&A, MID, and VW) MLP achieved a peak individual binary accuracy of 94.12%. Notably, the RNN voting SQI model also achieved a top accuracy of 94.12%, but with a higher extent accuracy of 94.12%, and an AUC of 0.99.

**Table 5.** Post-hoc statistic test results of the RNN[a] voting individual prediction extent accuracy between pairs of Q&A,[b] MID[c], and QI[d].

| Paradigm statistic item | *P* value[e] | Effect size (Cliff δ) |
|---|---|---|
| Q&A-MID[f] | .60 | 0.32 (small) |
| Q&A-QI[g] | .14 | –0.48 (large) |
| QI-MID[h] | .01 | 0.64 (large) |

[a]RNN: recurrent neural network.
[b]Q&A: question-and-answer.
[c]MID: mental imagery description.
[d]QI: combination of Q&A and MID.
[e]Nemenyi test.
[f]Q&A-MID: comparison between Q&A and mental imagery description.
[g]Q&A-QI: comparison between Q&A and combination of Q&A and mental imagery description.
[h]QI-MID: combination of Q&A and mental imagery description, and single paradigm mental imagery description.

**Table 6.** Performance of the best fold of the global feature MLP[a] SQIV[b] model, the MFCC-based[c] RNN[d] SQI[e] clip model, and the MFCC-based RNN SQI[e] voting model.

| Method | Paradigm performance | Sensitivity %, (95% CI) | Specificity %, (95% CI) | SA[f] %, (95% CI) | AUC[g] (95% CI) | EA[h] %(95% CI) |
|---|---|---|---|---|---|---|
| MLP[a] | SQIV[b] | 100.0 (63.06-100.0) | 88.89 (51.75-99.72) | 94.12 (71.31-99.85) | 0.97 (0.87-1.0) | 76.47 (50.10-93.19) |
| RNN[d] | SQI[e] | 80.69 (77.74-83.64) | 89.93 (88.60-91.26) | 87.36 (86.09-88.63) | 0.91 (0.90-0.92) | 83.03 (81.60-84.46) |
| RNN[d] voting | SQI[e] | 100.0 (63.06-100.0) | 88.89 (51.75-99.72) | 94.12 (71.31-99.85) | 0.99 (0.91-1.0) | 94.12 (71.31-99.85) |

| Method | Paradigm performance | Sensitivity %, (95% CI) | Specificity %, (95% CI) | SA[f] %, (95% CI) | AUC[g] (95% CI) | EA[h] %(95% CI) |
|---|---|---|---|---|---|---|

[a]MLP: multilayer perceptron.
[b]SQIV: combination of conventional scale, Q&A, mental imagery description, and video watching.
[c]MFCC: mel frequency cepstral coefficients.
[d]RNN: recurrent neural network.
[e]SQI: combination of conventional scale, Q&A, and mental imagery description
[f]SA: screen accuracy
[g]AUC: area under the curve
[h]EA: extent accuracy

The test results of comparison between the RNN voting and the proposed global feature method can be found in Table S4 in Multimedia Appendix 1. The input data are the same in Q&A and MID; for paradigm combinations, we compared RNN-voting QI and the global feature method (ie, combination paradigm of CS, Q&A, and MID; SQI), as they use the most comparable input data. No statistically significant differences were identified between the two methods across all three paradigms. Figures S1-S4 in Multimedia Appendix 1 illustrate the collective and individual learnings of the best MLP SQIV model. The mean attention scores of the features are sequenced by the nonMDD group, the MDD-sub1 group, and the MDD-sub2 group. The complete attention scores with Spearman correlation scores are mentioned in Table S5 in Multimedia Appendix 1, sorted in descending order for the MDD-sub2 group. Previously analyzed nonverbal markers in studies by Fiquer et al [20] and Girard et al [45] can be found in Table S6 in Multimedia Appendix 1. As shown in Figure S1 in Multimedia Appendix 1, the different groups exhibit varying levels of attention to specific features. The Spearman and Kendall correlation coefficients for each feature relative to the target extent of depression are available in Table S5 in Multimedia Appendix 1, where 18 features demonstrated a $P$ value<.05.

# Discussion

## Principal Findings

We aimed to evaluate the efficacy of different paradigms via AI on audiovisual signals. We aggregated the four paradigms within the ETD and held 5-fold cross-validation on the two AI models among the paradigm combinations. Our findings show that there are differences in paradigm efficacies, and the AI model learns knowledge consistent with prior human experience.

For the single paradigm with the MFCC-based RNN, Q&A outperformed MID in identifying patients but performed equally in distinguishing extent levels. The difference between Q&A and MID in clip sensitivity was significant, but nonsignificant in individual extent accuracy. This makes Q&A more precise in identifying MDD patients.

For paradigm combinations with the MFCC-based RNN, integrating MID with Q&A slightly decreased clip sensitivity significance but significantly improved individual extent accuracy significance compared with MID. Considering that the difference between QI and Q&A was nonsignificant in either clip sensitivity and extent accuracy, and the differences among Q&A, MID, and QI were nonsignificant in the other performance indexes, we conclude that Q&A demonstrated higher efficacy than MID and suggest that paradigm combinations perform better than a single paradigm. As known, Q&A is a simplified version of a clinical interview, and the questions are all symptom-related, which makes it the most relevant paradigm for MDD and the most important one.

In the individual prediction of the global feature MLP, no significant differences were observed across the paradigm combinations. When fixing the paradigms, no notable differences were found between the MFCC-based RNN voting and the global feature MLP, which validated the global features' effectiveness. Some large effect sizes were noted, particularly in binary AUC and extent accuracy for Q&A and QI, which may be attributable to feature granularity. The RNN feature integrates both local and global information, with local features benefiting from transfer learning, which enhances performance—achieving a top binary accuracy of 94.12%, a top mean binary accuracy of 82.36%, and a mean extent accuracy of 80.00%. In contrast, the global features may be coarse at the granular level. Even so, the MLP still achieved a mean binary accuracy ranging from 74.12% to 85.88%, with a 95% CI spanning 69.41% to 91.77%, and a top binary accuracy of 94.12%. Overfitting might exist and could result in wide CIs.

Compared with support vector machine models [7], which showed a mean binary accuracy of 78.95% on event-related fMRI [46] and 85.00% on block-related fMRI [47] and CNN [8], which achieved a mean binary accuracy of 85.62% on EEG data, the ETD demonstrated equivalent performance while relying on much more readily accessible daily data. Compared to SRSIs, the audiovisual data are more objective and easier to use. Compared to interview-based assessments such as the HAMD [2], the ETD required approximately 5 minutes, saving about 83% of the time. The ETD's performance and efficiency support its potential to objectively, accurately, and efficiently screen for MDD.

In the visualization, it is noteworthy that the high- and low-attention features did not intersect, particularly between the MDD group and the nonMDD groups, indicating that participants in different groups exhibited diverse behavioral patterns. Almost all 16 features mentioned by Fiquer et al [20] and Girard et al [21] exhibited high attention scores, with the lowest score being 0.68. Among these, 15 features ranked

in the top 20%, except for "head down" which ranked 31st in MDD-sub1, demonstrating consistency with prior studies. Additionally, "head motion velocity" was not included in this study. When comparing with correlation results, 18 features emerged as having significant positive or negative relationships with MDD extent, both in feature items and in correlation trends observed via Spearman and Kendall methods—differing only in specific weights. Of these, 11 aligned with the correlation trends; 5 showed patterns in which the attention scores for the MDD group were either higher or lower than those of the nonMDD group, and only 2 showed no clear trends. For instance, the mean attention score of AU4, interpreted as "frown" by Fiquer et al [20], increased with the extent of MDD, and the Spearman correlation for AU4 was positive ($P$=.005). These high-attention score and correlation-consistent features may serve as urgently needed objective markers and should be further investigated.

The MLP leverages global features representing statistical values throughout the process, sacrificing some detail at the granular level while maintaining low model and computational complexity. Despite potential overfitting, the alignment of the visualization results with correlation findings indicates that the MLP has acquired knowledge consistent with medical prior knowledge, supporting its performance and underscores its potential as a valuable tool. For the inconsistent elements, the neural network introduces significant nonlinearity and captures relationships in high-dimensional spaces. In contrast, Spearman and Kendall correlations are limited to assessing relationships between single inputs and targets. We propose that a trained model can reveal complex multi-input–target relationships that are difficult to define manually. Furthermore, results may vary with the accumulation of additional data.

The ETD's efficiency—requiring less time and energy—and its objectivity and accuracy make it a flexible and practical tool to be applied across diverse medical scenes that prioritize lightweight and quietness, particularly in screening and health monitoring scenes. Multimodal analysis may produce better results; for instance, AUC of binary depression status increased from 0.72 to 0.76 with networked smartphone sensors combining to text messages [12], and the binary accuracy increased from 76.27% to 95.60% when AU features were added to acoustic features in the baseline MFCC-based RNN [29]. As wearable devices continue to gain popularity, easily obtainable physical signals such as ECG and photoplethysmography can be integrated as additional modalities to enhance clinical outcomes. In our work, we currently use visual, acoustic, and text information jointly, which we believe may be a key point in the high performance observed and should receive more attention in future studies. As audiovisual features are also related to other conditions such as anxiety disorder and schizophrenia [48], or to distinguish between MDD or bipolar depression [49], aggregating multiple paradigms may further improve efficacy.

## Conclusions

The Q&A method showed greater efficacy compared to MID, and combining paradigms may yield better results than using individual paradigms alone. Visualization interpretation showed that the AI method acquired knowledge that aligns with medical expertise and identified several potentially significant markers. By applying AI to multimodal audiovisual signals, these findings position the ETD as a valuable, objective tool for screening MDD and show potential for applications across a broader spectrum of psychiatric disorders with various data modalities.

## Limitations

The efficacy of the modalities remains inadequately explored. Automatically detected AUs may not achieve the reliability of human-labeled results. Additionally, conclusions drawn from current analyses may require revision as sample sizes increase, particularly in deep learning frameworks.

### Acknowledgments

The study was supported by Sci-Tech Innovation 2030 – Major Project of Brain science and brain-inspired intelligence technology (2021ZD0200600) and the Beijing Municipal Administration of Hospitals Incubating Program (PX2019068).

### Data Availability

The datasets analyzed during this study are not publicly available as the agreement is only between the participants and researchers but are available from the corresponding author on reasonable request.

### Conflicts of Interest

None declared.

### Multimedia Appendix 1

Definitions, individual prediction results, and attention scores.
[DOCX File (Microsoft Word File), 565 KB-Multimedia Appendix 1]

### References

1. Depression and other common mental disorders: global health estimates. World Health Organization. 2017. URL: https://www.who.int/publications/i/item/depression-global-health-estimates [Accessed 2025-03-21]
2. HAMILTON M. A rating scale for depression. J Neurol Neurosurg Psychiatry. Feb 1960;23(1):56-62. [doi: 10.1136/jnnp.23.1.56] [Medline: 14399272]

3. Beck AT, Steer RA, Ball R, Ranieri W. Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. J Pers Assess. Dec 1996;67(3):588-597. [doi: 10.1207/s15327752jpa6703_13] [Medline: 8991972]
4. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med. Sep 2001;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
5. Pichot P. Self-report inventories in the study of depression. In: Hippius H, Klerman GL, Matussek N, editors. New Results in Depression Research. 1986:53-58. [doi: 10.1007/978-3-642-70702-5_7]
6. Ben-Porath YS. Assessing personality and psychopathology with self-report inventories. In: Handbook of Psychology. 2003:553-577. [doi: 10.1002/0471264385]
7. Rosa MJ, Portugal L, Hahn T, et al. Sparse network-based models for patient classification using fMRI. Neuroimage. Jan 15, 2015;105:493-506. [doi: 10.1016/j.neuroimage.2014.11.021] [Medline: 25463459]
8. Li X, La R, Wang Y, et al. EEG-based mild depression recognition using convolutional neural network. Med Biol Eng Comput. Jun 2019;57(6):1341-1352. [doi: 10.1007/s11517-019-01959-2]
9. Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. Sci Rep. Oct 11, 2017;7(1):13006. [doi: 10.1038/s41598-017-12961-9] [Medline: 29021528]
10. Schwartz HA, Eichstaedt J, Kern ML, et al. Towards assessing changes in degree of depression through facebook. Presented at: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology; Jun 2014; Baltimore, Maryland, USA. [doi: 10.3115/v1/W14-3214]
11. Chandra Guntuku S, Buffone A, Jaidka K, Eichstaedt JC, Ungar LH. Understanding and measuring psychological stress using social media. ICWSM. 2019;13(1):214-225. [doi: 10.1609/icwsm.v13i01.3223]
12. Liu T, Meyerhoff J, Eichstaedt JC, et al. The relationship between text message sentiment and self-reported depression. J Affect Disord. Apr 1, 2022;302:7-14. [doi: 10.1016/j.jad.2021.12.048] [Medline: 34963643]
13. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun. Jul 2015;71:10-49. [doi: 10.1016/j.specom.2015.03.004]
14. Scherer S, Morency LP, Gratch J, Pestian J. Reduced vowel space is a robust indicator of psychological distress: a cross-corpus analysis. Presented at: ICASSP 2015 - 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Apr 19-24, 2025; South Brisbane, Queensland, Australia. [doi: 10.1109/ICASSP.2015.7178880]
15. Logan B. Mel frequency cepstral coefficients for music modeling. Proc of Ismir. 2000. URL: https://ismir2000.ismir.net/papers/logan_paper.pdf [Accessed 2025-05-25]
16. Cummins N, Epps J, Breakspear M, Goecke R. An investigation of depressed speech detection: features and normalization. Presented at: INTERSPEECH 2011; Aug 27-31, 2011; Florence, Italy. URL: https://www.isca-archive.org/interspeech_2011 [doi: 10.21437/Interspeech.2011-750]
17. Kim AY, Jang EH, Lee SH, Choi KY, Park JG, Shin HC. Automatic depression detection using smartphone-based text-dependent speech signals: deep convolutional neural network approach. J Med Internet Res. Jan 25, 2023;25:e34474. [doi: 10.2196/34474] [Medline: 36696160]
18. Gratch J, Artstein R, Lucas G. The distress analysis interview corpus of human and computer interviews. Presented at: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); 3123-3128; Reykjavik, Iceland. Oct 15, 2019.URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf [Accessed 2025-03-21]
19. Ekman P, Friesen WV. Facial action coding system (FACS): a technique for the measurement of facial actions. APA PsycTests. URL: https://doi.org/10.1037/t27734-000 [Accessed 2025-03-21]
20. Fiquer JT, Moreno RA, Brunoni AR, Barros VB, Fernandes F, Gorenstein C. What is the nonverbal communication of depression? Assessing expressive differences between depressive patients and healthy volunteers during clinical interviews. J Affect Disord. Oct 1, 2018;238:636-644. [doi: 10.1016/j.jad.2018.05.071] [Medline: 29957481]
21. Girard JM, Cohn JF, Mahoor MH, Mavadati S, Rosenwald DP. Social risk and depression: evidence from manual and automatic facial expression analysis. Presented at: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG); Apr 22-26, 2013:1-8; [doi: 10.1109/FG.2013.6553748] [Medline: 24598859]
22. Gavrilescu M, Vizireanu N. Predicting depression, anxiety, and stress levels from videos using the facial action coding system. Sensors (Basel). Aug 25, 2019;19(17):3693. [doi: 10.3390/s19173693] [Medline: 31450687]
23. Melfsen S, Osterlow J, Florin I. Deliberate emotional expressions of socially anxious children and their mothers. J Anxiety Disord. 2000;14(3):249-261. [doi: 10.1016/s0887-6185(99)00037-7] [Medline: 10868983]
24. Smith MC, Smith MK, Ellgring H. Spontaneous and posed facial expression in Parkinson's disease. J Int Neuropsychol Soc. Sep 1996;2(5):383-391. [doi: 10.1017/s1355617700001454] [Medline: 9375163]

25. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J Neurolinguistics. Jan 2007;20(1):50-64. [doi: 10.1016/j.jneuroling.2006.04.001] [Medline: 21253440]
26. McIntyre G, Gocke R, Hyett M, Green M, Breakspear M. An approach for automatically measuring facial activity in depressed subjects. Presented at: 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009); Sep 10-12, 2009:1-8; Amsterdam. [doi: 10.1109/ACII.2009.5349593]
27. Cai H, Yuan Z, Gao Y, et al. A multi-modal open dataset for mental-disorder analysis. Sci Data. Apr 19, 2022;9(1):178. [doi: 10.1038/s41597-022-01211-x] [Medline: 35440583]
28. Ray A, Kumar S, Reddy R, Mukherjee P, Garg R. Multi-level attention network using text, audio and video for depression prediction. Presented at: AVEC '19: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop; Oct 15, 2019:81-88; Nice France. [doi: 10.1145/3347320.3357697]
29. Rejaibi E, Komaty A, Meriaudeau F, Agrebi S, Othmani A. MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. Biomed Signal Process Control. Jan 2022;71:103107. [doi: 10.1016/j.bspc.2021.103107]
30. Holmes EA, Mathews A, Mackintosh B, Dalgleish T. The causal effect of mental imagery on emotion assessed using picture-word cues. Emotion. Jun 2008;8(3):395-409. [doi: 10.1037/1528-3542.8.3.395] [Medline: 18540755]
31. Weßlau C, Cloos M, Höfling V, Steil R. Visual mental imagery and symptoms of depression - results from a large-scale web-based study. BMC Psychiatry. Dec 2, 2015;15:308. [doi: 10.1186/s12888-015-0689-1] [Medline: 26631081]
32. Andrade J, May J, Deeprose C, Baugh SJ, Ganis G. Assessing vividness of mental imagery: The Plymouth Sensory Imagery Questionnaire. Br J Psychol. Nov 2014;105(4):547-563. [doi: 10.1111/bjop.12050] [Medline: 24117327]
33. Tiggemann M, Kemps E. The phenomenology of food cravings: the role of mental imagery. Appetite. Dec 2005;45(3):305-313. [doi: 10.1016/j.appet.2005.06.004] [Medline: 16112776]
34. Ge Y, Zhao G, Zhang Y, Houston RJ, Song J. A standardised database of Chinese emotional film clips. Cogn Emot. Aug 2019;33(5):976-990. [doi: 10.1080/02699931.2018.1530197] [Medline: 30293475]
35. Liu YJ, Yu M, Zhao G, Song J, Ge Y, Shi Y. Real-Time movie-induced discrete emotion recognition from EEG Signals. IEEE Trans Affective Comput. 2018;9(4):550-562. [doi: 10.1109/TAFFC.2017.2660485]
36. International Advisory Group for the Revision of ICD-10 Mental and Behavioural Disorders. A conceptual framework for the revision of the ICD-10 classification of mental and behavioural disorders. World Psychiatry. Jun 2011;10(2):86-92. [doi: 10.1002/j.2051-5545.2011.tb00022.x] [Medline: 21633677]
37. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE. May 16, 2018;13(5):e0196391. [doi: 10.1371/journal.pone.0196391]
38. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. Presented at: Proceedings of the 36th International Conference on Machine Learning. 6105-6114; 2019.URL: https://proceedings.mlr.press/v97/tan19a.html [Accessed 2025-03-21]
39. Yin L, Chen X, Sun Y, Worm T, Reale M. A high-resolution 3D dynamic facial expression database. Presented at: 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition; Sep 17-19, 2008; Amsterdam, Netherlands. [doi: 10.1109/AFGR.2008.4813324]
40. Kellnhofer P, Recasens A, Stent S, Matusik W, Torralba A. Gaze360: physically unconstrained gaze estimation in the wild. Presented at: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 27 to Nov 2, 2019:6912-6921; Seoul, Korea (South. [doi: 10.1109/ICCV.2019.00701]
41. King DE. Dlib-ml: A machine learning toolkit. J Mach Learn Res. 2009;10(3):1755-1758. [doi: 10.1145/1577069.1755843]
42. Che W, Feng Y, Qin L, Liu T. N-LTP: an open-source neural language technology platform for chinese. Presented at: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; Nov 2021; Online and Punta Cana, Dominican Republic. [doi: 10.18653/v1/2021.emnlp-demo.6]
43. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); Oct 22-29, 2017:618-626; Venice. [doi: 10.1109/ICCV.2017.74]
44. Pedregosa F, Varoquaux G, Gramfort A. Scikit-learn: Machine learning in Python. JMLR. 2011;12:2825-2830. URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf [Accessed 2025-03-21]
45. Girard JM, Cohn JF, Mahoor MH, Mavadati SM, Hammal Z, Rosenwald DP. Nonverbal social withdrawal in depression: evidence from manual and automatic analysis. Image Vis Comput. Oct 2014;32(10):641-647. [doi: 10.1016/j.imavis.2013.12.007] [Medline: 25378765]

46. Fu CHY, Williams SCR, Cleare AJ, et al. Attenuation of the neural response to sad faces in major depression by antidepressant treatment: a prospective, event-related functional magnetic resonance imaging study. Arch Gen Psychiatry. Sep 2004;61(9):877-889. [doi: 10.1001/archpsyc.61.9.877] [Medline: 15351766]
47. Hahn T, Marquand AF, Ehlis AC, et al. Integrating neurobiological markers of depression. Arch Gen Psychiatry. Apr 2011;68(4):361-368. [doi: 10.1001/archgenpsychiatry.2010.178] [Medline: 21135315]
48. Abbas A, Hansen BJ, Koesmahargyo V, et al. Facial and vocal markers of schizophrenia measured using remote smartphone assessments: observational study. JMIR Form Res. Jan 21, 2022;6(1):e26276. [doi: 10.2196/26276] [Medline: 35060906]
49. Ruihua M, Meng Z, Nan C, et al. Differences in facial expression recognition between unipolar and bipolar depression. Front Psychol. 2021;12:619368. [doi: 10.3389/fpsyg.2021.619368] [Medline: 34335353]

## Abbreviations

**AI:** artificial intelligence
**AU:** action unit
**AUC:** area under the curve
**CNN:** conventional neural network
**CS:** conventional scale
**DL:** deep learning
**EEG:** electroencephalogram
**ETD:** electronic tool for depression
**fMRI:** functional magnetic resonance imaging
**HAMD:** Hamilton rating scale for depression
**MDD:** major depressive disorder
**MFCC:** mel frequency cepstral coefficients
**MID:** mental imagery description
**PHQ-8:** Patient Health Questionnaire-8
**PHQ-9:** Patient Health Questionnaire-9
**Q&A:** question and answering
**QI:** combination of Q&A and MID
**RNN:** recurrent neural network
**SOTA:** state-of-the-art
**SQ:** combination paradigm of CS and Q&A
**SQIV:** combination paradigm of CS, Q&A, MID, and VW
**SRSI:** Self-Report Scales and Inventories
**VW:** video-watching paradigm

*Edited by Amaryllis Mavragani; peer-reviewed by Soroosh Tayebi Arasteh, Zhongxia Shen; submitted 23.02.2024; final revised version received 02.05.2025; accepted 06.05.2025; published 30.05.2025*

*Please cite as:*
*Chen D, Wang P, Zhang X, Qiao R, Li N, Zhang X, Zhang H, Wang G*
*Comparative Efficacy of MultiModal AI Methods in Screening for Major Depressive Disorder: Machine Learning Model Development Predictive Pilot Study*
*JMIR Form Res 2025;9:e56057*
*URL: https://formative.jmir.org/2025/1/e56057*
*doi: 10.2196/56057*