



M3L: A Multi-Modal and Multi-Lingual Depression Detection Framework

Jiajun You¹, Shuai Wang^{2,1,3,*}, Xun Gong⁴, Xiang Wan³

¹School of Data Science, The Chinese University of Hong Kong, Shenzhen, China

²School of Intelligence Science and Technology, Nanjing University, Suzhou, China

³Shenzhen Research Institute of Big Data, Shenzhen, China

⁴Auditory Cognition and Computational Acoustics Lab, Shanghai Jiao Tong University, Shanghai, China,

122090684@link.cuhk.edu.cn, shuaiwang@nju.edu.cn, gongxun@sjtu.edu.cn, wanxiang@sribd.cn

Abstract

Early diagnosis are essential to reduce costs and improve treatment efficiency. Recently, automatic depression detection (ADD) based on audio and textual features from participant interviews has emerged as a promising approach, attracting significant attention. However, existing models are constrained to monolingual depression datasets, with limited exploration of multi-lingual scenarios. To investigate the effectiveness of multi-lingual data for the ADD task and its transferability in low-resource scenarios, in this paper, we propose a Multi-Modal Multi-Lingual (M3L) depression detection framework and an effective language adaptive fine-tuning (LAFT) to further boost the performance on the target language. M3L utilizes the pretrained speech model Whisper and the text model XLM-RoBERTa to enhance the encoding of multilingual information. Evaluations on the DAIC-WOZ (English) and EATD (Chinese) datasets demonstrate that M3L effectively integrates multi-lingual and multi-modal information, while the proposed LAFT consistently boosts performance across both datasets.

Index Terms: Depression detection, Cross-language, LoRA, Whisper, RoBERTa

1. Introduction

Depression is a critical mental health issue affecting approximately 322 million people globally, with nearly half located in Southeast Asia and the Western Pacific, including China [1]. In severe cases, it can lead to suicide [2, 3]. Current diagnostic methods primarily rely on self-reported questionnaires and professional evaluations, but these face challenges due to depression's effects on self-perception and cognitive awareness [4]. The subjectivity of these questionnaires can lead to variability in diagnoses, impacting treatment outcomes [5]. The lack of objective diagnostic tests necessitates time-consuming clinician screenings, underscoring the urgent need for more accurate early diagnostic techniques [6, 7].

Recently, artificial intelligence (AI) has made strides in ADD, leveraging various physiological and behavioural data modalities such as facial expressions, voice recordings, textual data, and EEG signals [8, 9]. Among these, audio and text data stand out due to their accessibility and rich representation of depressive symptoms. For instance, speech characteristics associated with depression include reduced pauses, increased speech errors [10], and a restricted pitch range [11]. Similarly, linguistic patterns in text, such as frequent use of personal pronouns and negative emotional words, have been linked to depression severity [12]. These observations have driven the development

of numerous audio-based and text-based methods for depression detection.

Multimodal approaches have demonstrated significant potential in depression detection, offering a more comprehensive perspective by integrating complementary information from various modalities, such as audio, text, and visual data. Recent studies have explored diverse techniques to improve multimodal fusion and model robustness. For example, Tuka et al. [13] utilized LSTMs for analyzing audio and text data, while Seneviratne et al. [14] developed TV-based ACFs to enhance classification performance. Advanced fusion frameworks, such as attention-based GRU/BiLSTM architectures [15] and multi-level attention mechanisms [16, 17], have further improved inter- and intra-modality correlations. Other notable efforts include speaker identity decoupling for robustness [18], time-aware attention networks [19], and large language model integration [20] to analyze multimodal data effectively.

Multilingual learning has gained attention as a promising approach for depression detection, addressing language-specific limitations and enhancing model generalizability across diverse populations. Liu et al. [21] proposed a CNN-based system with Chinese and English speech data, showcasing cross-language generalization and language-independent diagnostics. Cummins et al. [22] highlighted the robustness of multilingual depression diagnosis using a clinical speech dataset, identifying consistent markers, such as speech rate and intensity across languages. Malik et al. [23] developed a multilingual hope speech detection framework for English and Russian using a Russian YouTube comment corpus. Kiss et al. [24] presented a mono- and multi-lingual system predicting depression severity across German, Hungarian, and Italian. Demiroglu et al. [25] showed similarities in how depression manifests across languages, emphasizing the value of multilingual datasets. Pre-trained models like Whisper [26] and XLM-RoBERTa [27] enable efficient multilingual feature extraction, paving the way for cross-lingual depression detection frameworks.

In this study, we proposed a Multi-Modal and Multi-Lingual (M3L) automatic depression detection framework. The key contributions of this research can be summarized as follows:

- M3L leverages advanced multilingual pretrained models for both speech and text, including Whisper and XLM-RoBERTa, thereby inherently accommodating multilingual contexts.
- We developed different language-specific finetuning strategies that further improve depression detection performance for the target language.
- Experiments on Chinese and English datasets validated the effectiveness of the proposed M3L framework.

*Corresponding author

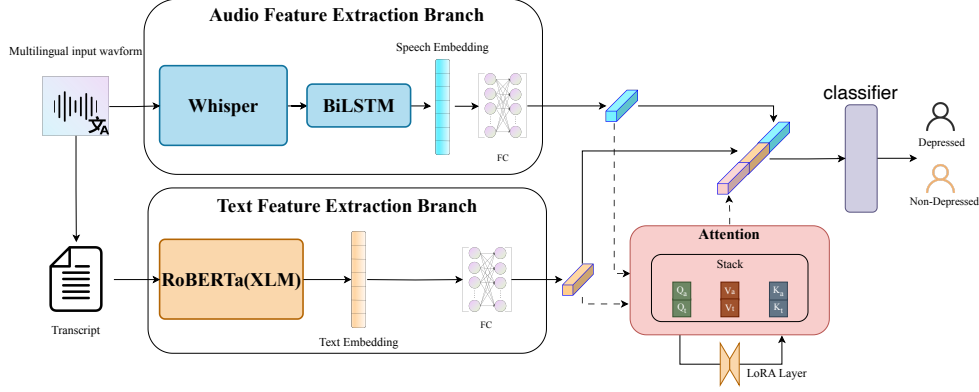


Figure 1: The overall structure of the proposed model.

2. Methods

2.1. Model Architecture

The proposed network framework is an adapted from CAMFM[28], depicted in Fig. 1, consists of two branches: an audio branch and a text branch. The audio branch processes speech recordings using the Whisper model to extract acoustic embeddings. The text branch generates sentence-level embeddings using a pre-trained XLM-RoBERTa model. At the fusion stage, we compared different strategies to combine audio and text features into a unified representation. This fused representation is then passed through an FC layer for the final classification.

In the feature extraction stage, we use the pre-trained Whisper and XLM-RoBERTa models, diverging from traditional methods like LLDs, MFCCs, or wav2vec embeddings. This approach simplifies the model structure and improves scalability to multilingual datasets.

2.1.1. Feature Extraction Branches

In the audio branch, we use the Whisper-large-V3 encoder [26] to extract robust acoustic embeddings from the raw audio signal $X_{\text{audio}} \in \mathbb{R}^T$, yielding the latent representation:

$$Z_{\text{audio}} = \mathcal{F}_{\text{FC}}(\mathcal{F}_{\text{BiLSTM}}(\mathcal{F}_{\text{whisper}}(X_{\text{audio}}))). \quad (1)$$

For the text branch, we employ the XLM-RoBERTa encoder [27] to capture semantic and syntactic features from the text sequence $X_{\text{text}} \in \mathbb{R}^L$, resulting in the latent representation:

$$Z_{\text{text}} = \mathcal{F}_{\text{FC}}(\mathcal{F}_{\text{XLM-RoBERTa}}(X_{\text{text}})). \quad (2)$$

Both branches operate simultaneously, extracting complementary features from audio and text.

2.1.2. Fusion Method

After obtaining the latent representations from the audio and text branches, we compare three fusion strategies to integrate the information from both modalities: the direct concatenation fusion method, the channel attention-based fusion method, and a hybrid approach combining both methods. The detailed structures of these fusion strategies are illustrated in Fig. 2.

Direct Concatenation: The acoustic latent representation Z_{audio} and Z_{text} are directly concatenated, obtaining Z_{concat} as

$$Z_{\text{concat}} = \text{cat}(Z_{\text{audio}}, Z_{\text{text}}). \quad (3)$$

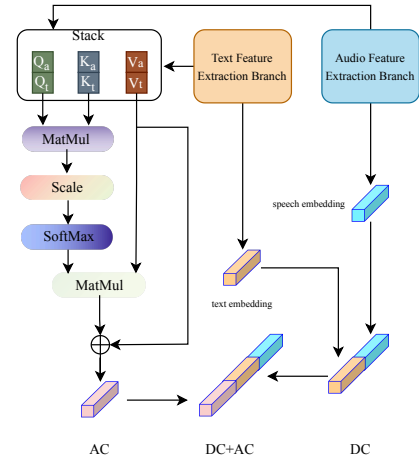


Figure 2: Three feature fusion methods flowchart.

Channel Attention: The channel attention mechanism, originally developed to capture weight relationships among image channels, has been effectively adapted for multimodal fusion tasks. We follow the method in [28] to integrate audio and text features for depression detection:

$$Q = \text{stack}(Q_{\text{audio}}, Q_{\text{text}}), \quad (4)$$

$$K = \text{stack}(K_{\text{audio}}, K_{\text{text}}), \quad (5)$$

$$V = \text{stack}(V_{\text{audio}}, V_{\text{text}}). \quad (6)$$

In this context, the function $\text{stack}(\cdot)$ represents a vertical stacking operation on vectors. The matrices Q , K and V are formulated by vertically stacking the corresponding matrices from the single-modality. Then, we apply the attention calculation formula to obtain Z_{att} :

$$Z_{\text{fusion}} = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V + V, \quad (7)$$

where d_k denotes the dimension of Q/K .

Hybrid Approach: In this approach, we concatenate Z_{concat} and Z_{fusion} from the previous two fusion approach,

$$Z_{\text{hybrid}} = \text{cat}(Z_{\text{concat}}, Z_{\text{fusion}}) \quad (8)$$

Note that Z_{hybrid} corresponds to the fusion method proposed in [28].

2.2. Loss functions

In addition to the standard cross-entropy loss, we apply an orthogonal loss as proposed in [29] to promote complementarity between audio and text modalities. We compute orthogonal losses among Z_{audio} , Z_{text} , and Z_{fusion} , encouraging independence between these representations. This ensures that each modality captures distinct aspects of depression detection.

$$L_{\text{diff}} = \|Z_{\text{audio}}^{\top} Z_{\text{text}}\|_F^2 + \|Z_{\text{fusion}}^{\top} (Z_{\text{audio}} \oplus Z_{\text{text}})\|_F^2 \quad (9)$$

The squared Frobenius norm, referred to as $\|\cdot\|_F^2$, calculates the sum of the squared elements in a matrix. The operator \oplus denotes the horizontal concatenation operation, merging features into a single vector. The overall loss function optimized by the feature fusion module is then defined as

$$L_{\text{fusion}} = L_{\text{ce}} + \alpha L_{\text{diff}}, \quad (10)$$

where α is a hyperparameter controlling the relative contribution of the difference loss. Notably, this loss formulation is the same as in [28], as both approaches employ the same feature fusion method.

3. Training Strategy

Our training process is divided into two phases. The first phase involves training a foundation model using the multilingual data. The second phase focuses on constructing language-specific models through various finetuning strategies to further enhance performance.

3.1. Multi-lingual Pre-training

To increase the effective data volume for training, we combined Chinese and English data to create a multilingual dataset during the multilingual pretraining phase. We utilized the previously mentioned model structure based on Whisper and XLM-RoBERTa, as shown in Fig. 1, for model training.

Table 1: Results for different models on the EATD dataset.

Model	Fusion	F1-Score	Recall	Precision
Whisper (Speech)	-	0.71	0.90	0.62
XLM (Text)	-	0.78	0.88	0.70
Multi-modal (ours)	DC	0.80	0.90	0.71
	CA	0.81	0.86	0.77
	CA+DC	0.83	0.86	0.80
multimodal LSTM[13]	-	0.57	0.67	0.49
GRU/BiLSTM[15]	-	0.57	0.67	0.49
TAMFN[19]	-	0.75	0.85	0.69
CAMFM[28]	-	0.77	0.84	0.73

3.2. Target Language Fine-tuning

After training the model on a multilingual dataset, we fine-tune it on the target dataset to improve its performance in depression diagnosis. Two fine-tuning strategies are explored: APFT (All Parameters Fine-Tuning) and LAFT (LoRA Adapter Fine-Tuning).

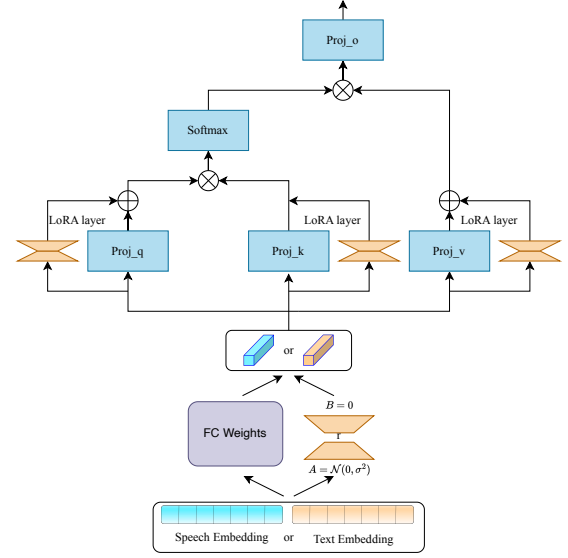


Figure 3: LoRA Parameter Tuning Architecture.

In APFT, all parameters of the multilingual model are updated during the fine-tuning process. Conversely, LAFT utilizes the LoRA fine-tuning approach [30], where only trainable low-rank factorization matrices are injected into the model architecture, while the remaining model parameters are kept frozen.

$$W = W_0 + \Delta W \quad (11)$$

$$\Delta W = A_k B_k \quad (12)$$

Where W_0 denotes the original parameter matrix and W represents the LoRA-adapted parameter matrix, with $\Delta W, W_0 \in \mathbb{R}^{m \times n}$. The matrices A_k and B_k are two low-rank matrices, where $A_k \in \mathbb{R}^{m \times r}$ and $B_k \in \mathbb{R}^{r \times n}$. Here, r refers to the rank of the low-rank adaptation, satisfying $r \ll \min(m, n)$.

As illustrated in Figure 3, the LoRA mechanism can be integrated into two different components: the fully connected layers and the attention layers. In the DC fusion mode, only the fully connected layers are LoRA-adapted, as the attention layers are not incorporated into this mode.

4. Experimental Setups And Results

4.1. Data

The DAIC-WOZ dataset [31] includes 189 English clinical interview transcripts, divided into a training set (107 samples), a development set (35 samples), and a test set. The EATD-corpus [15] is a Chinese depression dataset containing 162 participants, split into a training set (108 samples) and a test set (54 samples).

In the DAIC-WOZ dataset, audio recordings were segmented into question-answer pairs of varying lengths. To address class imbalance, we used the method from Shen et al. [15], forming samples with ten question-answer pairs and balancing the ratio of depressed to healthy participants by sampling multiple subsets for depressed individuals. Text data were processed similarly, segmenting each response into individual samples. For the EATD-Corpus, imbalance was mitigated by merging answers to three questions from depressed participants, following the same method. The feature processing mirrored that of

Table 2: Comparison of our framework’s performance on cross-language datasets using different strategies.

Fusion	Finetune	Parameters	EATD Test Set			DAIC Test Set		
			F1-Score	Recall	Precision	F1-Score	Recall	Precision
DC	-	2.42M	0.85	0.84	0.86	0.69	0.92	0.55
	LAFT	0.05M	0.87	0.86	0.88	0.79	0.92	0.69
	APFT	2.42M	0.88	0.86	0.90	0.81	0.92	0.73
CA	-	2.52M	0.85	0.86	0.84	0.71	0.83	0.63
	LAFT	0.11M	0.87	0.92	0.82	0.80	1.0	0.67
	APFT	2.52M	0.88	0.86	0.90	0.83	0.83	0.93
CA+DC	-	2.52M	0.87	0.86	0.88	0.71	0.92	0.58
	LAFT	0.11M	0.89	0.84	0.95	0.83	1.0	0.71
	APFT	2.52M	0.91	0.86	0.96	0.86	1.0	0.75

DAIC-WOZ. We then combined three sentences from each English sample with the Chinese training dataset to create a cross-lingual set, balancing data across languages for more comparable training conditions.

4.2. Configurations

The speech branch uses Whisper-large-v3 to extract 1280-dimensional features from its final encoder layer, capturing contextual and temporal information, which are then projected to 128 dimensions via a fully connected layer. Similarly, the text branch uses XLM-RoBERTa to extract 768-dimensional cross-lingual features, which are also projected to 128-dimensional embeddings through a fully connected layer.

For the loss function, when the model employs channel-based attention with the concatenation method, we set $\alpha = 0.01$. In all other cases, α is set to 0.

In the APFT stage, all model parameters are updated during training, regardless of whether the model uses connection fusion or channel attention fusion. In the LAFT stage, for the direct-connected (DC) fusion mode, trainable low-rank factorization matrices are incorporated into the fully connected layers of both the linguistic and textual feature extraction branches. In the Channel Attention (CA) fusion mode, as shown in Fig. 3, LoRA parameters are introduced into the attention mechanism by adding trainable low-rank matrices to the mapping layers of the query, key, and value, building upon the DC fusion mode. This approach is consistent with the DC+CA fusion mode. During training, the original parameters of the feature fusion module remain frozen, while the parameters of the factorization matrices are updated. In this study, we set the rank $r = 32$ and the learning rate to 3×10^{-6} , which is one-tenth of the learning rate used when training the original parameters.

4.3. Results

4.3.1. Validation of the proposed architecture

To validate the effectiveness of the proposed architecture, we first conduct a series of experiments on mono-lingual speech data, specifically using the EATD dataset. The corresponding results are presented in Table 1. From these results, it can be observed that our single-modality models (First two lines) achieves performance that is comparable to, or even surpasses, the results reported in the literature [13, 15, 19, 28].

Moreover, when different feature fusion techniques were employed to integrate the audio and text modalities, significant improvements were observed across all metrics, emphasizing the complementary nature of the two modalities and the effectiveness of the fusion strategies. For comparison, we included results from existing dual-branch multimodal mod-

els for depression detection. Our models, using three fusion methods, consistently outperformed existing dual-branch multimodal models in F1-score, recall, and precision, with the best model achieving an F1-score of **0.83**. These results demonstrate the effectiveness of our framework in a mono-lingual setting, showcasing its ability to combine multimodal information effectively.

4.3.2. Effectiveness of multi-language training

We expect a performance boost by adding more cross-lingual depression data. Comparing the results on the EATD test set in Table 1 and Table 2, we find that models trained on multilingual datasets consistently outperform those before fine-tuning, especially in terms of precision. This suggests that cross-lingual training improves the model’s prediction reliability. For example, the best model achieved an F1-score of 0.87 on the EATD dataset.

4.3.3. Effectiveness of language-specific finetuning

After multilingual training, we fine-tuned the models on target datasets using two strategies for domain adaptation. As shown in Table 2, all fine-tuned models outperformed their pre-fine-tuned versions, particularly on the DAIC dataset.

LAFT efficiently adapted by updating only a small subset of parameters, achieving an F1-score of 0.89 on EATD and 0.83 on DAIC with just 0.1M parameters (4% of the original). APFT, on the other hand, delivered the highest performance, with an F1-score of 0.91 on EATD and 0.86 on DAIC.

These findings emphasize the value of combining multi-lingual pretraining with language-specific fine-tuning for depression detection, where multilingual training provides strong feature representation, and fine-tuning drives substantial performance improvements.

5. Conclusion

In this work, we propose M3L, a multi-modal and multi-lingual depression detection framework. By leveraging the large pre-trained multilingual models, Whisper and XLM-RoBERTa, M3L achieves impressive performance on both English and Chinese benchmarks. Furthermore, we conducted a detailed validation and comparison of different modality fusion strategies and fine-tuning methods. The results highlight the effectiveness of combining multi-lingual pretraining with language-specific fine-tuning, demonstrating its ability to enhance performance across different languages and modalities.

6. Acknowledgements

This work is supported by National Natural Science Foundation of China (Grant No. 62401377), the Shenzhen Science and Technology Program (JCYJ20220818103001002), the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen, the Project (No. 20232ABC03A25), the Longgang District Special Funds for Science and Technology Innovation (LGKCSPT2023002) and Yangtze River Delta Science and Technology Innovation Community Joint Research Project (2024CSJGG01100).

7. References

- [1] W. H. Organization *et al.*, “Depression and other common mental disorders: global health estimates,” 2017.
- [2] F. Rice, L. Riglin, T. Lomax, E. Souter, R. Potter, D. Smith, A. K. Thapar, and A. Thapar, “Adolescent and adult differences in major depression symptom profiles,” *Journal of affective disorders*, vol. 243, pp. 175–181, 2019.
- [3] L. Orsolini, R. Latini, M. Pompili, G. Serafini, U. Volpe, F. Velante, M. Fornaro, A. Valchera, C. Tomasetti, S. Fraticelli *et al.*, “Understanding the complex of suicide in depression: from research to clinics,” *Psychiatry investigation*, vol. 17, no. 3, p. 207, 2020.
- [4] L. L. Craft and D. M. Landers, “The effect of exercise on clinical depression and depression resulting from mental illness: A meta-analysis,” *Journal of Sport and Exercise Psychology*, vol. 20, no. 4, pp. 339–357, 1998.
- [5] H. Wang, Y. Liu, X. Zhen, and X. Tu, “Depression speech recognition with a three-dimensional convolutional network,” *Frontiers in human neuroscience*, vol. 15, p. 713823, 2021.
- [6] K. M. Smith, P. F. Renshaw, and J. Bilello, “The diagnosis of depression: current and emerging methods,” *Comprehensive psychiatry*, vol. 54, no. 1, pp. 1–6, 2013.
- [7] X. Zhang, Z. Zhang, W. Diao, C. Zhou, Y. Song, R. Wang, X. Luo, and G. Liu, “Early-diagnosis of major depressive disorder: From biomarkers to point-of-care testing,” *TrAC Trends in Analytical Chemistry*, vol. 159, p. 116904, 2023.
- [8] S. Yasin, A. Othmani, I. Raza, and S. A. Hussain, “Machine learning based approaches for clinical and non-clinical depression recognition and depression relapse prediction using audiovisual and eeg modalities: A comprehensive review,” *Computers in Biology and Medicine*, vol. 159, p. 106741, 2023.
- [9] N. Cummins, F. Matcham, J. Klapper, and B. Schuller, “Artificial intelligence to aid the detection of mood disorders,” in *Artificial Intelligence in Precision Health*. Elsevier, 2020, pp. 231–255.
- [10] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, “Voice acoustical measurement of the severity of major depression,” *Brain and cognition*, vol. 56, no. 1, pp. 30–35, 2004.
- [11] J. K. Darby, N. Simmons, and P. A. Berger, “Speech and voice parameters of depression: A pilot study,” *Journal of Communication Disorders*, vol. 17, no. 2, pp. 75–85, 1984.
- [12] M. R. Morales and R. Levitan, “Speech vs. text: A comparative analysis of features for depression detection systems,” in *2016 IEEE spoken language technology workshop (SLT)*. IEEE, 2016, pp. 136–143.
- [13] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, “Detecting depression with audio/text sequence modeling of interviews,” in *Interspeech*, 2018, pp. 1716–1720.
- [14] N. Seneviratne and C. Espy-Wilson, “Multimodal depression classification using articulatory coordination features and hierarchical attention based text embeddings,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6252–6256.
- [15] Y. Shen, H. Yang, and L. Lin, “Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6247–6251.
- [16] M. Fang, S. Peng, Y. Liang, C.-C. Hung, and S. Liu, “A multimodal fusion model with multi-level attention mechanism for depression detection,” *Biomedical Signal Processing and Control*, vol. 82, p. 104561, 2023.
- [17] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, “Multi-level attention network using text, audio and video for depression prediction,” in *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, 2019, pp. 81–88.
- [18] J. Wang, V. Ravi, and A. Alwan, “Non-uniform speaker disentanglement for depression detection from raw speech signals,” in *Interspeech*, vol. 2023. NIH Public Access, 2023, p. 2343.
- [19] 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.
- [20] A. Anand, C. Tank, S. Pol, V. Katoch, S. Mehta, and R. R. Shah, “Depression detection and analysis using large language models on textual and audio-visual modalities,” *arXiv preprint arXiv:2407.06125*, 2024.
- [21] L. Liu, F. Tydeman, W. Xie, and Y. Wang, “Multilingual depression detection based on speech signals and deep learning,” in *2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService)*. IEEE, 2024, pp. 115–116.
- [22] N. Cummins, J. Dineley, P. Conde, F. Matcham, S. Siddi, F. Lamers, E. Carr, G. Lavelle, D. Leightley, K. M. White *et al.*, “Multilingual markers of depression in remotely collected speech samples: A preliminary analysis,” *Journal of affective disorders*, vol. 341, pp. 128–136, 2023.
- [23] M. S. I. Malik, A. Nazarova, M. M. Jamjoom, and D. I. Ignatov, “Multilingual hope speech detection: A robust framework using transfer learning of fine-tuning roberta model,” *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 8, p. 101736, 2023.
- [24] G. Kiss and K. Vicsi, “Mono-and multi-lingual depression prediction based on speech processing,” *International Journal of Speech Technology*, vol. 20, pp. 919–935, 2017.
- [25] C. Demiroglu, A. Beşirli, Y. Ozkanca, and S. Çelik, “Depression-level assessment from multi-lingual conversational speech data using acoustic and text features,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, p. 17, 2020.
- [26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [27] A. Conneau, “Unsupervised cross-lingual representation learning at scale,” *arXiv preprint arXiv:1911.02116*, 2019.
- [28] J. Xue, R. Qin, X. Zhou, H. Liu, M. Zhang, and Z. Zhang, “Fusing multi-level features from audio and contextual sentence embedding from text for interview-based depression detection,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6790–6794.
- [29] D. Hazarika, R. Zimmermann, and S. Poria, “Misa: Modality-invariant and-specific representations for multimodal sentiment analysis,” in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.
- [30] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, “Lora: Low-rank adaptation of large language models,” *JCLR*, vol. 1, no. 2, p. 3, 2022.
- [31] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, “The distress analysis interview corpus of human and computer interviews,” in *LREC*. Reykjavik, 2014, pp. 3123–3128.