ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Research Article

Open Access

TMAK; Emotion Estimation, Mental Health Application, Information Recommendation - Application for Stress Estimation Model and Mild Cognitive Impairment Detection

Kazuyuki Matsumoto[1*], Keita Kiuchi[2], Ryota Nishimura[3], Manabu Sasayama[4], Hidehiro Umehara[5] and Mikio Shindo[6]

[1]Graduate School of Advanced Science and Technology, Tokushima University, Japan

[2]National Institute of Occupational Safety and Health, Japan

[3]Graduate School of Engineering Department of Computer Science and Engineering, Toyohashi University of Technology, Japan

[4]Department of Information Engineering, National Institute of Technology, Kagawa College, Japan

[5]Graduate School of Biomedical Sciences, Health Service, Counseling and Accessibility Center, Tokushima University, Japan

[6]Wellfort Co., Ltd., Japan

ABSTRACT

TMAK (Trustworthy Multimodal Affective Intelligence and Knowledge Engineering Laboratory) is advancing research on several key themes. First, we are researching methods to estimate human emotions from multimodal information (audio, language, images) and apply this to support the diagnosis of mental disorders, mental health conditions, and dementia. Second, we are researching techniques to analyze interest and reputation information from diverse web reviews and social media posts and utilize it for information recommendation. Third, we are developing empathetic dialogue systems by leveraging the rapidly advancing large-scale multimodal language models. All these research areas require feature extraction from large-scale data, making big data analysis platforms indispensable. Our laboratory is also advancing research on lightweight AI models, developing language models, emotion estimation, and stress estimation algorithms capable of running on local edge devices. This presentation will introduce past research examples, outline solutions using our proprietary emotion estimation technology based on multi-stage fine-tuning, and discuss future prospects such as cognitive function prediction.

***Corresponding author**

Kazuyuki Matsumoto, Graduate School of Advanced Science and Technology, Tokushima University, Japan.

Received: October 18, 2025; **Accepted:** November 11, 2025; **Published:** December 20, 2025

Keywords: Mental Health Applications, Stress Estimation, Mild Cognitive Inpairment Detection

Introduction

The Trustworthy Multimodal Affective Intelligence and Knowledge Engineering Laboratory (TMAK) conducts research aimed at advancing the understanding of human affect and cognition through artificial intelligence [1]. Our overarching goal is to develop technologies that enable reliable and human-centered affective computing systems. The laboratory's current research can be broadly classified into three main domains.

First, we investigate methods for estimating human emotions from multimodal information—including speech, language, and images—and applying these methods to support the assessment and diagnosis of mental disorders, mental health conditions, and dementia.

Second, we analyze interest and reputation information obtained from diverse online data sources like web reviews and social media content. We perform preprocessing, such as advanced knowledge-based automatic construction, aiming to build data-driven models for information recommendation and reputation analysis.

Third, leveraging the rapid advancement of Multimodal Large Language Models (MLLMs), we develop empathetic dialogue systems capable of context awareness and emotional responses by utilizing technologies like LLM agent systems and deep reinforcement learning.

Across these research directions, efficient feature extraction and large-scale data analysis play an essential role. Accordingly, our laboratory is also engaged in the development of lightweight AI models that can operate effectively on local edge devices. This includes the implementation of compact language models, emotion estimation, and stress estimation algorithms that function independently of cloud-based computation.

In this paper, we present an overview of our ongoing research activities, focusing primarily on stress estimation models and

multimodal cognitive function estimation technologies. By introducing representative examples of our previous work, we describe our original technical approaches and discuss future perspectives toward the realization of trustworthy affective and cognitive AI systems.

Related Research

Emotion recognition in conversation is a key artificial intelligence task that aims to infer human affective states by integrating multimodal information. Previous studies have developed various benchmark datasets—such as EmotionLines, MELD, IEMOCAP, and CMU-MOSEI—containing audio, visual, and linguistic modalities annotated with emotional labels [2-5]. These resources have demonstrated the importance of contextual and multimodal cues for accurate emotion inference. The MuSe dataset further explored the interdependence between stress and emotion, highlighting the potential of multimodal approaches to capture complex psychological states [6]. Recent models, including attention-based fusion frameworks and the multimodal emotion estimation system MIST, have achieved high performance by integrating multiple modalities such as speech, text, facial expression, and body movement [7,8]. Nonetheless, challenges remain, including data imbalance, cultural bias, and limited model interpretability. In addition, contactless multimodal emotion recognition (CMER) has gained attention as a privacy-conscious alternative that avoids dependence on physiological sensors. Future work should focus on constructing multilingual and culturally diverse datasets, as well as developing ethical and generalizable multimodal emotion recognition models.

The COVID-19 pandemic has rapidly accelerated the adoption of online counseling across various industries, including information technology. However, one major challenge in online settings is that counselors often find it difficult to accurately perceive subtle emotional cues such as facial expressions and vocal tones. Consequently, technologies for automatically analyzing clients' emotions have become increasingly important for supporting counselors' decision-making. Human emotion recognition inherently relies on multiple modalities—voice, facial expressions, and linguistic content—and multimodal emotion recognition has been shown to achieve higher accuracy in AI-based estimation [9]. For such systems, multimodal datasets annotated with emotion labels are essential [10].

Our research group has been developing multimodal emotion recognition models that predict clients' stress levels to assist counselors in their assessments [11-14]. However, few emotion-labeled multimodal datasets focusing on counseling sessions are publicly available. To address this, we constructed a novel dataset of online counseling sessions, including video recordings of interactions between clients (remote workers) and counselors, emotion labels annotated by third-party evaluators, and stress labels derived from questionnaires and counselor assessments. This dataset provides comprehensive data for developing and evaluating stress prediction models. As one of the few multimodal datasets containing both emotion and stress labels in Japanese, it holds significant potential for advancing mental well-being assessment and supporting mental health management among workers.

This section introduces the Tokushima University Online Counseling Dialogue Corpus we constructed, presents the results of emotion estimation based on this corpus, and describes a stress estimation technique applying emotion estimation technology [15].

Dataset

This study aimed to analyze the relationship between stress and emotions in online counseling sessions by collecting data from 50 participants. Counseling sessions were conducted via Zoom, where clients completed an online stress questionnaire prior to the sessions. During the sessions, clients and counselors discussed work-related and daily stress. These sessions were recorded on the counselor's account, capturing both video and audio data.

- **Data Anonymization and Feature Extraction:** Feature extraction and anonymization of data in the TU-OCDC dataset were performed as follows.
  - ***Video Data:*** To protect the privacy of participants, anonymization was performed using the SimSwap library, which replaced facial features such as eyes, nose, and mouth with those of other individuals [16].
  - ***Audio Data:*** Features were extracted using Wav2Vec2.0, a self-supervised learning model, yielding 1024-dimensional feature vectors [17].
  - ***Text Data:*** Speech was transcribed into text using the NueASR Automatic Speech Recognition model, and linguistic features were obtained using OpenAI's text-embedding-3 [18,19].
- **Emotion Annotation:** Emotion annotation employed Russell's circumplex model. Seven annotators participated in recording the client's emotional valence (X-axis) and arousal (Y-axis) every second. The annotation tool, developed using JavaScript and HTML, displayed the counseling video alongside the circumplex model. Annotators could use playback, pause, and rewind functionalities to ensure accurate labeling. This annotation tool is publicly available on GitHub[1].
- **Data Evaluation:** The consistency of annotations was evaluated using Fleiss' kappa coefficient. The results showed a "slight agreement" overall, with higher consistency observed for arousal (Y-axis). In contrast, the consistency for valence (X-axis) was lower. These outcomes were attributed to limited changes in arousal during sessions and the annotators' tendency to position the pointer near the Y-axis center. We also constructed a viewer tool for the corpus as TU-OCDC Viewer. Figure 1 shows an example of displaying a portion of dataset by using TU-OCDC Viewer[2].

Figure 1: Example of Displaying a Portion of a Dataset Using the TU-OCDC Viewer

[1] https://github.com/A2TokushimaUniv/russell_emotion_annotation
[2] https://github.com/A2TokushimaUniv/tuocdcviewer

Statistics of Dataset: Table 1 presents the statistical information of the data included in the TU-OCDC corpus. Emotion labels were determined based on the average coordinates on Russell's circumplex model, calculated from multiple annotators' ratings assigned at one-second intervals. For each labeled coordinate value, the closest predefined sentiment class among 17 was assigned. For each utterance, these labels were converted into occurrence frequency vectors, and the emotion class with the highest frequency was identified and aggregated.

Table 1: Statistics of TU-OCDC

Number of Videos	Num. of Vocab.	Num. of Utterances (sentence/audio files)	Total time of video and audio(second)	Avg. time of video and audio(second)	Total Frames of videos
50	3665	3164	82078	1641.562	2462343
	Emotion	**Frequency**	**Emotion**	**Frequency**	
	Calm	3552	Satisfaction	70	
	Melancholy	239	Oppression	52	
	Worry	194	Easygoing	32	
	Unpleasant	97	Happiness	21	

Multimodal Stress Prediction Model

Takanabe et al. proposed a multimodal stress prediction model that integrates ten types of modality features, including linguistic, acoustic, and visual information, to perform binary stress classification [20]. To address the limited availability of training data, they utilized an auxiliary dataset, the Tokushima University Depressive State Corpus (TU-DEP), which is similar to TU-OCDC, and achieved a maximum accuracy of 80% [21]. Feature importance analysis revealed that high-dimensional acoustic and linguistic features were dominant contributors, while low-dimensional facial, speech-emotion, and language-emotion features also played a significant role in stress prediction. These results demonstrated that employing an auxiliary dataset effectively compensates for data scarcity and that combining multiple modalities produces complementary effects that enhance prediction performance.

Several other studies have also attempted to predict stress from human language, speech, and facial expressions [22-25]. In most of these studies, the degree of stress was determined through self-reports or by third-party evaluations (e.g., counselors or physicians), and the obtained labels were used to train machine learning models.

Given the difficulty of quantitatively measuring stress, alternative approaches have been explored in which datasets are labeled using psychological indices correlated with stress levels. Representative indices include the Patient Health Questionnaire-9 (PHQ-9) for depression diagnosis, and several scales assessing suicidal ideation, such as the Suicide Intent Scale (SIS), Scale for Suicide Ideation (SSI), Beck Scale for Suicide Ideation (BSI), and Suicidal Ideation Questionnaire (SIQ) [26-30]. For anxiety-related disorders, commonly used measures include the Generalized Anxiety Disorder scales (GAD-7 and GAD-2), Hamilton Anxiety Rating Scale (HAM-A), Hospital Anxiety and Depression Scale (HADS), Liebowitz Social Anxiety Scale (LSAS), and the State–Trait Anxiety Inventory (STAI) [31-36]. However, since these instruments rely on self-reported data and are influenced by the subject's psychological state at the time of assessment, they are not suitable as absolute indicators. In this study, we primarily use the self-reported PHQ-9 scores provided in the auxiliary dataset TU-DEP as an indicator to define stress levels [21].

Multimodal Mild Cognitive Impairment Prediction Model

Several studies have been conducted on cognitive function prediction. Most of these approaches can be categorized into two types: conventional task-based assessments that rely on performance scores, and dialogue-based approaches that focus on linguistic and acoustic information during interactive tasks [37-40]. In studies emphasizing speech and language features, typical acoustic features include pitch, formants, and prosody, while linguistic features involve lexical diversity, syntactic complexity, semantic coherence, and speech quantity. These extracted features are often used as input to machine learning models such as Support Vector Machines (SVM), Random Forests, and neural networks to estimate cognitive function. Features derived from speech and language have been shown to be effective for diagnosing dementia and mild cognitive impairment (MCI), with some studies reporting classification accuracies exceeding 80%.

However, most existing datasets target binary classification between dementia patients and healthy controls, and there is a lack of sufficient data representing early-stage cognitive decline, such as MCI [41-42]. To address this limitation, it is beneficial to simulate cognitive decline situations among healthy individuals through task performance, acquire multimodal behavioral data during such conditions, and analyze their characteristic patterns.

In this study, we propose a novel cognitive assessment approach that enables the acquisition of multimodal information in addition to conventional cognitive tests. The proposed system incorporates an AI agent that provides verbal encouragement and advice through spoken dialogue while participants sequentially perform multiple tasks. During these interactions, multimodal signals—including speech, facial expressions, utterance content, and gaze behavior—are captured. By associating these multimodal features with cognitive performance metrics such as task completion rate, a multimodal cognitive function prediction model is trained.
The overall framework of the proposed approach is illustrated in Figure 2.

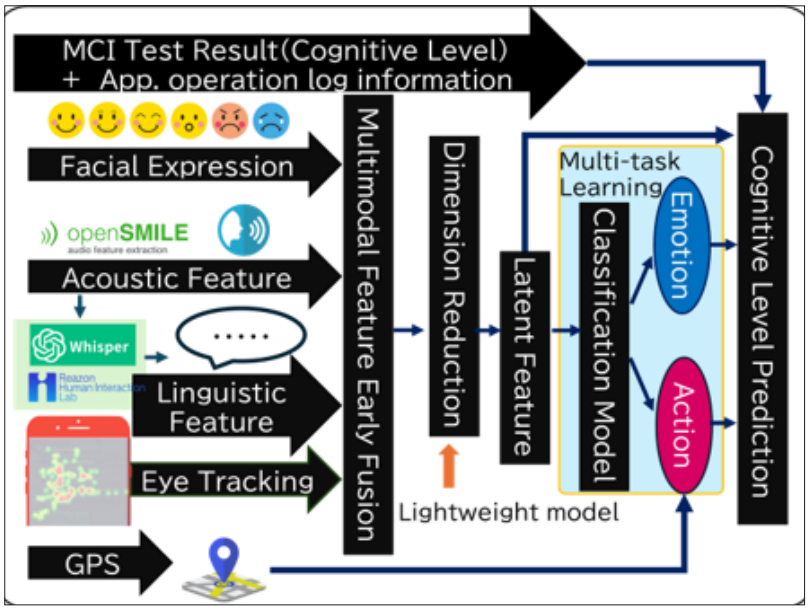


Figure 2: Training Flow of Multimodal Mild Cognitive Impairment Prediction Model

In this study, a mild cognitive impairment (MCI) testing task is conducted on portable devices such as smartphones, where multimodal data—including facial expressions, acoustic features, linguistic features, and eye movement—are collected during task performance. These multimodal features are integrated and subjected to dimensionality reduction to extract latent representations. The proposed model utilizes these latent features as inputs and employs multitask learning to predict target variables, including MCI test scores, behavioral information obtained from GPS and 3D motion sensors, and emotional indicators. The goal of this approach is to learn the interrelationships among behavior, emotion, and cognitive function from multimodal feature representations.

To acquire training data for implementing this technology, we are currently developing a web-based application accessible from edge devices such as smartphones. Since certain data, such as gaze and heart rate, require high-precision sensors, we first plan to develop a web-based cognitive assessment application for personal computers, enabling the collection of diverse multimodal information.

Experiment

In this section, we describe experiments conducted to construct a model for estimating emotional information that can be effectively utilized as features for cognitive function prediction, based on multimodal features extracted from the TU-OCDC dataset. In addition to TU-OCDC, the CMU-MOSI dataset and the TU-DEP dataset are employed as auxiliary data sources for training [43].

Multimodal features are extracted from three modalities: video, audio, and text. The models used for feature extraction in each modality are summarized in Table 2. For video features, we use VideoMAE [44]. For acoustic features, Kushinada-large, a HuBERT-based speech representation model, is utilized. For linguistic features, LUKE, a contextualized embedding model, is employed [45-46].

The emotion vector represents the outputs of pre-trained emotion recognition models applied to each modality, converted into a unified vector form. For facial expression analysis, we employ Py-Feat, which not only estimates facial expressions but also extracts head pose information (Pitch, Roll, Yaw) and Facial Action Units (AUs) representing facial muscle movements [47]. These additional features are treated as Other features. For audio (Kushinada-large, Kushinada-er) and text (LUKE, LUKE-emotion, LUKE-sentiment), Japanese language–specific models are used [48].

Table 2: Summary of Feature Extraction Models Used for Each Modality

Modality	Model Name	Dimension	Emotion vector	Other feature
Video	VideoMAE [44]	768	Py-Feat(7dim) [47]	Face direction(3dim) AU (20dim)
Audio	HuBERT(Kushinada-large) [45]	1024	Kushinada er(4dim) [48]	
Text	LUKE [46]	768	LUKE-emotion(8dim) [49]	
			LUKE-Sentiment(1dim) [50]	

In this study, we hypothesize that simply integrating multimodal information is insufficient for effectively learning emotion estimation models from low-resource modalities. To address this issue, we develop a novel model that follows a three-stage training process, as illustrated in Figure 3.

First, we pretrain Model A on CMU-MOSI, a large-scale multimodal dataset annotated with emotion polarity and intensity labels, to predict both the polarity and strength of emotions. Next, we fine-tune Model A using a depression dataset, where PHQ-9 scores are binarized by a predefined threshold, thereby constructing a binary classification model (Model B). Finally, based on Model B, we train Model C using the TU-OCDC dataset to predict the two-dimensional coordinates of emotions defined by Russell's circumplex model.

Through this stepwise training procedure, our framework enables the effective extraction of multimodal representations that are informative for both emotion estimation and mental disorder detection.

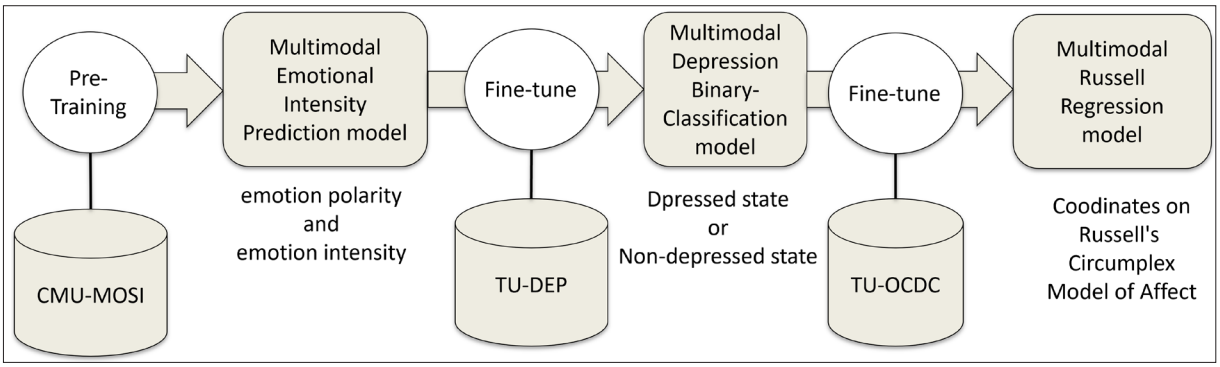


Figure 3: Overview of the Proposed Three-Stage Training Process

The proposed model consists of three stages:

- **Retraining Model A** on the CMU-MOSI dataset for emotion polarity and intensity prediction
- **Fine-tuning Model B** on a depression dataset to construct Model B, a binary classifier based on PHQ-9 scores; and
- **Training Model C** on the TU-OCDC dataset to predict emotional coordinates on Russell's circumplex model.

This stepwise process enables gradual knowledge transfer from emotion estimation to mental state prediction.

In previous studies, when dealing with low-resource settings, several approaches have been proposed to improve performance, such as data augmentation, as well as early and late fusion of multimodal features. Although some studies have explored the effective use of multiple resources, achieving significant performance improvement remains challenging. This difficulty arises from factors such as differences in languages and label structures across datasets, as well as the limited availability of multimodal corpora. In some cases, using heterogeneous datasets as complementary sources can even degrade model performance.

In this study, we utilize CMU-MOSI, a relatively large-scale multimodal dataset annotated with emotion labels relevant to mental health, to pretrain foundational multimodal representations [50]. The CMU-MOSI dataset contains English speech and language data, whereas TU-OCDC—our target dataset—comprises audiovisual and linguistic data from Japanese speakers. While acoustic features are less affected by language differences, linguistic features (text data) are inherently language-dependent. Although multilingual models could be used, they often fail to capture fine-grained semantic nuances specific to a single language. Therefore, we translate English text data into Japanese during pretraining and employ a Japanese-specific language model for feature extraction, allowing for more accurate semantic representation.

In addition, we train a binary classification model using PHQ-9 scores based on the TU-DEP dataset, which contains Japanese participants' data collected in collaboration with the Tokushima University Faculty of Medicine. This enables the model to learn Japanese-specific modality characteristics and integrate multimodal features related to both emotion and mental health.

Finally, we train and evaluate the model through cross-validation to predict emotional coordinates on Russell's circumplex model, which represent the direction and intensity of emotion. This model is expected to serve as a foundation for subsequent stress estimation systems.

The four models evaluated in this study are summarized as follows:

- ***Model A*:** Pretrained on CMU-MOSI (emotion polarity/intensity), fine-tuned on TU-DEP (depression classification), and further fine-tuned on TU-OCDC (emotion coordinate prediction).
- ***Model B*:** Pretrained on CMU-MOSI and fine-tuned directly on TU-OCDC.

- ***Model C*:** Pretrained on TU-DEP and fine-tuned on TU-OCDC.
- ***Model D*:** Trained solely on TU-OCDC without any pretraining.

During the pretraining of CMU-MOSI, the dataset was partitioned to ensure that participant IDs did not overlap between the training and test sets, and the best-performing model was selected. Similarly, for TU-DEP, both pretraining and fine-tuning were conducted using participant-level data partitioning, followed by cross-validation to determine the best model.

For the TU-OCDC dataset, the data were split by participant, and a 50-fold cross-validation was performed. In addition, to examine the contribution of each modality, an ablation study was conducted in which the features of all but one modality (Visual, Audio, or Text) were masked by replacing them with zero vectors. This analysis was used to evaluate the relative importance of each modality.

Figure 4 illustrates a simplified representation of the architecture of the emotion estimation model for two-dimensional coordinate values in the Russell circle model defined in this study.

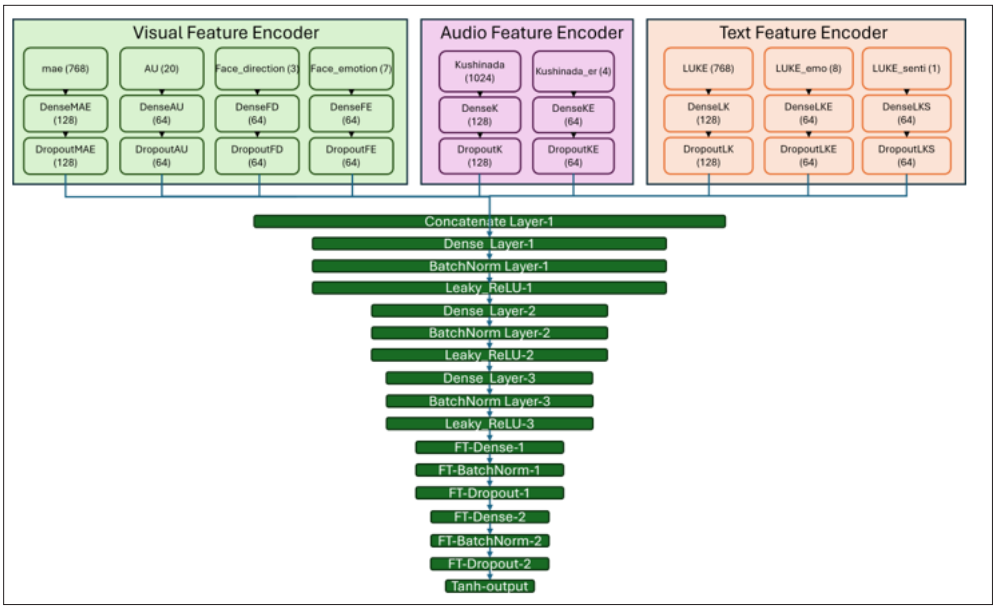


Figure 4: Network Architecture for Russell's Two-dimensional Emotion Prediction

Results

This section summarizes the results of the evaluation experiments described in the previous section.

Figures 5–8 graphically represent the MSE, MAE, and RMSE for each model across modalities.

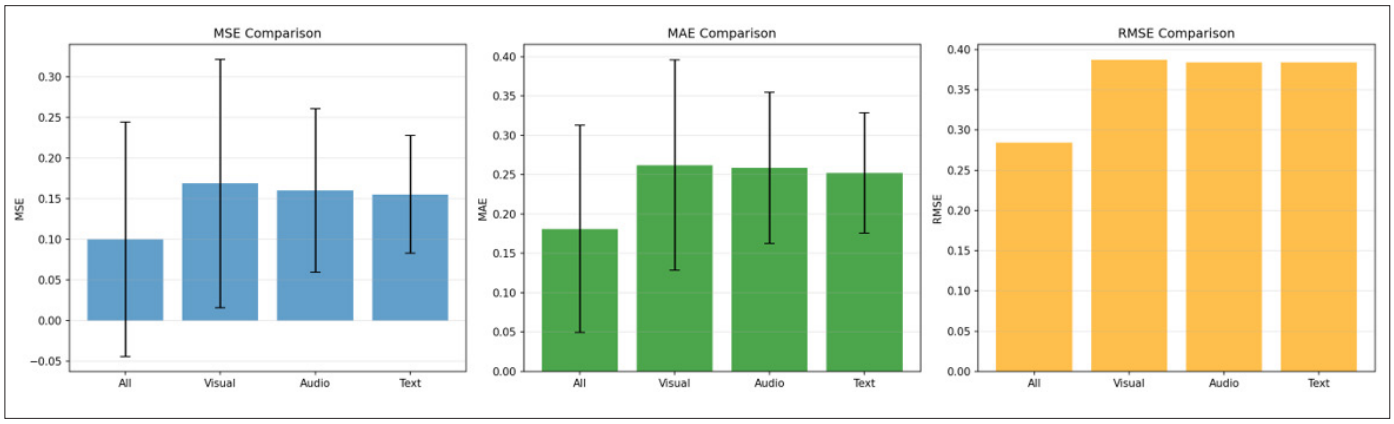


Figure 5: MSE, MAE, RMSE; Model-A (CMU-MOSI+TU-DEP+TU-OCDC)

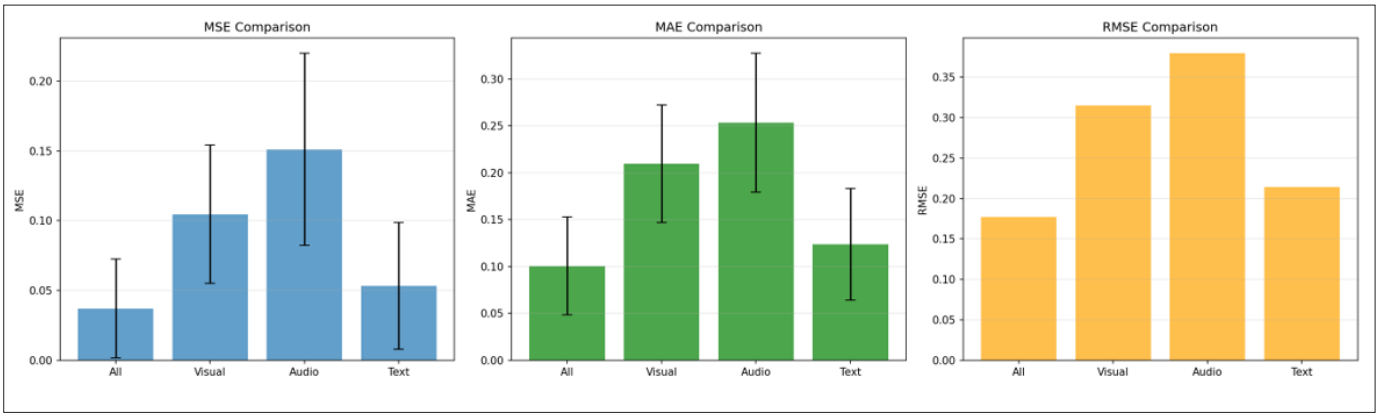


Figure 6: MSE, MAE, RMSE; Model-B (CMU-MOSI+TU-OCDC)

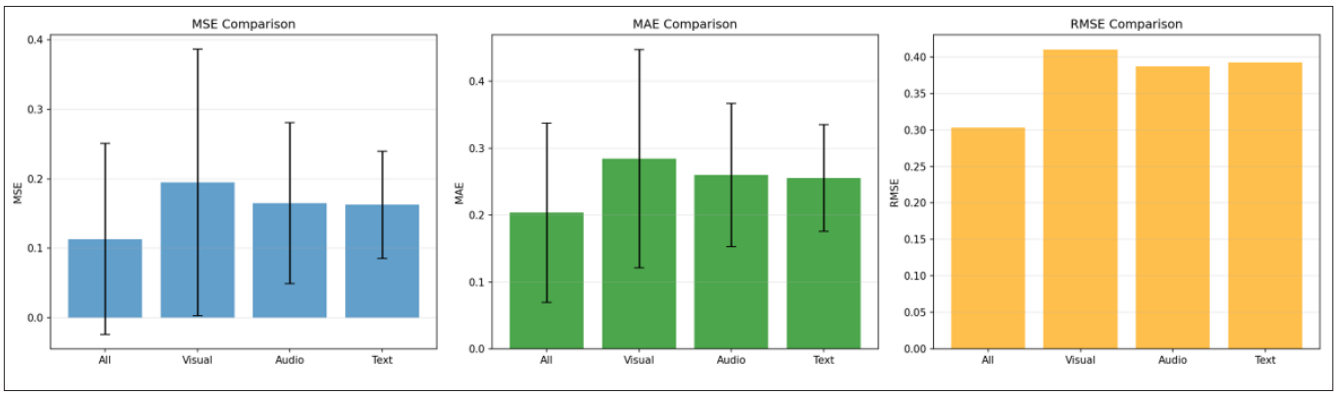


Figure 7: MSE, MAE, RMSE; Model-C (TU-DEP+TU-OCDC)

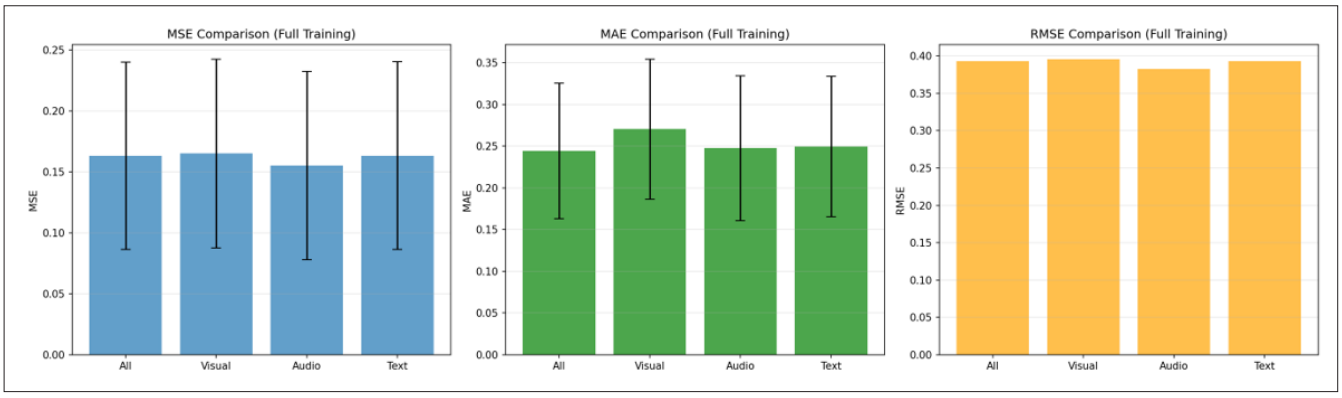
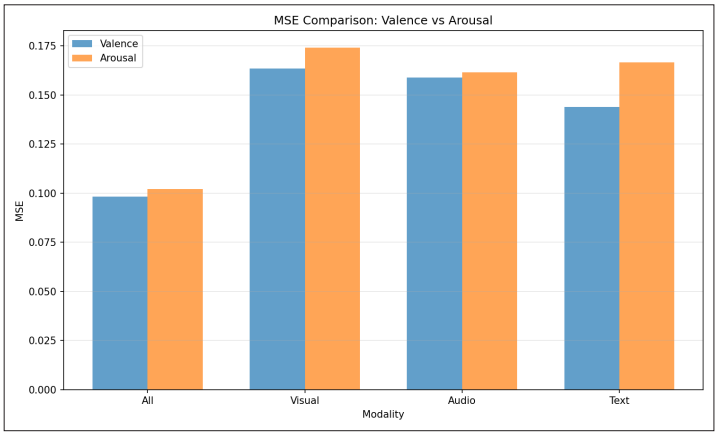


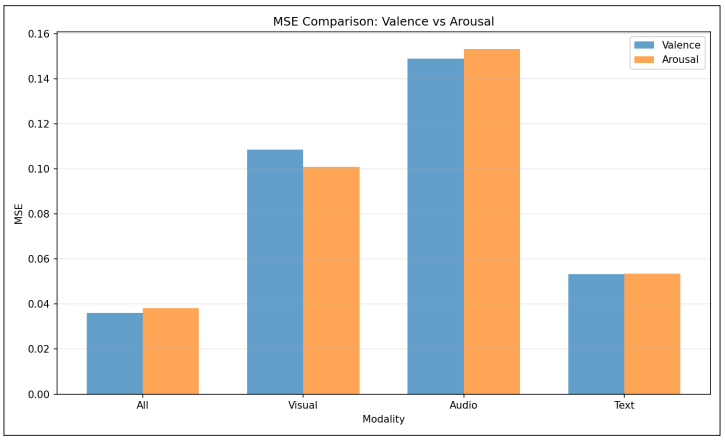
Figure 8: MSE, MAE, RMSE; Model-D (TU-OCDC)

These results indicate that high performance cannot be achieved using a single modality alone, and that performance can be improved by utilizing all three modalities: Visual, Audio, and Text. Furthermore, Model-A, which underwent two-stage fine-tuning, demonstrated higher performance than Model-C and Model-D, but it could not surpass Model-B, which was fine-tuned in a single stage using a pre-trained model based on CMU-MOSI.

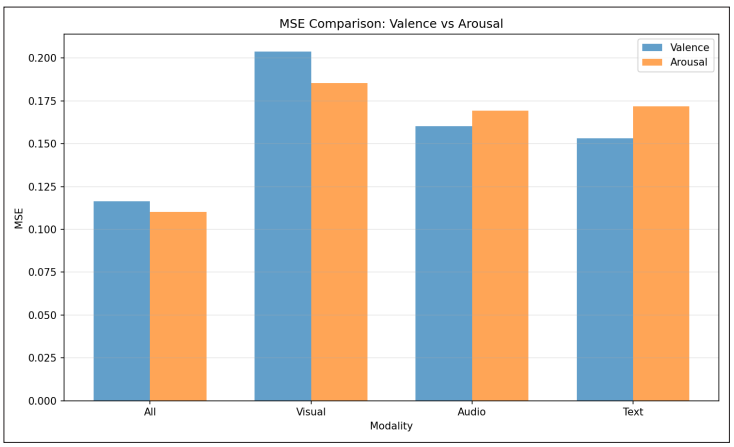
Figure 9 shows the comparison results for Valence and Arousal.



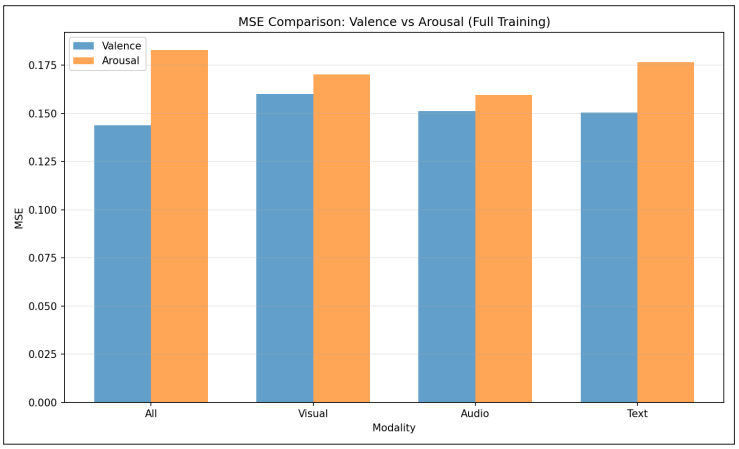
(a) Model-A



(b) Model-B



(c) Model-C



(d) Model-D

Figure 9: Comparison between MSE of Valence and Arousal

These results indicate that while no significant difference was observed between the Valence and Arousal dimensions, overall, Valence demonstrated slightly higher accuracy.

Figure 10 compares the performance of Model-D (Full-Training) without pre-training against Model-A (Fine-Tuning). The bottom row shows the difference in performance between Model-A and Model-D for each modality. Smaller values indicate a greater improvement in performance due to fine-tuning. These results indicate that while the effect of fine-tuning is not particularly noticeable when evaluated by modality, combining the three modalities significantly enhances the effectiveness of fine-tuning.

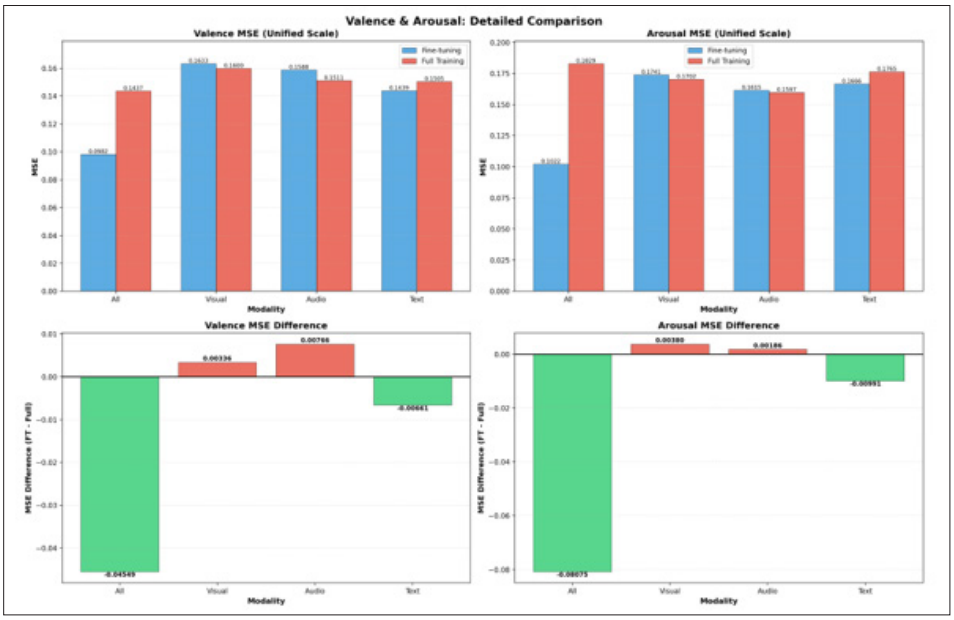


Figure 10: Comparison between Performance of the Fine-Tuned Model and Full-Training Model

Discussion

Evaluation experiments demonstrated that performance improvements could be achieved by fine-tuning models pre-trained on large-scale datasets for coordinate estimation in the Russell Emotion Circle Model using multimodal information. However, performing two-stage fine-tuning resulted in lower performance compared to single-stage fine-tuning. This degradation is likely primarily caused by data bias within the first-stage TU-DEP corpus. Indeed, in the PHQ-9 binary classification task, we observed a problem where models tended to predict almost exclusively one label, suggesting a significant impact from data imbalance. While weight adjustment was applied to address class imbalance, data from subjects with high PHQ-9 scores tended to exhibit greater feature variability compared to other subjects. This is likely because subjects with extremely high PHQ-9 scores and severe depressive tendencies have voice, image, and language features distinct from other subjects, making them prone to becoming outliers. Selecting data based on severity level is expected to become important in the future.

The distribution of emotion labels in the TU-OCDC dataset used in this experiment exhibits significant bias. Therefore, in addition to fine-tuning, data augmentation is also crucial. At present, we have no choice but to employ methods such as adding noise to the features. The reason for this is that video generation AI is still in its developmental stages. Adding audio would further degrade accuracy and compromise data quality. Generating data modality-specific also raises concerns about accuracy loss due to inconsistencies between modalities.

Conclusion

This paper describes our research on multimodal information analysis, focusing on stress estimation, cognitive function prediction, and emotion estimation—our primary research themes. We present practical examples, including evaluation experiments using our proprietary multimodal dataset. Specifically, we addressed the challenge of insufficient data for target tasks by leveraging multimodal prediction models pre-trained on labeled multimodal training datasets collected for different purposes.

As a future direction, we aim to build high-precision models pre-trained on various modality features to predict human stress levels, which are considered highly correlated with cognitive decline. Achieving this approach requires high-quality, large-scale datasets. However, datasets providing all three modalities—audio, image, and text—in a single set remain scarce. To address this, we are experimenting with a method that combines and fuses representations from the intermediate layers of pre-trained emotion estimation models (audio emotion analysis, linguistic emotion analysis, facial expression analysis) across single modalities. Through fine-tuning and transfer learning, we aim to construct a high-accuracy cognitive function prediction model.

Acknowledgements

This work was supported by Japan's National Research and Development Agency New Energy and Industrial Technology Development Organization (NEDO)(JPNP20004) and The General Insurance Association of Japan (GIAJ).

References

1. TMAK. Available at: https://www-a2.is.tokushima-u.ac.jp/.
2. Calzolari N, Choukri K, Cieri C, Declerck T, Goggi S, et al. (2018) EmotionLines: An Emotion Corpus of Multi-Party Conversations. Proc Eleventh Int Conf Language Resources and Evaluation. Available at: https://aclanthology.org/L18-1252/.
3. Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R (2019) MELD: Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Proc 57th Annual Meeting Assoc Computational Linguistics 527-536.
4. Zadeh A, Liang PP, Poria S, Vij P, Cambria E, Morency L-P (2018) Multi-attention Recurrent Network for Human Communication Comprehension. Proc Thirty-Second AAAI Conf Artificial Intelligence 5642-5649.
5. Zadeh AB, Liang PP, Poria S, Cambria E, Morency L-P (2018) Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. Proc 56th Annual Meeting Assoc Computational Linguistics 1: 2236-2246.
6. Jaiswal M, Bara CP, Luo Y, Burzo M, Mihalcea R, Provost EM (2020) MuSE: a Multimodal Dataset of Stressed Emotion. Proc Twelfth Language Resources and Evaluation Conf 1499-1510.
7. Mamieva D, Abdusalomov AB, Kutlimuratov A, Muminov B, Whangbo TK (2023) Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features. Sensors 23: 5475.
8. Boitel E, Mohasseb A, Haig E (2025) MIST: Multimodal Emotion Recognition Using DeBERTa for Text, Semi-CNN for Speech, ResNet-50 for Facial, and 3D-CNN for Motion Analysis. Expert Syst Appl 270: 126236.
9. Stappen L, Baird A, Christ L, Schumann L, Sertolli B, et al. (2021) The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress. Available at:10.48550/arXiv.2104.07123.
10. Schmidt P, Reiss A, Duerichen R, Marberger C, Laerhoven KV (2018) Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection. Proc 20th ACM Int Conf Multimodal Interaction 400-408.
11. Takanabe T, Matsumoto K, Kiuchi K, Kang X, Nishimura R, Sasayama M (2024) Construction and Evaluation of a Multimodal Counseling Dataset for Emotion Analysis. Proc IPSJ National Conf 4: 83-84.
12. Takanabe T, Kashihara K, Matsumoto K, Kiuchi K, Kang X, Nishimura R, et al. (2024) Multimodal Emotion Recognition and Dataset Construction in Online Counseling. Proc 38th Pacific Asia Conf Language, Information and Computation 1-9.
13. Kashihara K, Takanabe T, Kiuchi K, Umehara H, Irizawa K, et al. (2024) Constructing Multimodal Counseling Dataset for Depressive State and Feature Analysis. Proc 8th Int Conf Natural Language Processing and Information Retrieval.
14. Tan Y, Zhou J, Matsumoto K, Kang X, Yoshida M (2025) Proposal of a Multimodal Emotion Recognition Model Based on Fusion of Audio and Text. Proc 39th Annual Conf Japanese Society for Artificial Intelligence. Available at: https://www.jstage.jst.go.jp/article/pjsai/JSAI2025/0/JSAI2025_3G5GS605/_article/-char/en.
15. TU-OCDC. Tokushima University Online Counseling Corpus. Available at: https://www.nii.ac.jp/dsc/idr/rdata/TU-OCDC/.
16. SimSwap. Available at: https://github.com/neuralchen/SimSwap.
17. Baevski A, Zhou H, Mohamed A, Auli M (2020) wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Adv Neural Information Processing Systems. Available at: arXiv:2006.11477.
18. NueASR. Available at: inna/nue-asr. https://huggingface.co/rinna/nue-asr.
19. text-embedding-3. Available at: https://huggingface.co/datasets/Qdrant/dbpedia-entities-openai3-text-embedding-3-large-3072-1M.
20. Takanabe T, Matsumoto K, Kiuchi K, Kashihara K, Umehara H, et al. (2025) Construction of a Stress Estimation Model Using Multimodal Features. 24th Forum on Information Science and Technology.
21. Kashihara K, Takanabe T, Kiuchi K, Umehara H, Irizawa K, et al. (2024) Construction of Multimodal Dataset for Early Depression Detection and Performance Evaluation of Depression Detection Model. ResearchSquare. Available at: https://www.researchsquare.com/article/rs-6344200/v1.
22. Kerz E, Zanwar S, Qiao Y, Wiechmann D (2023) Toward Explainable AI (XAI) for Mental Health Detection Based on Language Behavior. Front Psychiatry 14: 1-20.
23. Kappen M, Hoorelbeke K, Madhu N, Demuynck K, Vanderhasselt M-A (2022) Speech as an Indicator for Psychosocial Stress: A Network Analytic Approach. Behav Res Methods 54: 910-921.
24. Zhang H, Feng L, Li N, Jin Z, Cao L (2020) Video-Based Stress Detection through Deep Learning. Sensors 20: 5552.
25. Ciharova M, Amarti K, van Breda W, Peng X, Lorente-Català R, et al. (2024) Use of Machine Learning Algorithms Based on Text, Audio, and Video Data in the Prediction of Anxiety and PTSD: A Systematic Review. Biol Psychiatry 96: 519-531.
26. Kroenke K, Spitzer RL, Williams JB (2001) The PHQ-9: Validity of a Brief Depression Severity Measure. J Gen Intern Med 16: 606-613.
27. Beck A, Steer R (1989) Clinical Predictors of Eventual Suicide: A 5- to 10-year Prospective Study. J Affect Disord

17: 203-209.
28. Beck A, Brown G, Steer R (1997) Psychometric Characteristics of the Scale for Suicide Ideation with Psychiatric Outpatients. Behav Res Ther 35: 1039-1046.
29. Beck A, Ward C, Mendelson M, Mock J, Erbaugh J (1961) An Inventory for Measuring Depression. Arch Gen Psychiatry 4: 561-571.
30. Reynolds WM. Available at: Suicidal Ideation Questionnaire (SIQ). https://teams.semel.ucla.edu/sites/default/files/pdf/SuicidalIdeationQuestionnaire(SIQ)(Child).PDF.
31. Spitzer RL, Kroenke K, Williams JB, Löwe B (2006) A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. Arch Intern Med 166: 1092-1097.
32. Kroenke K, Spitzer R, Williams J, Monahan P (2007) Anxiety Disorders in Primary Care: Prevalence, Impairment, Comorbidity, and Detection. Ann Intern Med 146: 317-325.
33. Hamilton M (1960) A Rating Scale for Depression. J Neurol Neurosurg Psychiatry 23: 56-62.
34. Zigmond A, Snaith R (1983) The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 67: 361-370.
35. Liebowitz MR (1987) Social Phobia. Modern Probl Pharmacopsychiatry 22: 141-173.
36. Spielberger CD (1971) The State-Trait Anxiety Inventory. Interamerican J Psychol 5: 3-4.
37. Voleti R, Liss JM, Berisha V (2020) A Review of Automated Speech and Language Features for Assessment of Cognitive and Thought Disorders. IEEE J Sel Top Signal Process 14: 282-298.
38. Gosztolya G, Vincze V, Tóth L, Pákáski M, Kálmán J, et al. (2019) Identifying Mild Cognitive Impairment and Alzheimer's Disease Based on Spontaneous Speech. Comput Speech Lang 53: 181-197.
39. Luz S, De La Fuente Garcia S, Haider F, Fromm D, MacWhinney B, et al. (2024) Connected Speech-Based Cognitive Assessment in Chinese and English. InterSpeech. Available at: arXiv:2406.10272.
40. Agbavor F, Liang H (2024) Multilingual Prediction of Cognitive Impairment with Large Language Models and Speech Analysis. Brain Sci 14: 1292.
41. Shibata D, Ito K, Wakamiya S, Aramaki E (2019) Detecting Early Stage Dementia Based on Natural Language Processing. Trans Japanese Society for Artificial Intelligence 34: B-J11_1–9.
42. Becker JT, Boller F, Lopez OL, Saxton J, McGonigle KL (1994) The Natural History of Alzheimer's Disease: Accuracy of Diagnosis. Arch Neurol 51: 585-594.
43. Zadeh A, Zellers R, Pincus E, Morency L-P (2016) MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. Available at: arXiv:1606.06259.
44. Tong Z, Song Y, Wang J, Wang L (2022) VideoMAE: Masked Autoencoders for Self-Supervised Video Pre-Training. Proc 36th Adv Neural Information Processing Systems 10078-10093.
45. Kushinada-large. Available at: https://huggingface.co/imprt/kushinada-hubert-large.
46. LUKE. Available at: https://huggingface.co/studio-ousia/luke-japanese-base.
47. Py-Feat. Available at: https://py-feat.org/.
48. Kushinada-er. Available at: https://huggingface.co/imprt/kushinada-hubert-base-jtes-er.
49. LUKE-emotion. Available at: https://huggingface.co/Mizuiro-sakura/luke-japanese-large-sentiment-analysis-wrime.
50. LUKE-sentiment. Available at: https://huggingface.co/Mizuiro-sakura/luke-japanese-base-marcja.