T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

論題	
Title	Detection of Depression Using Web-Interview Data by LLM Enhanced with Multimodal Features
著者	篠田 浩一
Authors	Isaac MORALES NOLASCO, Koichi SHINODA, Momoko KITAZAWA, Yuriko KAISE, Shunsuke TAKAGI, Genichi SUGIHARA, Taishiro KISHIMOTO
出典	電子情報通信学会技術研究報告, Vol. 125, no. 348, pp. 49-54
Citation	IEICE technical report, Vol. 125, no. 348, pp. 49-54
発行日 / Pub. date	2026, 1
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright(c) 2026 IEICE

Powered by T2R2 (Science Tokyo Research Repository)

一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
PRMU2025-33(2026-01)

Detection of Depression Using Web-Interview Data by LLM Enhanced with Multimodal Features

Isaac MORALES NOLASCO†, Koichi SHINODA†, Momoko KITAZAWA††, Yuriko KAISE††, Shunsuke TAKAGI†††, Genichi SUGIHARA†††, and Taishiro KISHIMOTO††

† Department of Computer Science, School of Computing, Institute of Science Tokyo W8-81, Ookayama 2–12–1, Meguro-ku, Tokyo, 152–8552 Japan

†† Center for Promotion of Interdisciplinary Research in Medicine and life Science, Keio University School of Medicine, Tokyo Japan Mori JP Tower F7, 1-3-1, Azabudai, Minato-ku, Tokyo 106-0041, Japan

††† Department of Psychiatry and Behavioral Sciences, School of Medical and Dental Sciences, Institute of Science Tokyo 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan

E-mail: †isaac@ks.c.titech.ac.jp, ††shinoda@c.titech.ac.jp,
†††{m.kitazawa,ykaise,tkishimoto}@keio.jp, ††††{takagi.s.659d,sugihara.g.44d0}@m.isct.ac.jp

Abstract Depression is a complex mental disorder that has been widely studied. Using various machine learning techniques, researchers have been able to predict whether an individual is healthy or experiencing depression. The most common approach involves analyzing a person's voice and speech content. Multimodal approaches improve prediction accuracy by incorporating facial features. With the rising popularity of large language models (LLMs), these models have recently been applied to evaluate symptoms of depression. In this paper, we explore the use of LLMs combined with audio and facial features for binary classification of depression using a Japanese dataset. The current method obtained an average accuracy of 0.7956 using the DSM-5 labeling with a 5-fold cross-validation.

Key words LLM, Depression Detection, Multimodal fusion, Web-Interview

1. Introduction

Depression, clinically defined in the DSM-5 [1], is characterized by the presence of at least five symptoms within a two-week period, including persistent sadness, loss of interest, sleep disturbances, appetite or weight changes, psychomotor agitation or retardation, fatigue, impaired concentration, feelings of worthlessness, and recurrent thoughts of death or suicide. It is a serious medical condition that often co-occurs with chronic illnesses such as diabetes, cardiovascular disease, Parkinson's disease, and Alzheimer's disease. Studying depression has a profound impact, as it remains one of the leading contributors to global disability and reduced quality of life. Given the rise in depression and the shortage of trained mental-health professionals, particularly in low and middle-income regions, there is a clear need to develop a fast and efficient diagnosis method [2]. Human assessments are time consuming, subjective, and difficult to scale, whereas AI systems can process large amounts of data efficiently, providing consistent and accessible screening tools that complement clinical judgment.

The vast majority of research has explored automated depression detection using cues from voice, text, and facial behavior [3], [4]. These approaches have shown that linguistic and acoustic patterns, as well as facial expressions, are closely related to depressive states. However, despite promising results, existing methods still face important challenges. Although several recent studies have incorporated LLMs into depression detection—for example, Sadeghi *et al.* used an LLM to clean and improve E-DAIC transcripts before combining them with facial features [5], Zhang *et al.* proposed SpeechT-RAG, which augments an LLM with retrieved speech-timing information [2], Hong *et al.* generated personalized textual descriptions from multimodal signals and fed them to an LLM [6], and Dong *et al.* used an LLM to denoise transcripts and construct an emotion lexicon that is then passed to a separate neural classifier [7]—these approaches still treat the LLM as an auxiliary pre-processing or feature-extraction module rather than operating directly on the patient's raw interview transcripts together with multimodal behavioral cues to make the final prediction. Additionally, limited availability of public data complicate the generalization and comparison of techniques across diverse populations.

Motivated by these gaps, our work aims to effectively exploit the natural-language understanding ability of LLMs by focusing on the patient's actual speech content during interviews. We build on the intuition that the linguistic content itself contains rich psychological indicators that LLMs are uniquely suited to interpret. At the same time, we recognize that depression manifests through multiple behavioral channels, so relying on a single modality may fail to capture

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2026 by IEICE

the full complexity of the condition.

The contribution of our research is as follows:

- Use the LLM directly on raw ASR transcripts without any intermediate NLP processing such as rephrasing, summarization, or key-phrase extraction, while at the same time feeding audio–visual embeddings straight into a fine-tuned LLM backbone instead of relying on separate fusion and classification heads. In our framework, high-level speech and visual representations are injected as additional tokens alongside the original transcripts, so that the LLM itself learns to jointly integrate all modalities and produce the final "depressed" vs. "healthy" decision in a single step. To the best of our knowledge, this is the first work to apply such an end-to-end multimodal LLM formulation to automatic depression detection.

2. Previous studies

2.1 Depression detection

Nowadays there are many modern technology-driven approaches to predict depression to increase the number of diagnoses that are done directly by a specialist. Examples include prediction of depression using wearable devices [8], analysing social media posts [9], or medical imaging such as functional Magnetic Resonance Imaging (fMRI) [10]. However, the majority of papers that predict depression use interviews as input; this might be due to the availability of interview-datasets. Furthermore, predicting depression through online interviews offers an easily accessible and cost-effective tool that can facilitate faster treatment, thus helping to prevent worst-case scenarios.

In recent years, prediction of depression using only one modality has become less common. Rodriguez *et al.* [11] use only speech to predict depression by analysis of spectrogram using CNN and CNN-LSTM. Tao *et al.* [12] use Multi Local Attention, and LSTM to predict depression using features extracted from audio only. Pan *et al.* [13] focus on the image modality to predict depression while developing a strategy to maintain anonymized physical features of the volunteers.

Multimodal approaches have shown a stronger performance compared to solo modality. Hao Sun *et al.* [14] pointed out that using crossmodal attention is limited to two modalities, and they proposed using a tensor instead of a traditional matrix to better capture the relationship between three modalities. The paper written by Kumar *et al.* [15] used MTCNN for video, TS-CAN for physiological, ResNet-18 for audio, and RoBERTa for text modalities, and a strategic fusion of modality-specific networks including CNN-RNN, Transformer, MLP, and ResNet-18.

The previous work using the Japanese multimedia Mental-health Interview Dataset (J-MIND) [16] also uses a multimodal combination extracted features from speech (spectrogram encoder-decoder), text (BERT), and video (Video Swim Transformers) and merged them using a MLP, their main novelty was comparing the utterance analysis vs the full interview session.

2.2 Multimodal LLMs

A Multimodal LLM (MLLM) typically consists of: (i) modality encoder(s), (ii) an input projection module that concatenates, aligns, or fuses multimodal features, (iii) an LLM backbone, (iv) an output projection, and (v) a modality generator [17], [18]. The following studies are close to MLLMs, but because their output is only a prediction rather than regenerated modalities, we regard them as LLMs enhanced with multimodal features.

Zhang *et al.* [19] presented SpeechT-RAG, a Retrieval-Augmented Generation framework that leverages acoustic temporal patterns for depression detection. They used this SpeechT-RAG to improve the performance of the LLM to predict depression. Hong *et al.* [6] compared different approaches to predict depression, one approach uses LLM enhanced with a multimodal approach. They include in the prompt a personalized description of the volunteer that was automatically generated by another subprocess according to the characteristics of the person (age, gender, native, *etc.*).

After reviewing the current State Of the Art, there is still a lack of application of LLM to address the depression prediction. The actual LLM depression prediction papers do not use the multimodal approach with the direct transcript interview. Considering the high potential of the LLM, we decided to test the performance of LLM in the J-MIND. Inspired by the work of EmotionLlama [20] that uses the LLM to predict emotion in videos, we decided to apply a similar approach to predict binary depression.

3. J-MIND Dataset

3.1 COI-NEXT project

The COI-NEXT Project is a national Japanese program designed to create industry–academia–government co-creation platforms that drive innovation, regional revitalization, and long-term societal transformation. It supports universities as central hubs that work with companies, local governments, and other organizations to develop sustainable, independent innovation centers. In particular for this project, COI-NEXT represents an ongoing effort for a comprehensive and authentic approach in mental health research in Japan, and the final goal is to be able to diagnose depression in normal or daily conversations.

3.2 Criteria HAMD

It is necessary to differentiate the definition of a depression from how to measure its severity. Nowadays, the most popular metrics are the Hamilton Depression Rating Scale (HAMD-17) and the Patient Health Questionnaire (PHQ-8). HAMD-17 is used by clinicians, has 17 items, and according to the answers a person can be labeled as normal (0-7 points), mild depression (8-13 points), moderate depression (14-18 points), severe depression (19-22 points) and very severe depression (more than 22 points). PHQ-8 is a self-rating questionnaire with 8 items on a 0-3 Likert scale.

3.3 Database design

Zoom interviews offer a unique perspective, where the patient can be in the comfort of their home and have a easier diagnose than going to the clinic. The J-MIND: Japanese multimedia Mental-health Interview Dataset is built from interviews conducted across multiple hospitals, enriching the diversity of the data. The hospitals involved are Asaka Hospital, Science Tokyo Hospital, Nagatsuda Ikoinomori Clinic, Keio University Hospital, and Tsurugaoka Garden Hospital.

The first version of the J-MIND depression dataset contained 91 patients, but due to the withdrawal of some volunteers, some data have been deleted. For the current paper

we present a second version and a third version of J-MIND depression dataset. The second version objective was to recreate as much as possible the first dataset. This version contains 89 patients, with a gender distribution of 40 male and 49 female participants, and age from 20 to 76. The third version, contains 168 patients, the gender distribution revealed 71 male and 98 female participants, and mantains the same age range.

For the labeling, we consider two proposals: HAMD and DSM. For a binary classification, the person is labeled as healthy if the HAMD score is less than or equal to 7, and if it is greater than 7, the person is considered depressed. The second approach is to divide the data according to the diagnosis made by the physician according to the DSM-5 manual.

The preprocessing of the data is described as follows: The audio is originally separated by Zoom downloading process, so the dataset has the patient-only audio, the physician-only audio, and both person audio. In this research only the patient audio is used. The videos were cropped to focus only on the patient using YoLov8 nano. The speech was divided into different utterances using the Voice Acitivity Detection (VAD) based on the Malaya speech Toolkit. For textual representation we used Google ASR transcripts with a random prompt of our bank of formats as an example: *The person in video says: ... "Determine the emotional state shown in the video, choosing from healthy, or depressed.".*

3.4 Related datasets

In this work, we use the E-DAIC dataset as a benchmark to compare our proposed LLM-based multimodal depression detection method with previous approaches. The E-DAIC dataset contains extracted video features, not the raw video (*e.g.* Action Units, head Pose, Eye-Gaze, vgg, ResNet), acoustic features extracted using openSMILE, the raw audio of the interview and their ASR transcripts. The PHQ-8 metric is used to classify volunteers, considering 11 points or more as depressed. These interviews were conducted by a virtual interviewer, who can be totally AI or controlled by a human. The dataset contains 275 interviews divided into three subsets: a trainset of 163 instances, a development set of 56 instances and a test set of 56 instances. Gender distribution revealed 170 male and 105 female participants, and the age variability within the dataset range from 18 to 69 years old. For the isolated version, the preprocessing focused on the audio, filtering it to only hear the voice of the patient, and the transcripts were generated by the ASR of Google Speech Recognition.

4. Proposed method

We proposed the Depression-LLaMA framework to predict binary depression based on the work of Emotion LLaMA [20], which consists of inputting video, audio, and transcripts into their respective encoders, concatenating the feature embeddings of each and adding those with the encoded prompt, to have a final response of the LLM about the state of the patient.

For audio we used the HuBERT Large for the English volunteers and HuBERT base Japanese from [21] for Japanese speakers. We loaded each wav audio file, extracted HuBERT-Large hidden-state features, applied mean pooling over time to produce a single 1024-dimensional embedding, and saved it as a .npy file.

The visual pipeline consists of three encoders: Video Masked Autoencoders (MAE) [22]: each video is divided into overlapping 16-frame clips, from which 768-dimensional MAE features are extracted and saved as .npy files. Face-MAE embeddings [23], an MAE variant optimized for facial dynamic expressions: the script extracts features from every second frame, averages them across tokens and time, projects them to a 1024-dimensional vector, and saves the resulting embedding as a .npy file. EVA embeddings [24] : the first frame of each video is extracted, passed through a frozen EVA-ViT model to obtain its feature representation, and saved as .npz files.

The HuBERT, Video-MAE, and Face-MAE features are then adapted to 1024-dimensional embeddings. Each embedding is split along the channel dimension, projected through dedicated linear layers, stacked, and concatenated with the EVA and CLS-token features to form the final multimodal input to LLaMA. This feature vector, combined with the prompt, is fed into the LLaMA LLM to predict the patient's state. The LLaMA model is fine-tuned using the LoRA method. The final output of the system is a binary LLM response: "depressed" or "healthy". The overall process is represented in the Figure 1.

5. Experiments

5.1 Experimental conditions

All experiments were conducted on the TSUBAME supercomputer at the Institute of Science Tokyo. We used one *f*-node equipped with 2 NVIDIA H100 GPUs and CPU AMD EPYC 9654 with 384 GB of memory with PyTorch in distributed data-parallel mode (world size = 4) using the NCCL backend and CUDA devices. As backbone, we used LLaMA-2-7B; all LLaMA weights were frozen and we applied LoRA to the self-attention projections (q proj and v proj), with rank r=64, scaling α=16, and dropout p =0.05. We fine-tuned the model with the AdamW optimizer and a linear-warmup cosine-decay learning-rate schedule for at most 10 epochs with 1,000 iterations per epoch. The random seed was set to 42 for all experiments.

For the J-MIND, we used an initial learning rate (lr) of 5.0×10^{-6}, a minimum lr of 1.0×10^{-6}, and a warm-up lr of 5.0×10^{-7}. The lr was linearly increased from the warm-up value during the first 1,000 warm-up steps and then decayed with a cosine schedule down to the minimum learning rate. We performed subject-independent 5-fold cross validation. The same folds were used for both HAMD and DSM labelings to enable a fair comparison. For the "long" setting we extracted features from the first 5 minutes of each interview (2 minutes for HuBERT due to computational constraints), while for the "utterance" setting we applied the VAD-based segmentation described in Section 3.3 and aggregated utterance-level predictions by majority vote. For the E-DAIC dataset, the initial lr was set to 3.0×10^{-6}, the minimum lr to 5.0×10^{-7}, and the warm-up lr to 1.0×10^{-7}. We followed the official train/test split.

5.2 Results

5.2.1 Speech Length Analysis

In the paper of Lam [16], the best results were obtained using short utterances. In order to analyze the effect of the utterance we proposed two different approaches: long and utterance. For the Utterance approach, first all the data were divided according to the detected utterances by the VAD. For

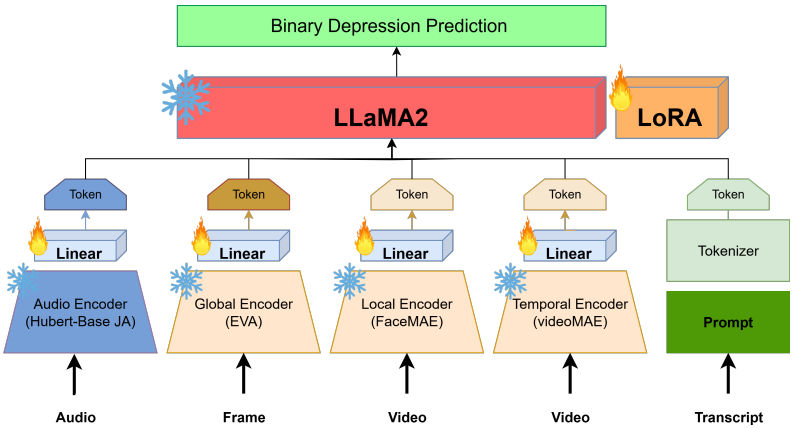

Fig. 1: Diagram of Depression-LLaMA. Based on [20], we extract different features from the interview; a linear block generates the tokens, which are concatenated with the prompt. The LLaMA model is fine-tuned with LoRA to obtain the binary depression prediction from the LLM response.

Dataset (num of subjects)	Avg Acc (5-fold cross val)		
	Part B long	Part B utte	Part A long
V2_HAMD (89)	0.7418	0.6493	0.7162
V2_DSM (74)	0.7428	-	0.6885
V3_HAMD (168)	0.7205	0.5889	0.6983
V3_DSM (132)	**0.7956**	-	0.7198

Table 1: Depression Llama performance on J-MIND

Dataset (num of volunteers)	Avg Acc (5-fold cross validation)	
	Part B long	Part A long
V3_HAMD (132) without bipolar	0.6515	0.6740
V3_DSM (132)	**0.7956**	0.7198

Table 2: Label performance.

the final prediction, the most common label among all the utterances is considered as the final prediction. For the long approach, the first 5 minutes of the video were used to obtain the features embeddings, except for the Hubert embeddings that were limited to 2 minutes due to computational limits.

The results in Table 1 reflect that LLM method has a better performance with the long approach than the utterance approach. The experiment was repeated using the V3 with more patients and the behaviour was consistent, the utterance approach had a lower accuracy (0.59) than the long approach (0.72).

5.2.2 Different Interview Section

The interview is divided into 5 sections. In this study we focus only on the first two sections; the first section A is a free talk where the psychologist asks the volunteer questions regarding patient's worries, the second section B is a structured part where the patient talks about depression episode. Comparing these two sections we can see how the model behaves between a non-structured vs a structured interview.

According to the results in Table 1 the structured part B, that was used originally in the previous study [16], performs better than the part A unstructurated. This pattern repeats independently from the label strategy used. The comparison was made only using long approach; the omission of utterance approach was due to the fact that it showed lower performance in comparison to the long approach.

5.2.3 Comparison of Label

From the results of the Table 1 the DSM label showed a higher accuracy than the HAMD dataset. However, the HAMD dataset also included bipolar people who had a HAMD score, so in order to have a fair comparison we trained and evaluated the model using the same volunteers. The results in Table 2 show that better classification performance using the DSM label prevails.

5.2.4 Comparison with the SOTA

To analyze how this LLM method's behavior with the current methods we decided to test it with the E-DAIC dataset. However, as the dataset has no raw videos we used the ResNet50 and VGG features, because those are the closer representation to our current embeddings. Using the raw audio and transcripts we obtained an accuracy of 0.6071, and with the isolated audio and isolated ASR transcripts it increased to 0.6250.

In the Table 3 we can see that our method has results comparable to the state of the art. The work by Dong *et al.* [7] attains 0.8214 accuracy by first using an LLM (ChatGLM3-6b) to restate the interview texts and then constructing a fine-grained emotion matrix that is processed by a dedicated CNN–attention classifier. Their architecture is highly optimized for E-DAIC: it focuses solely on text, explicitly models the temporal trajectory of emotions with an emotion lexicon, and is trained on carefully cleaned transcripts, which likely explains its superior performance on this text-only setting. In contrast, our approach feeds raw transcripts together with audio–visual embeddings directly into a LLaMA-2 backbone, asking the LLM itself to integrate all modalities and output the final decision. Thus, our framework is more general and naturally suited to scenarios where rich non-verbal cues are available, and it is complementary to Dong *et al.*'s emotion-centric design: their interpretable emotion matrix could be incorporated as an additional modality in our LLM, potentially combining their strong text-based emotion representation with our capability to jointly reason over multimodal signals.

As the EDAIC dataset does not provide the raw video, we could not extract the same visual features. In order to know how much affects the specific visual features of the method, we also test our method replacing the visual features with ResNet50 and VGG features. In Table 4 we can see how much the LLM enhanced by multimodal inputs degrades by replacing these visual features.

To evaluate how each factor contributes to the model performance, the model trained on the V2 with the HAMD

Method	Modality	(test set) Accuracy	Precision	Recall	F1
Depression-llama	A+V+T	0.6250	0.6489	**0.6250**	0.6264
Gimeno [25]	A+V	–	0.59	0.58	0.56
Dong [7]	T	**0.8214**	**0.7692**	0.5882	**0.6667**

Table 3: Binary classification results on E-DAIC.

Dataset (num of volunteers)	Visual features	Accuracy (5-fold cross validation)
V3_DSM_partB (132)	MAE, FACE_MAE, EVA	**0.7956**
V3_DSM_partB (132)	VGG16, ResNet	0.6583
EDAIC	VGG16, ResNet	0.6250 (test set)

Table 4: Performance of Depression LlaMA using different visual features.

labeling was tested when inputting just one feature and also ommitting one feature, see Table 5. The raw Llama-2 with 7 billion parameters had the lowest performance prediction with an accuracy of 0.3088 with a consistent standard deviation of 0.06. The modality that contributes the most were the face MAE with an accuracy of 0.5717 ±0.11.

5.3 Discussion

Regarding the difference in performance comparing the utterance vs the long approach, the long approach shown better results, we believe is due to LLM has a better comprehesion of the context if it includes the whole monologue instead of cropped words, that can be seen as noise to the LLM.

The section B is a structured interview, while the section A has no structured. The section B show better performance, because it is easier to compare the answers to almost the same questions (the questionnaire changes according to each person's answer, so it is personalized).

In this study the initial binary diagnosis made by the DSM label showed a better performance compared to the HAMD, the HAMD measures how severe a person has depression symptoms; however, in the case of the DSM label, there might be cases where a person has depression, while having low "healthy" HAMD score, in other words, depression without clear symptoms. It is necesary to continue the evaluation of the labeling in order to decide which label is the most effective to predict depression. When the data was increased, using the HAMD score, the accuracy values decreased. However, it is important to continue to increase the dataset, so that the model can have a stronger generalization.

The limited features of EDAIC dataset restrict the research to only using those available features. Working with more efficient visual features (MAE, EVA and FaceMAE) can increase the prediction performance, so we can hypothesize that the performance of our approach can be even greater if we were able to extract these features. The EDAIC dataset in the audio and transcripts contains: interviewer, patient and sometimes a third person such as a secretary; our method was developed based on the COINEXT dataset where the patient audio and transcripts is the only input. Then, when isolated the transcripts and audio to only be from the patient, the accuracy increased.

The work of Gimeno *et al.* [25] maintains a more traditional approach: extracting features, aligning them and using a transformer encoder for classification. Our approach not only uses feature extraction, but let the LLM predict the final prediction by inputing all the features. Compared with the state of the art, the model developed by Dong *et al.* [7] shows a really big gap from the binary classification SOTA. In their work, they use LLM to clean up the messy transcripts and develop an emotion lexicon. With the new transcripts and this emotion lexicon, they were able to construct an emotion matrix describing a person's emotional state and mood swings over time. The depression prediction used this matrix as input to a neural network. Their works focus on structured interviews, making difficult to use this model to predict depression in daily or normal conversation. The gap in the results can also be explained by the use of different LLM, LlaMA 7b vs ChatGLM3-6b, but further experiments are required to confirm it.

6. Conclusion

In this work, we proposed a multimodal LLM-based framework for binary depression detection using web-interview data. By combining audio, facial, and textual features with a LLaMA-based backbone enhanced via LoRA, our method achieves results comparable to existing state-of-the-art approaches. On the J-MIND, the proposed model attains up to 79.6% accuracy using DSM-based labels and the structured Part B section in the long-segment setting, and we showed that long segments outperform utterance-level segmentation and that structured interviews are more informative than free conversation for this task. In future work, we plan to extend this framework beyond binary classification to predict depression severity scores (HAMD, MADRS). Furthermore, most existing studies, including ours, analyze only a single visit; an important next step is to model longitudinal changes in depression severity over multiple interviews from the same patient over time. This could provide a more realistic view of how depressive symptoms evolve and help move toward continuous real-world monitoring of mental health.

7. Acknowledgement

This work was supported by JST COI-NEXT Grant Number JP MJPF2101, JAPAN.

References

[1] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision, 5th, text revision ed., American Psychiatric Association, Washington, DC, 2022.

[2] Y. Zhang, X. Jia, Y. Yang, N. Sun, S. Shi, and W. Wang, "Change in the global burden of depression from 1990–2019 and its prediction for 2030," Journal of Psychiatric Research, vol.178, pp.16–22, 2024.

[3] Y. Li, S. Kumbale, Y. Chen, T. Surana, E.S. Chng, and

Llama+LoRA (text)	Hubet (Visual)	EVA (Visual)	face MAE (Visual)	video MAE (Visual)	5 Folds Average Accuracy
Raw model without LoRA	-	-	-	-	0.3088 SD 0.06
✓	-	-	-	-	0.5465 SD 0.10
-	✓	-	-	-	0.4630 SD 0.14
-	-	✓	-	-	0.4850 SD 0.10
-	-	-	✓	-	0.5717 SD 0.11
-	-	-	-	✓	0.4961 SD 0.13
✓	✓	✓	✓	-	0.6939 SD 0.10
✓	✓	✓	-	✓	0.6939 SD 0.08
✓	✓	-	✓	✓	0.4606 SD 0.08
✓	-	✓	✓	✓	0.6695 SD 0.12
✓	✓	✓	✓	✓	0.7162 SD 0.09

Table 5: Modalities analysis. Average Accuracy 5-fold cross validation

C. Guan, "Automated depression detection from text and audio: A systematic review," IEEE Journal of Biomedical and Health Informatics, vol.29, no.10, pp.7498–7513, 2025.

[4] P. Zhai and L. He, "Deep learning-based depression recognition through facial expression: A systematic review," Neurocomputing, vol.627, p.129605, 2025.

[5] M. Sadeghi, R. Richer, B. Egger, L. Schindler-Gmelch, L.H. Rupp, F. Rahimi, M. Berking, and B.M. Eskofier, "Harnessing multimodal approaches for depression detection using large language models and facial expressions," npj Mental Health Research, vol.3, no.1, p.66, 2024.

[6] J.S.Z. Hong, T.Z. Delaya, S. Chan Yin Kit, P.C. Ng, and X. Miao, "Exploring machine learning and language models for multimodal depression detection," arXiv preprint arXiv:2508.20805, 2025.

[7] Z. Dong, Z. Wang, T. Yuan, B. Ma, N. Yan, and M. Sun, "Interview-based depression detection using llm-based text restatement and emotion lexicon," ICCC, p.1–15, October 2025.

[8] X. Xu, H. Zhang, Y. Sefidgar, Y. Ren, X. Liu, W. Seo, J. Brown, K. Kuehn, M. Merrill, P. Nurius, S. Patel, T. Althoff, M.E. Morris, E. Riskin, J. Mankoff, and A.K. Dey, "Globem dataset: Multi-year datasets for longitudinal human behavior modeling generalization," NeurIPS 2022 Workshop on Datasets and Benchmarks, University of Washington, USA, 2022.

[9] A. Zafar, D. Aftab, R. Qureshi, Y. Wang, and H. Yan, "Multi-explainable temporalnet: An interpretable multimodal approach using temporal convolutional network for user-level depression detection," Proceedings of the IEEE CVPR Workshops, p.2258–2265, June 2024.

[10] H. Ye, Y. Zheng, Y. Li, K. Zhang, Y. Kong, and Y. Yuan, "Rh-brainfs: Regional heterogeneous multimodal brain networks fusion strategy," NeurIPS 2023 Conference, Southeast University; Zhongda Hospital, China, 2023.

[11] S. Rodriguez, S.H. Dumpala, K. Dikaios, S. Rempel, R. Uher, and S. Oore, "Predicting individual depression symptoms from acoustic features during speech," arXiv preprint, vol.arXiv:2406.16000, 2024.

[12] F. Tao, X. Ge, W. Ma, A. Esposito, and A. Vinciarelli, "Multi-local attention for speech-based depression detection," 2023 IEEE ICASSP, p.–, 2023.

[13] Y. Pan, J. Jiang, K. Jiang, Z. Wu, K. Yu, and X. Liu, "Opticaldr: A deep optical imaging model for privacy-protective depression recognition," arXiv preprint, vol.arXiv:2402.18786, 2024.

[14] H. Sun, Y.W. Chen, and L. Lin, "Tensorformer: A tensor-based multimodal transformer for multimodal sentiment analysis and depression detection," IEEE Transactions on Affective Computing, vol.14, no.4, pp.2776–2786, Oct. 2023.

[15] P. Kumar, S. Misra, Z. Shao, B. Zhu, B. Raman, and X. Li, "Multimodal interpretable depression analysis using visual, physiological, audio and textual data," Proceedings of the IEEE WACV, pp.5305–5315, 2025.

[16] C.H. Lam, N. Nah, K. Shinoda, M. Kitazawa, Y. Kaise, S. Takagi, G. Sugihara, and T. Kishimoto, "Detection of depression using web-interview data," Technical Report Vol. 124, No. 23, IEICE, May 2024.

[17] J. Wang, H. Jiang, Y. Liu, C. Ma, X. Zhang, Y. Pan, M. Liu, P. Gu, S. Xia, W. Li, Y. Zhang, Z. Wu, Z. Liu, T. Zhong, B. Ge, T. Zhang, N. Qiang, X. Hu, X. Jiang, X. Zhang, W. Zhang, D. Shen, T. Liu, and S. Zhang, "A comprehensive review of multimodal large language models: Performance and challenges across different tasks," arXiv preprint, vol.arXiv:2408.01319, 2024.

[18] R. AlSaad, A. Abd-alrazaq, S. Boughorbel, A. Ahmed, M.A. Renault, R. Damseh, and J. Sheikh, "Multimodal large language models in health care: Applications, challenges, and future outlook," Journal of Medical Internet Research, vol.26, p.e59505, 2024.

[19] X. Zhang, H. Liu, Q. Zhang, B. Ahmed, and J. Epps, "Speecht-rag: Reliable depression detection in llms with retrieval-augmented generation using speech timing information," Findings of the Association for Computational Linguistics: ACL 2025, pp.10019–10030, 2025.

[20] Z. Cheng, Z.Q. Cheng, J.Y. He, J. Sun, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann, "Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning," arXiv preprint arXiv:2406.11161, 2024.

[21] K. Sawada, T. Zhao, M. Shing, K. Mitsui, A. Kaga, Y. Hono, T. Wakatsuki, and K. Mitsuda, "Release of pre-trained models for the Japanese language," LREC-COLING 2024, pp.13898–13905, 5 2024.

[22] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," arXiv preprint arXiv:2203.12602, 2022.

[23] L. Sun, Z. Lian, B. Liu, and J. Tao, "Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition," arXiv preprint arXiv:2307.02227, 2023.

[24] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," Proceedings of the IEEE CVPR, p.——, 2023.

[25] D. Gimeno-Gómez, A.M. Bucur, A. Cosma, C.D. Martínez-Hinarejos, and P. Rosso, "Reading between the frames: Multi-modal depression detection in videos from non-verbal cues," arXiv preprint, vol.arXiv:2401.02746, 2024.