AdoDAS: A Privacy-Preserving Multimodal Challenge for Adolescent Depression, Anxiety, and Stress Assessment

Abstract

We present **AdoDAS**, a privacy-preserving multimodal grand challenge dataset for adolescent Depression/Anxiety/Stress (D/A/S) assessment. **AdoDAS contains 6,000 child and adolescent participants and 24,000 audio-video segments** collected via a controlled school-based protocol that combines a standardized reading passage with open-ended interview prompts. Ground-truth labels are derived from DASS-21, providing both three subscale scores (D/A/S) and 21 item-level responses to support coarse screening and fine-grained, interpretable symptom modelling. Unlike recent grand challenges that focus on social-media text or adult interview settings, AdoDAS targets minors and addresses stringent privacy constraints by withholding raw recordings and releasing reproducible pre-computed representations and temporal metadata. We provide two benchmark tracks with strong baselines to facilitate robust, subject-disjoint evaluation and advance safe multimodal mental-health research.

1 Introduction

Adolescent mental health screening is critical for early identification and timely intervention, yet scalable assessment remains challenging in real-world practice. Depression, anxiety, and stress (D/A/S) are prevalent during adolescence and frequently co-occur, while routine screening still relies heavily on questionnaires and limited clinical resources. Recent progress in multimodal affective computing suggests that audio-visual behavioural cues during communication can provide signals complementary to self-reports, motivating standardised, reproducible benchmarks for multimodal D/A/S *screening and assessment*.

Despite growing interest, reproducible multimodal benchmarks for *minors* remain scarce. A key barrier is privacy and ethics: releasing raw audio/video recordings of children and adolescents can enable re-identification even after conventional de-identification, making open data sharing difficult. In addition, existing mental-health resources often differ in task definitions, data splits, and evaluation protocols, which hinders fair comparison and may introduce subject leakage. Prior shared tasks such as the MPDD Challenge [6] have advanced multimodal depression detection with individual-difference annotations, but a minors-oriented benchmark that jointly targets D/A/S under realistic privacy constraints is still missing.

Existing depression-related benchmarks span controlled clinical interviews and in-the-wild social media/video. Clinical interview corpora (e.g., DAIC-WOZ) have been widely used for speech-/text-based modelling and robustness studies [2, 8, 11, 14, 16–18, 21]. In-the-wild datasets (e.g., D-Vlog) support non-verbal behaviour analysis beyond lab settings [19], while longitudinal multimodal datasets (e.g., MUD3) provide richer temporal coverage for user-level assessment [15]. Social-media shared tasks and recent EMNLP work further explore scalable detection from posts and highlight generalisation issues [1, 9, 12, 13]. Compared with these resources, we focus on *adolescent* multimodal D/A/S screening with a controlled elicitation protocol and a privacy-preserving release strategy that supports subject-disjoint evaluation at scale.

To address this gap, we introduce **AdoDAS**, a feature-centric, privacy-preserving multimodal Grand Challenge for adolescent D/A/S assessment with labels derived from DASS-21 (Appendix A). Data are collected using a tightly controlled school-based protocol that combines a standardised reading passage and open-ended question answering, balancing controlled elicitation and spontaneous expression while improving comparability across participants and sites. To enable reproducible research without exposing raw signals, AdoDAS withholds raw audio/video and releases pre-computed acoustic descriptors, visual representations, and cross-modal temporal metadata (e.g., VAD-based alignment) to support temporal modelling and multimodal fusion. In total, AdoDAS includes 6,000 participants and 24,000 audio-video segments across four sessions per participant.

The challenge provides two tracks: (i) **Track A1** performs multi-task binary screening for D/A/S; (ii) **Track A2** predicts DASS-21 item-level responses with subscale reconstruction. We further provide official subject-disjoint splits, standardised evaluation tools, and baseline implementations to support fair comparison and reproducible research.

In summary, the AdoDAS Challenge contributes: (1) a large-scale adolescent multimodal dataset collected with a controlled protocol; (2) a privacy-preserving design with temporal metadata for multimodal modelling; (3) two complementary tasks with official splits, evaluation scripts, baseline methods, and a leaderboard protocol to facilitate future benchmarking.

2 Dataset

The dataset is available on **https://AdoDAS Dataset**.

2.1 Participants and Ethics

AdoDAS is collected from 6,000 child and adolescent participants, yielding 24,000 audio-video segments (Detailed at Figure 1). All procedures follow institutional/school ethics requirements for minors. Written informed consent is obtained from legal guardians and assent from participants. The released benchmark contains no raw audio/video and is designed to minimise re-identification risks (Section 2.5).

2.2 Recording Protocol

Each participant contributes four sessions under two complementary elicitation conditions: one scripted reading session (A01) and three interview-style open-ended sessions (B01–B03) designed to elicit spontaneous behaviours (Table 1). Recordings are acquired under standardised conditions to reduce nuisance variability: participants are seated in a quiet classroom, centred on screen, with a fixed mouth-to-microphone distance. A tablet-based prompting system is used; each session is capped at 60 s, and re-recording is permitted upon interruption.

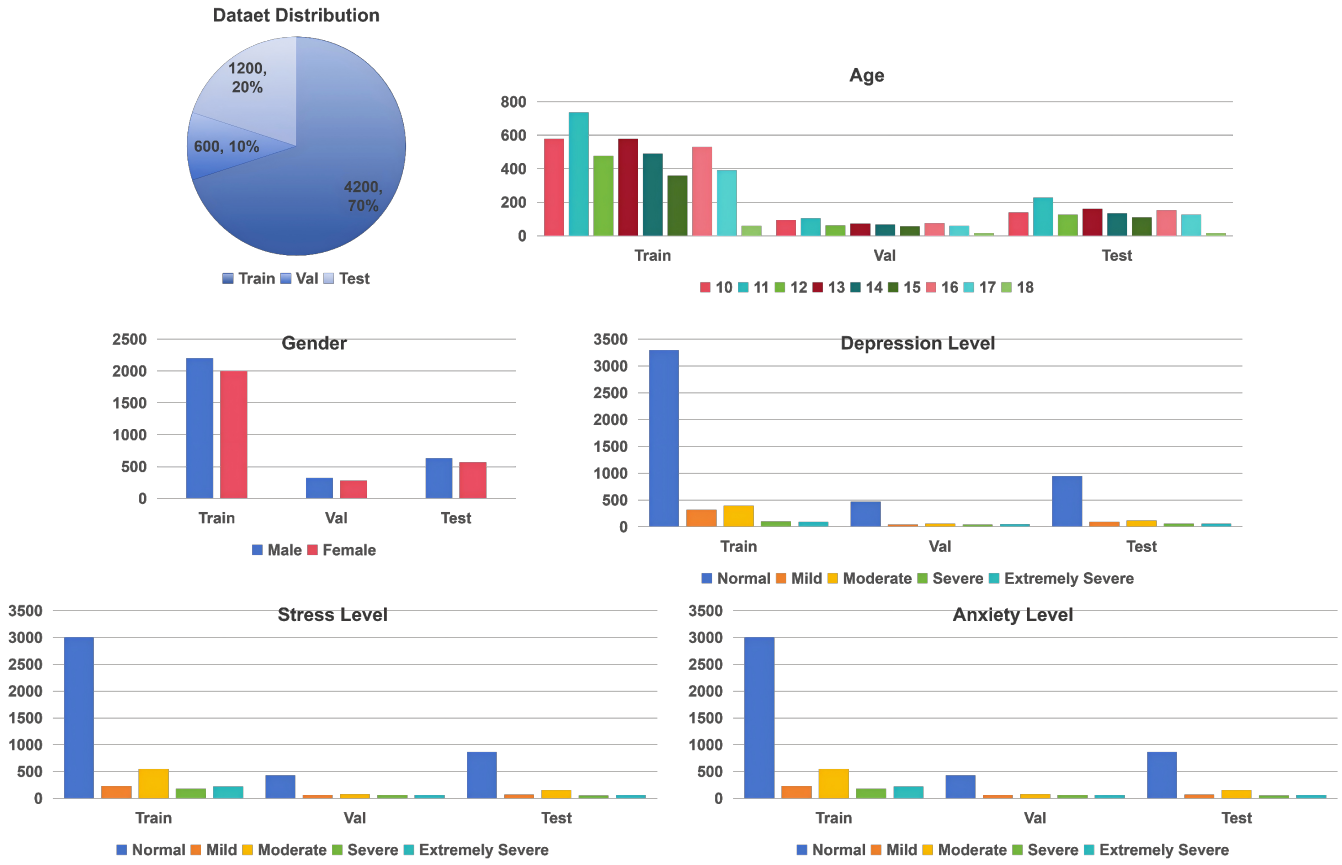


Figure 1: AdoDAS data distribution across sessions and label groups (D/A/S).

Table 1: Elicitation prompts used for recording.

Order	Type	Contents
A01	Fixed Text Reading	Once, the North Wind and the Sun argued about who was stronger. Suddenly, a traveler passed by, wearing a thick coat. They agreed to a challenge: whoever could make the traveler take off his coat first would be proven stronger. The North Wind blew fiercely, but the harder he blew, the tighter the man wrapped his coat around himself. In the end, the North Wind gave up. Soon after, the Sun came out and shone warmly. Instantly, the man took off his coat. Thus, the North Wind had to admit that the Sun was stronger.
B01	Open-ended Question	Please describe how your day went yesterday.
B02	Open-ended Question	Please describe your happiest memory from the past week.
B03	Open-ended Question	Please describe your saddest memory from the past week.

2.3 Annotation and Labels

Ground-truth labels are derived from DASS-21 self-report assessments, which provide 21 item-level ordinal responses (0–3) and three subscale scores for Depression, Anxiety, and Stress (D/A/S) (Appendix A, Table 4). Compared to diagnosis-oriented scales such as PHQ-9 grounded in DSM criteria, DASS-21 focuses on symptom severity across multiple affective dimensions and offers an itemised structure, enabling both coarse screening and fine-grained symptom modelling from a single annotation source.

For binary screening labels (used in Track A1), we map `Normal` to the negative class and `Mild-or-above` to the positive class for each subscale. We report label prevalence and class imbalance statistics in Figure 1.

2.4 Auxiliary Attributes

In addition to DASS-21, we collect several optional demographic/context attributes (Table 2) for *stratified analysis* (e.g., subgroup evaluation). **These attributes are not intended as model inputs for leaderboard ranking** to reduce shortcut learning risks and to preserve participant privacy; they are provided for post-hoc analysis only.

Table 2: Optional auxiliary attributes in AdoDAS.

Attribute	Description
family structure	1=Nuclear; 2=Extended; 3=Single-parent; 4=Blended; 5=Skipped-generation; 6=Other.
only child status	Whether the respondent is an only child (Yes/No).
parental favoritism	If not an only child: 1=Favoring siblings; 2=No favoritism; 3=Favoring self.
academic performance change	Compared with the previous semester: 1=Improved; 2=Declined; 3=Stable.
emotional state change	Compared with the previous semester: 1=Better; 2=Worse; 3=No change.

2.5 Privacy-Preserving Release

To mitigate re-identification risks for minors, AdoDAS adopts a privacy-preserving, representation-level release: no raw audio, raw video, or identifiable frame images are distributed. Instead, each segment is provided as: (i) pre-computed acoustic descriptors; (ii) pre-computed visual behaviour representations; and (iii) cross-modal temporal metadata (e.g., speech-activity/alignment information) to support temporal modelling and multimodal fusion. All resources are indexed by consistent identifiers to enable deterministic joins across modalities.

2.6 Official Splits and Access

For fair comparison and to prevent identity leakage, official data splits are subject-disjoint: all segments from the same participant (A01 and B01–B03) appear in exactly one split (train/validation/test: 7/1/2). We release official loaders and evaluation scripts to ensure standardised access and reproducible reporting.

3 Tracks

All tracks share the same privacy-preserving, feature-only audio-visual inputs and follow official subject-disjoint splits: all four sessions (A01, B01–B03) from the same participant appear in exactly one split. The leaderboard evaluation is computed by the official script to ensure deterministic metric computation and ranking.

Track A1: Multi-task Binary Screening (D/A/S).

Task For each segment (`file_id`), predict three binary outcomes for Depression (D), Anxiety (A), and Stress (S): `Normal` (0) vs. `Mild-or-above` (1). Ground-truth targets are provided by the released severity fields (e.g., `dass\_depression\_level\_3`, `dass\_anxiety\_level\_3`, `dass\_stress\_level\_3`); `Normal` is mapped to 0 and all other levels are mapped to 1.

Output Three real-valued confidence scores per `file_id`, one for each target (D/A/S). Scores are interpreted as positive-class probabilities in $[0, 1]$ by the official script (logits must be transformed by participants).

Primary metric Mean F1 over the three targets:

$$\text{Score}_{\text{A1}} = \frac{1}{3} \sum_{t \in \{D,A,S\}} \text{F1}_t,$$

where F1_t is the binary F1 computed after thresholding the predicted probability at a fixed threshold $\tau = 0.5$.

Secondary metric Mean AUROC over D/A/S using the submitted scores.

Ranking Rank by Score_{A1} (higher is better); break ties by mean AUROC (higher is better).

Submission A CSV file with one row per `file_id` and three columns `p_D`, `p_A`, `p_S` (positive-class probabilities). The official script validates column names and ranges.

Track A2: DASS-21 Item Response Prediction and Subscale Reconstruction.

Task For each segment (`file_id`), predict the 21 DASS-21 item responses, each taking an ordinal value in $\{0, 1, 2, 3\}$. Outputs correspond to `dass\_d01\_score\_3`–`dass\_d21\_score\_3`. This track encourages learning symptom structure and enables interpretable item-level analysis.

Output Twenty-one predicted item scores per `file_id`. The official script expects integer outputs in $\{0, 1, 2, 3\}$. If a method produces real values, participants should convert them to integers (e.g., clipping to $[0, 3]$ and rounding) before submission.

Subscale reconstruction Depression/Anxiety/Stress subscale scores are deterministically reconstructed from the predicted items using the standard DASS-21 rule (Appendix A; Eq. (1)). Reconstructed scores may be mapped to severity levels for analysis; *ranking remains item-based.*

Primary metric Mean Quadratic Weighted Kappa (QWK) over 21 items:

$$\text{Score}_{\text{A2}} = \frac{1}{21} \sum_{i=1}^{21} \text{QWK}(y_i, \hat{y}_i),$$

computed per item with label set $\{0, 1, 2, 3\}$ and quadratic weights, then averaged across items.

Secondary metric Mean MAE over 21 items:

$$\text{MAE}_{\text{A2}} = \frac{1}{21} \sum_{i=1}^{21} \mathbb{E}\left[|y_i - \hat{y}_i|\right],$$

reported by the official script (lower is better).

Auxiliary metric Mean CCC over reconstructed continuous subscale scores (D/A/S), computed separately for each subscale and then averaged (higher is better).

Ranking Rank by Score_{A2} (higher is better); break ties by mean MAE (lower is better), then mean CCC (higher is better).

Submission A CSV file with one row per `file_id` and 21 integer columns `d01`, `d02`, …, `d21`. The official script computes item-level metrics and reconstructed-score diagnostics.

4 Baseline Methods

We encourage participants to build on recent advances in speech/text foundation models and multimodal fusion for depression screening [3, 4, 7, 10, 20]. Given privacy and domain-shift concerns, methods

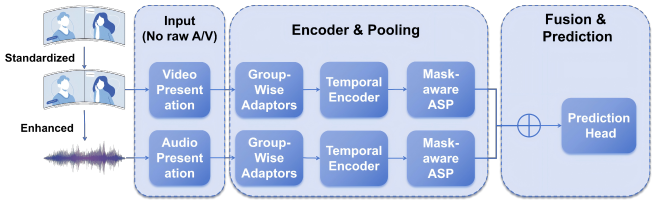


Figure 2: Overview of Baseline

that reduce speaker leakage and improve robustness (e.g., speaker disentanglement, test-time adaptation, and ordinal modeling) are also relevant [5, 11, 16, 21].

The baseline is a feature-only, privacy-preserving temporal model that leverages 25 Hz aligned audio-visual sequences together with validity/quality masks. Given heterogeneous feature groups, it adopts a modular design with modality-specific temporal encoders, mask-aware pooling, and prediction heads tailored to Track A1 (binary D/A/S screening) and Track A2 (DASS-21 item response prediction).

4.1 Components.

Group-wise adapters Each feature group is normalized and linearly projected into a shared latent width, alleviating scale mismatch between low-dimensional behavioral cues and high-dimensional SSL embeddings.

Temporal encoders Projected groups are merged within each modality to form audio and visual streams on the same 25 Hz timeline, then processed by stacks of residual dilated temporal convolution blocks to capture both local dynamics and longer-range trends with linear-time complexity. Invalid frames are zeroed and masked throughout.

Mask-aware attentive statistics pooling To obtain clip-level representations, an attention module computes frame weights under validity masks and aggregates temporal embeddings via weighted mean and weighted standard deviation. The attention can be informed by temporal metadata and visual quality cues to reduce the impact of missing/low-quality frames.

Fusion and multi-task heads Audio and visual clip-level embeddings are concatenated with optional pooled/tabular descriptors, followed by a lightweight MLP trunk. For Track A1, three task-specific heads output binary logits for D/A/S (`Normal` vs. `Non-normal`). For Track A2, an item-response head predicts the 21 DASS-21 item scores, implemented as 21 four-way classifiers (0-3) or ordinal-regression heads. AdoDAS evaluation additionally reconstructs D/A/S subscale scores from the predicted items for analysis.

4.2 Feature sets

Audio inputs include 25 Hz spectral descriptors (log-mel/MFCC), 25 Hz-aligned self-supervised speech embeddings, and VAD-derived speech/pause cues. Visual inputs summarize facial and motion behaviors without releasing raw frames, including quality-control signals, head-pose/geometry cues, facial-behavior sequences, global motion descriptors, body-pose landmarks, and deep visual embeddings. Cross-modal temporal metadata (e.g., VAD aggregated to video frames) is also provided to support speech-conditioned modeling and robust pooling.

4.3 Others

We train the baseline with AdamW and early stopping on the public development split; class weighting is supported to mitigate label imbalance, and all evaluation is performed by the official script for reproducibility.

5 Schedule

Data, website, baseline & code available	15 Mar, 2026
Results submission start	09 May, 2026
Results submission deadline	20 May, 2026
Deadline for paper submission	28 May, 2026
Paper acceptance notification	16 Jul., 2026
Deadline for camera-ready papers	06 Aug., 2026

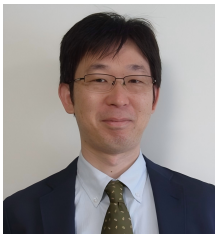
6 People

6.1 Organizer

Haizhou Li (Fellow, IEEE) received the B. Sc., M. Sc., and Ph.D degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China in 1984, 1987, and 1990 respectively. He is currently a Presidential Chair Professor and the Executive Dean of the School of Data Science, The Chinese University of Hong Kong, China. He is also an Adjunct Professor with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests include automatic speech recognition, speaker and language recognition, natural language processing. He was the Editor-in-Chief of IEEE/ACM Transactions on Audio, Speech and Language Processing during 2015–2018, an elected Member of IEEE Speech and Language Processing Technical Committee during 2013–2015, the President of the International Speech Communication Association during 2015–2017, President of Asia Pacific Signal and Information Processing Association during 2015–2016, and President of Asian Federation of Natural Language Processing during 2017–2018. Since 2012, he has been a Member of the Editorial Board of Computer Speech and Language. He was the General Chair of ACL 2012, INTERSPEECH 2014, ASRU 2019 and ICASSP 2022. He is a Fellow of the ISCA, and a Fellow of the Academy of Engineering Singapore. He was the recipient of the National Infocomm Award 2002, and President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, and U Bremen Excellence Chair Professor in 2019.

Hiroshi Ishiguro received a Ph. D. from Osaka University, Japan in 1991. He is currently Professor of Department of Systems Innovation at Osaka University, Visiting Director of Hiroshi Ishiguro Laboratories at the Advanced Telecommunications Research Institute (ATR), Project Manager of MOONSHOT R&D Project, Thematic Project Producer of EXPO 2025 Osaka, Kansai, Japan, and CEO of AVITA, Inc. His research interests are interactive robotics, avatar, and android science. Geminoid is an avatar android that is a copy of himself. In 2011, he won the Osaka Cultural Award. In 2015, he received the Prize for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology. He was also awarded the Sheikh Mohammed Bin Rashid Al Maktoum Knowledge Award in Dubai in 2015. Tateisi Award in 2020, and

honorary doctorate of Aarhus university in 2021.



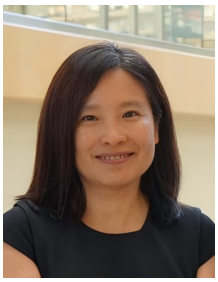
Tetsuya Takiguchi received the M. Eng. and Dr. Eng. degrees from Nara Institute of Science and Technology, Japan, in 1996 and 1999, respectively. From 1999 to 2004, he was a Researcher with Tokyo Research Laboratory, IBM Research. From 2004 to 2016, he was an Associate Professor with Kobe University. From May 2008 to September 2008, he was a Visiting Scholar with the Department of Electrical Engineering, University of Washington. From March 2010 to a Visiting Scholar with the Institute for Learning and Brain Sciences, University of Washington. From April 2013 to October 2013, he was a Visiting Scholar with the Laboratoire d'InfoRmatique en Image et Systemes d'information, INSA Lyon. Since 2016, he has been a Professor with Kobe University. His research interests include speech, image, and brain processing, and multimodal assistive technologies for people with articulation disorders. He received the Best Paper Award from the IEEE ICME 2008.

Zhaojie Luo (Member, IEEE) received the M.Eng. and Dr.Eng. degrees from Kobe University, Kobe, Japan, in 2017 and 2020, respectively. He is currently an Associate Professor with Southeast University, Nanjing, China. From 2020 to 2024 he was an assistant Professor with Osaka University, Suita, Japan. From 2019 to 2020, he was a Researcher with the Department of Electrical & Computer Engineering, National University of Singapore, Singapore. His research interests include voice conversion, speech synthesis, facial expression recognition, multimodal emotion recognition, and statistical signal processing. He has published more than 20 papers in top-tier speech/multimedia journals and international conferences, such as IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, IEEE TRANS. MULTIMEDIA, EURASIP JASMP, ACM Multimedia, INTERSPEECH, ICASSP, SSW, ICME, and ICPR. He is a member of ISCA and ASJ, and is the Reviewer for many major referred journal and conference papers.

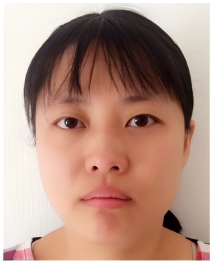
Tomoko Matsui is a Professor at the Shenzhen Loop Area Institute (SLAI). Her research focuses on statistical machine learning and the statistical foundations of artificial intelligence, with particular emphasis on speech and language data, time-series and longitudinal modeling, and uncertainty-aware learning. Her core interests include spatio-temporal and change-point modeling, and statistical learning from longitudinal data, with applications to speech and language analytics and conversational AI. She is particularly interested in developing reliable and trustworthy AI systems by evaluating uncertainty, handling non-stationary or unexpected data conditions, and integrating large language models (LLMs) with principled statistical methods for real-world decision-making. She has been actively engaged in international and interdisciplinary research across statistics and AI, has published extensively in leading journals and conferences, co-edited scholarly volumes, and has served on IEEE technical committees in Speech and Language Processing and Machine Learning for Signal Processing, including as Chair of the 2017 IEEE International Workshop on Machine Learning for Signal Processing.

Kun Qian received his doctoral degree for his study on automatic general audio signal classification in 2018 in electrical engineering and information technology from Technische Universität München (TUM), Germany. He is a Full Professor and the Vice Dean of the School of Medical Technology at Beijing Institute of Technology, China. He is a Senior Member of the IEEE. He won the ACM Beijing Rising Star Award in 2025, Beijing Science and Technology Award (Second Prize for Scientific and Technological Progress) in 2023, Invention Entrepreneurship and Innovation Award in 2023, and the Excellent Achiement Award of Zhejiang Lab in 2020. He was elected as a Forbes China "100 Outstanding Overseas Returnees" in 2023 and a MIT Technology Review (China) & DeepTech "Intelligent Computing Innovators 2023 of China". Prof. Qian serves as an Associate Editor for the IEEE Transactions on Affective Computing, Frontiers in Digital Health, and BIO Integration. He (co-)authored more than 200 publications in peer reviewed journals, and conference.

6.2 Data Chair



Fei Wang is a professor and Director of Early Intervention Unit, Affiliated Nanjing Brain Hospital, Nanjing Medical University. She is a psychiatric scientist with expertise in neuroimaging, genetics, mental health screening and intervention research. She has been engaged in engaged in clinical, translational and mulit-disciplinary research for more than 30 years. The first line of related work is a molecular-neural circuitry-clinical manifestation association study for studying the shared neuropathological characteristics across major psychiatric disorder. Bringing together artificial intelligence (AI) and the findings in neuroimaging and multi-omics data, Prof. Wang also studied the clinical safety and efficacy responses to conventional physiotherapy (rTMS) and psychotherapy (CBT) with the goal of developing multidimensional biomarkers that can be used to personalize treatment and prevention strategies. The second is large-scale mental health screening (SEARCH cohort) and online stratified intervention research outside the hospital to develop new approaches that take advantage of emerging internet informatics technologies, AI.



Shuqiong Wu received her B.E. and M.E. degrees from Beihang University, Beijing, China, in 2008 and 2011, respectively. She obtained her Ph.D. degree in Computational Intelligence and Systems Science from the Institute of Science Tokyo, Tokyo, Japan, in 2015. From 2015 to 2020, she was a research fellow at the Graduate School of Informatics, Kyoto University. From 2020 to 2025, she served as an assistant professor at SANKEN (The Institute of Scientific and Industrial Research), the University of Osaka. She is currently an associate professor at the Graduate School of Engineering Science, the University of Osaka. Her current research topics include dualtask-based cognitive impairment detection, cognitive status monitoring, medical image reconstruction, and contactless biometric sensing. Her research interests include biomedical signal processing, image processing, three-dimensional reconstruction, pattern recognition, and machine learning.

Zhengjun Yue is an Assistant Professor at Shenzhen Loop Area Institute at the Center of Language, Intelligence and Machines. She worked at the Delft University of Technology, the Netherlands from 2022 to 2025 as an Assistant Professor and as a Research Associate at King's College London (KCL), UK. She received her Ph.D. (funded by the Marie-Curie H2020 TAPAS project) from the University of Sheffield, UK in 2022, with a focus on continuous speech recognition for people with dysarthria. She received her M.Sc. degree in Artificial Intelligence from the University of Edinburgh, UK in 2018. She has published in IEEE-TASLP, Computer Speech & Language, ICASSP, INTERSPEECH, and ASRU. Her research interests lie in acoustic modelling and multi-modal speech recognition for atypical speech, end-to-end and robust ASR and atypical speech processing.

Junkun Wang received his B.E. degree from University of Electronic Science and Technology of China, Chengdu, currently working toward M.E. degree of Electronic Information in Southeast University, Nanjing. He is a second-year master's student in Health and Affective Intelligent Lab (HAI LAB), advised by Prof. Zhaojie Luo and Prof. Zhongze Gu. His Research interests includes affective computing, deep learning, MLLM.

Tianhua Qi is currently working toward the PhD degree with the School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China, and also with the Key Laboratory of Child Development and Learning Science (Southeast University), Ministry of Education, China. He is the general co-chair of The 12th ISCA-SAC Doctoral Consortium. His research interests include affective computing, deep learning, and speech signal processing.

References

[1] Nuredin Ali Abdelkadir, Charles Zhang, Ned Mayo, and Stevie Chancellor. 2024. Diverse Perspectives, Divergent Models: Cross-Cultural Evaluation of Depression Detection on Twitter. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. 672–680. doi:10.18653/v1/2024.naacl-short.58

[2] Catarina Botelho, Tanja Schultz, Alberto Abad, and Isabel Trancoso. 2022. Challenges of using longitudinal and cross-domain corpora on studies of pathological speech. In *Interspeech 2022*. 1921–1925. doi:10.21437/Interspeech.2022-10995

[3] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2022. SpeechFormer: A Hierarchical Efficient Framework Incorporating the Characteristics of Speech. In *Interspeech 2022*. 346–350. doi:10.21437/Interspeech.2022-74

[4] Qingkun Deng, Saturnino Luz, and Sofia de la Fuente Garcia. 2025. An interpretable speech foundation model for depression detection by revealing prediction-relevant acoustic features from long speech. In *Interspeech 2025*. 5248–5252. doi:10.21437/Interspeech.2025-1968

[5] Sri Harsha Dumpala, Chandramouli S. Sastry, Rudolf Uher, and Sageev Oore. 2025. Test-Time Training for Speech-based Depression Detection. In *Interspeech 2025*. 479–483. doi:10.21437/Interspeech.2025-2378

[6] Changzeng Fu, Zelin Fu, Xinhe Kuang, Jiacheng Dong, Qi Zhang, Kaifeng Su, Yikai Su, Wenbo Shi, Junfeng Yao, Yuliang Zhao, Shiqi Zhao, Jiadong Wang, Siyang Song, Chaoran Liu, Yuichiro Yoshikawa, Björn Schuller, and Hiroshi Ishiguro. 2025. The First MPDD Challenge: Multimodal Personality-aware Depression Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*. Association for Computing Machinery, 13924–13929. doi:10.1145/3746027.3762020

[7] Lucía Gómez-Zaragozá, Javier Marín-Morales, Mariano Alcañiz, and Mohammad Soleymani. 2025. Speech and Text Foundation Models for Depression Detection: Cross-Task and Cross-Language Evaluation. In *Interspeech 2025*. 5253–5257. doi:10.21437/Interspeech.2025-1035

[8] Zhaocheng Huang, Julien Epps, Dale Joachim, Brian Stasak, James R. Williamson, and Thomas F. Quatieri. 2020. Domain Adaptation for Enhancing Speech-Based Depression Detection in Natural Environmental Conditions Using Dilated CNNs. In *Interspeech 2020*. 4561–4565. doi:10.21437/Interspeech.2020-3135

[9] Xiaochong Lan, Zhiguang Han, Yiming Cheng, Li Sheng, Jie Feng, Chen Gao, and Yong Li. 2025. Depression Detection on Social Media with Large Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 2155–2171. doi:10.18653/v1/2025.emnlp-industry.151

[10] Bubai Maji, Monorama Swain, Shazia Nasreen, Debabrata Majumdar, Rajlakshmi Guha, Aurobinda Routray, and Anders Søgaard. 2025. A Study on The Impact of Foundation Models on Automatic Depression Detection from Speech Signals. In *Interspeech 2025*. 5258–5262. doi:10.21437/Interspeech.2025-1789

[11] Vijay Ravi, Jinhan Wang, Jonathan Flint, and Abeer Alwan. 2022. A Step Towards Preserving Speakers' Identity While Detecting Depression Via Speaker Disentanglement. In *Interspeech 2022*. 3338–3342. doi:10.21437/Interspeech.2022-10798

[12] Kayalvizhi S, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C. 2022. Findings of the Shared Task on Detecting Signs of Depression from Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 331–338. doi:10.18653/v1/2022.ltedi-1.51

[13] Kayalvizhi Sampath, Durairaj Thenmozhi, Bharathi Raja Chakravarthi, Jerin Mahibha C, Kogilavani Shanmugavadivel, and Pratik Anil Rahood. 2023. Overview of the shared task on Detecting Signs of Depression from Social Media Text. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*. 25–30. https://aclanthology.org/2023.ltedi-1.4/

[14] Esaú Villatoro-Tello, S. Pavankumar Dubagunta, Julian Fritsch, Gabriela Ramírez-de-la Rosa, Petr Motlicek, and Mathew Magimai-Doss. 2021. Late Fusion of the Available Lexicon and Raw Waveform-Based Acoustic Modeling for Depression and Dementia Recognition. In *Interspeech 2021*. 1927–1931. doi:10.21437/Interspeech.2021-1288

[15] Bichen Wang, Yixin Sun, Yanyan Zhao, and Bing Qin. 2025. Beyond Snapshots: A Multimodal User-Level Dataset for Depression Detection in Dynamic Social Media Streams. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM '25)*. doi:10.1145/3746027.3758236

[16] Jinhan Wang, Vijay Ravi, and Abeer Alwan. 2023. Non-uniform Speaker Disentanglement For Depression Detection From Raw Speech Signals. In *Interspeech 2023*. 2343–2347. doi:10.21437/Interspeech.2023-2101

[17] Jinhan Wang, Vijay Ravi, Jonathan Flint, and Abeer Alwan. 2022. Unsupervised Instance Discriminative Learning for Depression Detection from Speech Signals. In *Interspeech 2022*. 2018–2022. doi:10.21437/Interspeech.2022-10814

[18] Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. 2020. Affective Conditioning on Hierarchical Attention Networks Applied to Depression Detection from Transcribed Clinical Interviews. In *Interspeech 2020*. 4556–4560. doi:10.21437/Interspeech.2020-2819

[19] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal Vlog Dataset for Depression Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12226–12234. doi:10.1609/aaai.v36i11.21483

[20] Jiajun You, Shuai Wang, Xun Gong, and Xiang Wan. 2025. M3L: A Multi-Modal and Multi-Lingual Depression Detection Framework. In *Interspeech 2025*. 5268–5272. doi:10.21437/Interspeech.2025-329

[21] Lishi Zuo and Man-Wai Mak. 2025. Leveraging Ordinal Information for Speech-based Depression Classification. In *Interspeech 2025*. 484–488. doi:10.21437/Interspeech.2025-638

A DASS-21 Background, Scoring, and Item List

Instrument overview and response scale. DASS-21 is a 21-item self-report instrument measuring Depression, Anxiety, and Stress over the *past week*. Each item is rated on a 4-point ordinal scale: 0 (did not apply to me at all), 1 (applied to me to some degree / some of the time), 2 (applied to me to a considerable degree / a good part of the time), 3 (applied to me very much / most of the time). In AdoDAS (Track A2), we denote the item responses as $d01, d02, \ldots, d21$ with $d_i \in \{0, 1, 2, 3\}$.

Subscale composition and reconstruction. DASS-21 contains three subscales with 7 items each: Depression: {3, 5, 10, 13, 16, 17, 21}; Anxiety: {2, 4, 7, 9, 15, 19, 20}; Stress: {1, 6, 8, 11, 12, 14, 18}. Following

the standard DASS-21 convention, subscale sums are multiplied by 2 to obtain scaled scores:

$$\begin{aligned} \text{Depression} &= (d03 + d05 + d10 + d13 + d16 + d17 + d21) \times 2 \\ \text{Anxiety} &= (d02 + d04 + d07 + d09 + d15 + d19 + d20) \times 2 \\ \text{Stress} &= (d01 + d06 + d08 + d11 + d12 + d14 + d18) \times 2 \end{aligned} \quad (1)$$

Severity cutoffs (scaled scores). Scaled subscale scores can be mapped into five conventional severity levels (Normal/Mild/Moderate/Severe/ Extremely Severe) using standard DASS-21 cutoffs (Table 3).

Table 3: Standard DASS-21 severity cutoffs on scaled subscale scores.

Severity	Depression	Anxiety	Stress
Normal	0-9	0-7	0-14
Mild	10-13	8-9	15-18
Moderate	14-20	10-14	19-25
Severe	21-27	15-19	26-33
Extremely Severe	28+	20+	34+

Table 4: DASS-21 item definitions (0-3 ordinal scale)

Item	Paraphrase	Item	Paraphrase
D01	Difficulty calming down and settling myself.	D02	Mouth felt dry.
D03	Could not experience positive feelings.	D04	Breathing discomfort without exertion.
D05	Struggled to get started on tasks (e.g., studying).	D06	Reacted too strongly to situations.
D07	Experienced trembling (e.g., in my hands).	D08	Used a lot of nervous energy.
D09	Worried about panicking or embarrassment.	D10	Little to look forward to.
D11	Agitated or unable to stay still.	D12	Found it hard to relax.
D13	Felt down-hearted / depressed.	D14	Intolerant of interruptions.
D15	Close to breaking down / losing control.	D16	Could not become enthusiastic about anything.
D17	Felt not worth much as a person.	D18	Easily irritated or touchy.
D19	Unusual heart sensations without exertion.	D20	Scared without an obvious reason.
D21	Life lacked meaning.		

Item list for Track A2. Table 4 summarises the item definitions used in Track A2. *Note:* To avoid reproducing copyrighted questionnaire text verbatim, we provide an English reference paraphrase aligned to the item IDs. For the exact authorised wording of DASS-21 items, please refer to the official DASS-21 form/manual.