論文 / 著書情報
Article / Book Information

論題	ウェブ面接データを用いたうつ病の検出
Title	Detection of Depression Using Web-Interview Data
著者	Lam Cheuk Hee, Nah Nathania, 篠田 浩一, 北沢 桃子, 貝瀬 有里子, 高木 俊輔, 杉原 玄一, 岸本 泰士郎
Authors	Cheuk Hee Lam, Nathania Nah, 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. 124, no. 23, pp. 36-40
Citation	IEICE Technical Report, vol. 124, no. 23, pp. 36-40
発行日 / Pub. date	2024, 5
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) 2024 IEICE

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
PRMU2024-7(2024-05)

Detection of Depression Using Web-Interview Data

Cheuk Hee LAM†, Nathania NAH†, Koichi SHINODA†, Momoko KITAZAWA††, Yuriko KAISE††, Shunsuke TAKAGI†††, Genichi SUGIHARA†††, and Taishiro KISHIMOTO††

† Shinoda Laboratory, Department of Computer Science,
School of Computing, Tokyo Institute of Technology
W8-81, Ookayama 2–12–1, Meguro-ku, Tokyo, 152–8552 Japan
†† Hills Joint Research Laboratory for Future Preventive Medicine and Wellness, Keio University School of Medicine
Mori JP Tower F7, Azabudai 1–3–1, Minato-ku, Tokyo, 106–0041 Japan
††† Department of Psychiatry and Behavioral Sciences, Graduate School of Medical and Dental Sciences,
Tokyo Medical and Dental University
Yushima 1–5–45, Bunkyo-ku, Tokyo, 113–8510 Japan

E-mail: †{chlam,nathania}@ks.c.titech.ac.jp, ††shinoda@c.titech.ac.jp,
†††{m-kitazawa,ykaise,tkishimoto}@keio.jp, ††††{stakagi.psyc,gen-psyc}@tmd.ac.jp

Abstract This paper presents a method for integrating speech, text, and video modalities for multimodal depression detection. Our work leverages shorter utterances to enhance depression detection accuracy, rather than relying on traditional long-term approaches. We introduce the COI-NEXT dataset, comprising authentic clinical interviews conducted through Zoom. Our experiments show that video modalities, particularly when using shorter utterances, lead to improved accuracy for depression detection in patients. Despite limitations due to data scarcity, this work offers valuable insights into multimodal depression detection, emphasizing the significance of multimodal integration in mental health research.

Key words Depression Detection, Web-Interview, Multimodal Learning

1. Introduction

Depression is a pervasive mental health disorder characterized by persistent feelings of sadness, hopelessness, and disinterest in daily activities [1]. It affects millions of people worldwide and poses significant challenges to both individuals and society as a whole. Early detection and intervention are crucial for effective management and treatment of depression, yet accurately diagnosing this complex condition remains a formidable task.

Traditional approaches to depression detection often rely on assessments by healthcare professionals, which can be time-consuming and subjective. Moreover, the heterogeneous nature of depression symptoms further complicates diagnosis, as individuals may present with varying combinations and severity of symptoms, as delineated in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [2].

In recent years, advancements in technology and data analysis techniques have opened new avenues for improving depression detection. Multimodal approaches that leverage data from diverse sources, such as speech, text, and video, hold promise for enhancing the accuracy and reliability of depression diagnosis [13]–[15]. By integrating information from multiple modalities, these approaches can capture a more comprehensive picture of an individual's mental health status and aid in the early identification of depressive symptoms.

This paper explores multimodal depression detection, with a focus on leveraging speech, text, and video data. We investigate how machine learning algorithms can analyze patterns and features within these modalities to infer indicators of depression. By examining the effectiveness of multimodal approaches in real-world settings, we aim to contribute to the development of more accurate and accessible tools for depression screening and diagnosis.

2. Related Work

2.1 Depression Evaluation Metrics

Depression assessment tools are vital in evaluating and quantifying depressive symptoms in patients. The Patient Health Questionnaire-9 (PHQ-9) [3] and the Hamilton Depression Rating Scale-17 (HAMD-17) [4] are widely used measures by researchers and clinics for this purpose.

The PHQ-9 is a self-reported questionnaire aligned with DSM-5 criteria for Major Depressive Disorder (MDD). It assesses depressive symptoms over the past two weeks, with respondents rating each item on a severity scale. In comparison, the HAMD-17 is a clinician-administered tool comprising 17 items covering various symptom domains of depression. Clinicians rate each item based on the patient's responses during a structured interview, providing a comprehensive evaluation of symptom severity. The HAMD-17 scores range from 0 to 52, enabling clinicians to categorize depression severity and guide treatment decisions. This work uses HAMD-17 scores as a means to evaluate depressive symptoms.

2.2 Depression Detection

In recent years, depression detection methodologies have undergone significant advancements, particularly with the integration of machine learning algorithms. Previous studies have demonstrated

Category	HAMD-17 Score
Healthy	0-7
Mild Depressive	8-13
Moderate Depressive	14-23
Severe Depressive	24+

Table 1 Depression Severity Categories over Hamilton Depression Rating Scale (HAMD-17) according to the UK National Institute for Health & Clinical Excellence

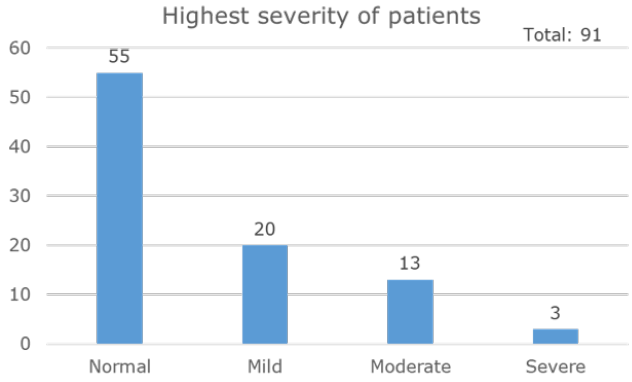

Fig. 1 Depression Severity Distribution of Patients in the COI-NEXT Dataset

the effectiveness of these algorithms in capturing subtle variations in depressive states through the analysis of acoustic features and linguistic cues in speech data [5], [6]. Similarly, text-based approaches leverage natural language processing techniques to be able to analyze sentiment, language patterns, and semantic content in social media posts [8], showing promise in identifying linguistic markers associated with depressive symptoms [7]. Video-based methodologies have also gained traction, combining facial expressions, body language, and speech cues to provide a comprehensive perspective [9], with multimodal integration improving accuracy, facilitated by advancements in computer vision and deep learning.

However, challenges persist within the domain of depression detection. The heterogeneous nature of depressive symptoms presents difficulties in establishing universal features indicative of depression. Moreover, temporal dynamics within speech data, encompassing short-term fluctuations and longer-term patterns, necessitate models capable of capturing these nuances effectively. Current methods [13], [14] predominantly rely on Long-term Temporal Feature Extraction, potentially overlooking subtle short-term temporal variations within facial expressions. To address these challenges, we introduce a Short-term Multimodal Correlation methodology, focusing on short sequences of patient data.

3. Methodology

This work is inspired by a multimodal system of audio and text [13] for emotion recognition. Unlike traditional approaches that rely on global temporal feature extraction, our method focuses on capturing short sequences of patient data, allowing for the detection of subtle, short-term variations and nuances in facial expressions and other modalities. By adopting this approach, we aim to enhance the model's sensitivity to short-term temporal variations, which can help improve our understanding of how depressive symptoms manifest among individuals. Our method integrates audio, text, and video data, and analyzes these multimodal features to aid in the development of robust depression detection models.

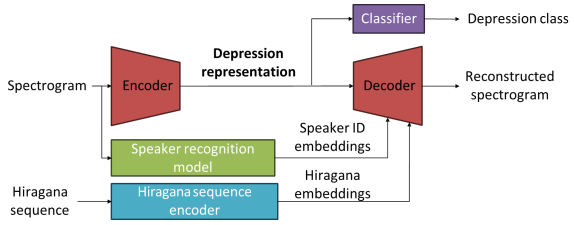

Fig. 2 Speech Model Architecture

3.1 Original Depression Dataset

The COI-NEXT dataset used in this work represents an ongoing effort for a comprehensive and authentic approach in mental health research in Japan. Zoom interviews offer a unique perspective, enhancing data quality and patient comfort. The data contained in this dataset originates from multiple medical institutions, contributing to a diverse collection. Specifically, datasets were acquired from Asaka Hospital, Tokyo Medical and Dental University Hospital, Nagatsuda Ikoinomori Clinic, Keio University Hospital, and Tsurugaoka Garden Hospital.

The dataset used in this work contains the initial five minutes of interview data for sessions conducted by a healthcare professional in Japanese, in which the individual's evaluation consists of the highest severity for depression among multiple visits. Patients are evaluated according to the HAMD-17 scale.

The data includes a total of 91 individual patients whose severity of depressive symptoms are as follows: 55 patients categorized as normal, 20 as mild, 13 as moderate, and 3 as severe. These categories are determined using their HAMD-17 scores as described in Table 1. Gender distribution revealed 42 male and 49 female participants. Patient types encompassed 17 individuals diagnosed with bipolar disorder, 49 with depression, and 25 classified as healthy with other non-psychiatric diseases. Age variability within the dataset ranged from 20 to 76, with the majority falling within the 45-50 age bracket.

In the data pre-processing phase, where both video and audio data are available for each sample, we take specific steps to ensure the data's readiness for subsequent analyses. The video data processed by cropping the patient's video, then extracting individual images corresponding to frames, allowing for a more granular analysis. The audio data is isolated and organized into utterance data, distinguishing the distinct speech segments within the dataset. To enhance the quality of the data, a noise and silence removal process is applied, refining the audio data for more accurate analyses. The utterances are then split, enabling a more focused examination of individual speech segments. Additionally, speech recognition techniques are employed to transcribe the spoken content into text data. During this process, both patient and interviewer timestamps are extracted to facilitate precise duration calculations.

3.2 Speech Depression Detection

The proposed speech depression detection model is based on an encoder-decoder architecture, drawing inspiration from a disentanglement representation model [10]. Our model is specifically crafted to discern depression-related features within speech data. In the encoding phase, the model takes 1024-dimensional wav2vec 2.0 features per speech frame, employing a weighted average computation of these features. This information is concatenated with a 256-dimensional speaker identity embedding, and the resulting data is processed through ConvNorm layers, BLSTM layers, and a downsampler operation. The downsampling operation is a key element, as it contributes to the creation of a feature array with controlled bottleneck dimensions. These dimensions are designed to selectively retain depression-related information while filtering out speaker identity and phonetic details.

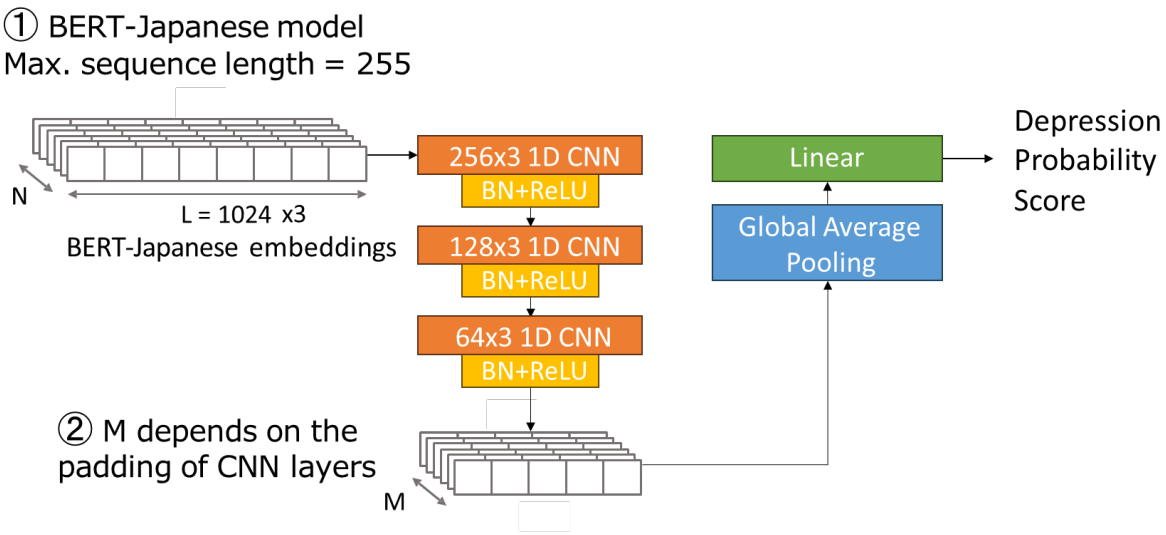

Fig. 3 Text Model Architecture

Moving on to the decoding phase, the model reconstructs mel-frequency spectrograms from the encoded information. This process involves the use of reconstructed loss functions ($\mathcal{L}_{r1}$ and $\mathcal{L}_{r2}$) to fine-tune the model's parameters. In addition to the encoder-decoder structure, the model incorporates a classifier component. This component consists of fully-connected layers, a dropout layer, and a final layer dedicated to depression class logits. The classifier's role is pivotal in encouraging the inclusion of depression-related information in the encoder's outputs.

Because interviews are conducted in Japanese, a Japanese Hiragana Encoder is used to encode phone information. This encoder serves a crucial purpose in the model architecture by making the encoder's representations phone-independent. This is achieved by taking a sequence of hiragana embeddings as input and outputting a representation for the entire hiragana sequence. Each hiragana is defined as a 128-dimensional feature array, contributing to a holistic understanding of depression in speech data.

During the training phase, the model employs mean-squared error loss ($\mathcal{L}_r$) for reconstruction losses and cross-entropy loss ($\mathcal{L}_e$) for depression class classification, as expressed by the sum of the reconstruction loss and the depression prediction loss:

$$\mathcal{L} = \mathcal{L}_r + \mathcal{L}_e \tag{1}$$

3.3 Text Depression Detection

Our proposed model for text-based depression detection leverages pre-trained Japanese BERT models and shares similarities with [11] research but introduces modifications in the number of CNN layers, filter configurations, kernel sizes, and the activation function for the final linear layer responsible for emotion classification.

The text embeddings, shaped as [N;L] for each utterance, are extracted using pre-trained Transformer-based models. Here, N represents the number of tokens in the utterance (excluding special tokens like [CLS], [SEP]), and L is the size of each token's feature. The selected pre-trained models are based on the Transformer architecture, trained on extensive text corpora to capture a broad understanding of textual data.

The text model builds upon the architecture inspired by [10] and incorporates several key modifications to enhance its performance. Our architecture is adjusted to have four 1D CNN layers with output channels set to 512, 256, 128, and 64, respectively. Each convolutional layer is followed by batch normalization for better stability, and ReLU activation functions are applied to capture non-linear relationships within the data.

Additionally, global average pooling is introduced to replace fully connected layers, enhancing the model's ability to capture the most salient features while reducing the risk of overfitting. The global average pooling is applied after the convolutional layers, and the resulting feature array is fed into a linear layer with a sigmoid activation function for depression severity predictions.

3.4 Video Depression Detection

The video model for depression detection, inspired by Video Swin Transformer Model [12], incorporates various components to form a comprehensive architecture. In this model, the overall structure consists of multiple stages, each characterized by a specific number of layers. The Swin-B version of video swin transformer was adopted, where the number of layers is distributed as 2, 2, 18, 2, and the channel number of hidden layers in the first stage (C) is set to 128.

The model's initialization involves adapting weights from a pre-trained Swin Transformer model, with adjustments made to the linear embedding layer and relative position biases to align with the depression detection task. The output layer is tailored to the specific needs of depression detection, accommodating both regression (predicting depression severity) and classification (representing different depression severity classes).

3.5 Multimodal Fusion

Our fusion model introduces the concept of "Short-term Multimodal Correlation" by ensuring strict temporal alignment through inputting short sequences and thus enhancing the model's ability to capture nuanced and contextually relevant information across multiple modalities.

Figure 5 presents an outline of our proposed approach. First, we extract feature representations from the processed speech, text, and video data. These feature vectors are then concatenated to make a unified multimodal representation. For multimodal fusion, we introduce an MLP architecture which allows our model to capture intricate dependencies among the input features. The output layer of the MLP produces a probability distribution across different depression severity classes. To predict the severity of depression, we apply softmax to obtain normalized probabilities and then select class with

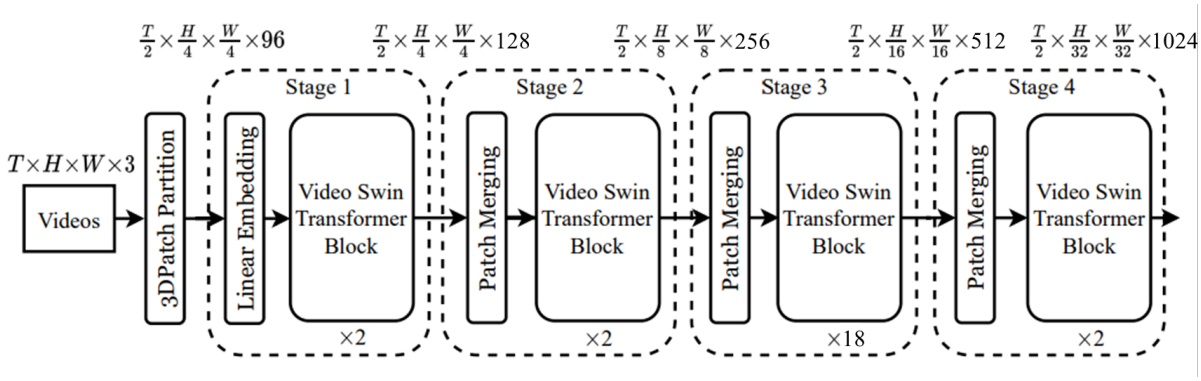

Fig. 4 Video Model Architecture

Model	Average
Speech	0.6691
Text	0.6145
Video	0.6360
Speech + Text	0.6731
Speech + Video	0.7405
Text + Video	0.7725
Speech + Text + Video	**0.8128**

Table 2 Comparison of Accuracy on Different Modalities

Maximum Duration of Utterances	Accuracy
3 seconds	0.6924
5 seconds	0.8313
10 seconds	0.8239
15 seconds	0.8190
20 seconds	0.8169
40 seconds (All utterances used)	0.8128

Table 3 Comparison of Accuracy in Utterance Duration

the highest probability as the final prediction.

By employing an MLP for multimodal fusion, our model can effectively leverage information from different modalities to make a unified and informed prediction about depression severity.

4. Experiments

4.1 Experimental Setup

The experimental configuration for the multimodal depression detection system adheres to a 5-fold cross-validation strategy. We evaluate our performance by comparing overall accuracy for each patient over their utterances, as well as comparing accuracy based on utterance duration as well as against other current depression detection methods.

4.2 Multimodal Performance

Our results in Table 2 show that the Speech model consistently performs well across folds, with an average accuracy of 66.91%. The Text model follows closely with an average accuracy of 61.45%, while the Video model achieves an average accuracy of 63.60%. Notably, the Speech model exhibits higher accuracy in most folds, indicating its proficiency in capturing depression-related cues from speech data. The Video model's performance is competitive, showcasing the significance of visual cues in depression detection. Combining modalities yields improvements, with the best-performing combination being Speech + Text + Video, achieving an average accuracy of 81.28%. This underscores the complementary nature of modalities, enhancing the overall detection capabilities.

Analyzing these results, each modality contributes unique information to the depression detection task. Speech and Video modalities, in particular, exhibit strong individual performances, and their combination with Text further enhances accuracy with the aids of MLP Score Fusion. The higher accuracy in speaker-based assessments suggests that the models are proficient in capturing distinctive speaker-specific patterns associated with depression.

4.3 Comparison in Utterance Duration

The analysis of depression detection accuracy based on different utterance duration thresholds is presented in Table 3. The table displays the accuracy scores corresponding to varying maximum durations for each utterance considered in the experiment. Notably, the accuracy scores exhibit discernible fluctuations across different utterance duration thresholds. For instance, shorter thresholds, such as 3 seconds, yield relatively lower accuracy scores compared to longer thresholds. Conversely, there is a notable improvement in accuracy as the utterance threshold increases, particularly up to 10 seconds. However, beyond this threshold, there is a slight decline in accuracy, indicating a potential diminishing returns effect with longer utterances.

Of particular interest is the observation that the highest accuracy score of 0.8313 is achieved with an utterance threshold of 5 seconds. This finding suggests that a this duration of utterances appears to be optimal for depression detection in the context of the experiment, which implies that utterances within this duration range may contain a sufficient amount of meaningful emotional cues and depressive symptoms for effective detection. However, it is worth noting that beyond a threshold of 10 seconds, the improvement in accuracy becomes marginal. Utterances exceeding this threshold may include additional irrelevant information or noise, potentially hindering the effectiveness of the depression detection model.

These findings have important implications for optimizing depression detection models. They indicate that tailoring model architectures and feature extraction methods to focus on shorter utterances may lead to improved performance. By concentrating on shorter utterances where meaningful emotional cues are concentrated, models could better capture the essential features associated with depression.

4.4 Comparison with Previous Methods

Table 4 presents a comprehensive comparison of accuracy across various depression detection models, where our proposed Short-term Multimodal Correlation model, denoted as "Ours(Short-term SA)," stands out prominently. The listed models, including Digital Biomarker [15], TAMFN [13], and DepMSTAT [14], predominantly

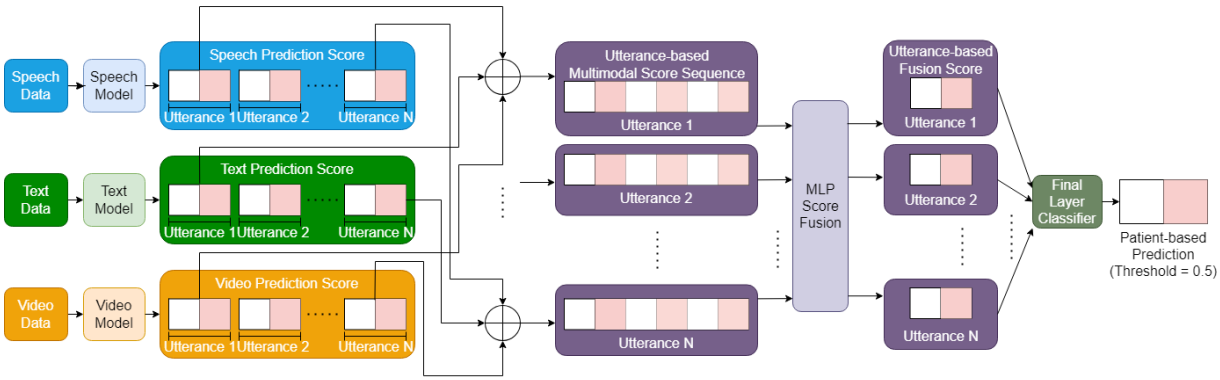

Fig. 5 Multimodal Architecture

Model	Accuracy
Digital Biomarker [15])	0.6413
TAMFN [13]	0.6923
DepMSTAT [14]	0.7363
Ours (Interview based)	0.6551
Ours(Short Utterance based)	**0.8128**

Table 4 Comparison of Accuracy with Previous Methods on the COI-NEXT Dataset

employ long-term methodologies in their approaches.

Notably, the accuracy achieved by our proposed model is significantly higher, marked at an impressive 81.28%. This notable improvement underscores the efficacy of our Short-term Multimodal Correlation approach. While traditional long-term methods, yield accuracies ranging from 64.13% to 73.63%, our model demonstrates a substantial leap in performance.

The improved accuracy of our Short-term Multimodal Correlation model can be attributed to its focus on short sequences, allowing our model to capture dynamic variations in patients' expressions and behaviors over brief intervals, offering a nuanced understanding of short-term multimodal correlations.

5. Conclusion

In conclusion, our study has demonstrated notable improvements in depression detection accuracy, particularly through the usage of short-term utterances in our dataset when compared to other long-term methods. We demonstrate the effectiveness of short utterances in capturing meaningful emotional cues and depressive symptoms, suggesting a promising avenue for optimizing depression detection models. However, it is important to acknowledge that the constraints of our dataset, such as the size and patient's Zoom setting may impact the generalizability of the findings. Despite these limitations, our study contributes valuable insights into the optimization of depression detection methodologies, paving the way for future advancements in the critical area of mental health research.

6. Acknowledgement

This work was supported by JST COI-NEXT Grant Number JPMJPF2101, JAPAN.

References

[1] World Health Organization, *The ICD-10 classification of mental and behavioral disorders: clinical descriptions and diagnostic guidelines*, vol. 1. World Health Organization, 1992.

[2] American Psychiatric Association, DSMTF, and others, *Diagnostic and statistical manual of mental disorders: DSM-5*, vol. 5, no. 5. American Psychiatric Association Washington, DC, 2013.

[3] B. Löwe, J. Unützer, C. M. Callahan, A. J. Perkins, and K. Kroenke, "Monitoring depression treatment outcomes with the patient health questionnaire-9," *Medical care*, pp. 1194–1201, 2004.

[4] M. Hamilton, "Development of a rating scale for primary depressive illness," *British journal of social and clinical psychology*, vol. 6, no. 4, pp. 278–296, 1967.

[5] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, vol. 71, pp. 103107, 2022.

[6] S. A. Almaghrabi, S. R. Clark, and M. Baumert, "Bio-acoustic features of depression: A review," *Biomedical Signal Processing and Control*, vol. 85, pp. 105020, 2023.

[7] Z. N. Vasha, B. Sharma, I. J. Esha, J. Al Nahian, and J. A. Polin, "Depression detection in social media comments data using machine learning algorithms," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 2, pp. 987–996, 2023.

[8] A. Trifan, R. Antunes, S. Matos, and J. L. Oliveira, "Understanding depression from psycholinguistic patterns in social media texts," in *European Conference on Information Retrieval*, 2020, pp. 402–409.

[9] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in reddit social media forum," *Ieee Access*, vol. 7, pp. 44883–44893, 2019.

[10] M. R. Makiuchi, K. Uto, and K. Shinoda, "Multimodal emotion recognition with high-level speech and text features," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 350–357.

[11] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6484–6488.

[12] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[13] L. Zhou, Z. Liu, Z. Shangguan, X. Yuan, Y. Li, and B. Hu, "TAMFN: Time-Aware Attention Multimodal Fusion Network for Depression Detection," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 669–679, 2022.

[14] Y. Tao, M. Yang, H. Li, Y. Wu, and B. Hu, "DepMSTAT: Multimodal Spatio-Temporal Attentional Transformer for Depression Detection," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[15] Z. Jiang, S. Seyedi, E. Griner, A. Abbasi, A. B. Rad, H. Kwon, R. O. Cotes, and G. D. Clifford, "Multimodal Mental Health Digital Biomarker Analysis from Remote Interviews using Facial, Vocal, Linguistic, and Cardiovascular Patterns," *IEEE Journal of Biomedical and Health Informatics*, 2024.