Cuet_Neural_Navigators@DravidianLangTech 2026: Depression Detection from Malayalam and Tamil Speech using Self-Supervised Acoustic Models

Shuva Dey, Abir Dey, Sha Newaz Mahmud, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2104001,u2104005,u2004081}@student.cuet.ac.bd,
hasanmurad@cuet.ac.bd

Abstract

Depression detection from speech aims to find signs of depression using behavioral signals. This approach enables early mental health screening and makes it scalable. However, the task is tough because of subtle acoustic cues, differences among speakers, and language-specific patterns. In this work, we introduce our system for the Shared Task on Depression Detection in Dravidian Languages (DD-DL) at DravidianLangTech@ACL 2026. We focus on speech in Tamil and Malayalam. We explore pretrained self-supervised speech encoders, including HuBERT, XLS-R, and Whisper, to identify acoustic patterns related to depression directly from raw audio. Our method combines these models through ensembling to capture different acoustic features. The experiments use stratified evaluation and cross-lingual analysis to check how well the models work across languages. Results show that pretrained acoustic representations effectively capture vocal features of depression, achieving Macro-F1 scores of 0.9058 for Tamil and 0.9396 for Malayalam. However, cross-lingual transfer faces challenges because of phonetic and prosodic differences.

1 Introduction

Depression constitutes a significant worldwide mental health problem which requires immediate detection to enable appropriate treatment and assistance. While textual signals have been widely studied for automatic depression detection, speech offers additional insights through acoustic and paralinguistic elements including prosody, pitch, and speaking rate that reflect emotional states.

However, speech-based depression detection has not received sufficient research attention for low-resource Dravidian languages such as Tamil and Malayalam. The Shared Task on Depression Detection from Malayalam and Tamil Speech Data at DravidianLangTech@ACL 2026 (Jyothish Lal et al., 2026) introduces a benchmark dataset and evaluation framework for this problem, where the goal is to automatically determine whether a speaker shows depressive symptoms from their speech recordings.

We present our system based on pretrained self-supervised speech encoders used as frozen feature extractors, with a lightweight classification head trained on Tamil and Malayalam speech data to identify depression-related acoustic patterns. We examine how different evaluation settings affect model performance and study how speaker differences and language variations create obstacles.

Our research contributions include the following:

- We develop a depression detection system for Tamil and Malayalam speech using pretrained self-supervised acoustic representations.
- We evaluate model robustness under stratified and speaker-disjoint evaluation settings.
- We analyze cross-lingual transferability of depression-related acoustic representations across Dravidian languages.

For implementation details and access to the complete codebase, please refer to our GitHub repository. [1]

2 Related Work

Automatic depression detection has been explored using textual, acoustic, and multimodal signals. Surveys highlight the growing role of AI in identifying depression-related patterns while noting challenges such as annotation noise and cultural variability (Babu and Kanaga, 2021). Transformer-based models have been applied for depression detection in Tamil text (Hemalatha et al., 2025).

[1] https://github.com/Snikdrek/Deppression-Detection-of-Dravidian-speech-Tamil-Malayalam-CUET_Neural_Networks

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 207–211

Speech-based approaches utilize paralinguistic cues such as prosody and pitch variation to capture depressive characteristics. Recent work shows that acoustic features alone can distinguish depressive speech across languages (Binu et al., 2024). The InStant-EMDB corpus introduced English–Malayalam speech data for depression detection (Mathew et al., 2024), while other studies explored multimodal systems (Reddy et al., 2024) and feature-fusion methods for Dravidian speech classification (Kritika et al., 2025).

Shared tasks have further advanced low-resource Dravidian language research (Premjith et al., 2024; Anilkumar et al., 2026), with earlier text-based depression detection tasks demonstrating the effectiveness of transformers (Sampath et al., 2023). The DravidianLangTech@ACL 2026 shared task (Jyothish Lal et al., 2026) provides a standardized speech-based benchmark. Our work builds on these efforts using pretrained self-supervised speech encoders.

3 Dataset

We use the dataset released for the Shared Task on Depression Detection from Malayalam and Tamil Speech Data at DravidianLangTech@ACL 2026 (Jyothish Lal et al., 2026). Depressed utterances were recorded at 16 kHz (2–5 seconds each), while non-depressed recordings were captured at 48 kHz and resampled for consistency. Tables 1 and 2 summarize the data distribution and speaker statistics.

Language	Train	Test	Total
Tamil	1374	160	1534
Malayalam	1888	200	2088

Table 1: Train–test distribution.

Parameter	Malayalam	Tamil
Total Samples	1,888	1,534
Depressed Samples	888	534
Unique Speakers	8	9
Depressed Speakers	3	4
Non-depressed Speakers	5	5

Table 2: Class and speaker statistics across languages.

4 Methodology

Our approach models depression detection as a binary classification task over raw speech signals. Given an input utterance x, the objective is to predict a label $y \in \{0, 1\}$, where 1 denotes depressed and 0 denotes non-depressed speech. Figure 1 illustrates the overall architecture of the proposed framework.

4.1 Preprocessing

All audio recordings were resampled to 16 kHz for consistency. Depressed samples were originally recorded at 16 kHz while non-depressed samples were recorded at 48 kHz; resampling ensured uniform input representation. Each utterance was truncated or padded to 5 seconds to standardize input length. No handcrafted acoustic features were used; models operate directly on raw waveforms.

4.2 Speech Representation Models

We employ pretrained self-supervised speech encoders to obtain acoustic representations:

- **HuBERT** (Hsu et al., 2021): A self-supervised speech representation model trained via masked prediction of discrete speech units.[2]
- **XLS-R** (Babu et al., 2022): A multilingual extension of wav2vec 2.0 pretrained on large-scale cross-lingual speech data.[3]
- **Whisper** (Radford et al., 2023): A transformer-based multilingual speech model trained on large-scale audio data.[4]

Pretrained backbones were initialized with publicly available weights. To reduce overfitting on the small dataset, encoder parameters were frozen and only a lightweight classification head was trained.

4.3 Classification Layer

Hidden representations from each encoder were mean-pooled across the temporal dimension and passed to a linear classification layer. Cross-entropy loss was used for optimization with the AdamW optimizer. Gradient accumulation was applied to simulate larger batch sizes.

4.4 Ensemble Strategy

To improve robustness, predictions from HuBERT, XLS-R, and Whisper were combined via logit-level averaging:

[2] https://huggingface.co/facebook/hubert-base-ls960

[3] https://huggingface.co/facebook/wav2vec2-xls-r-300m

[4] https://huggingface.co/openai/whisper-base

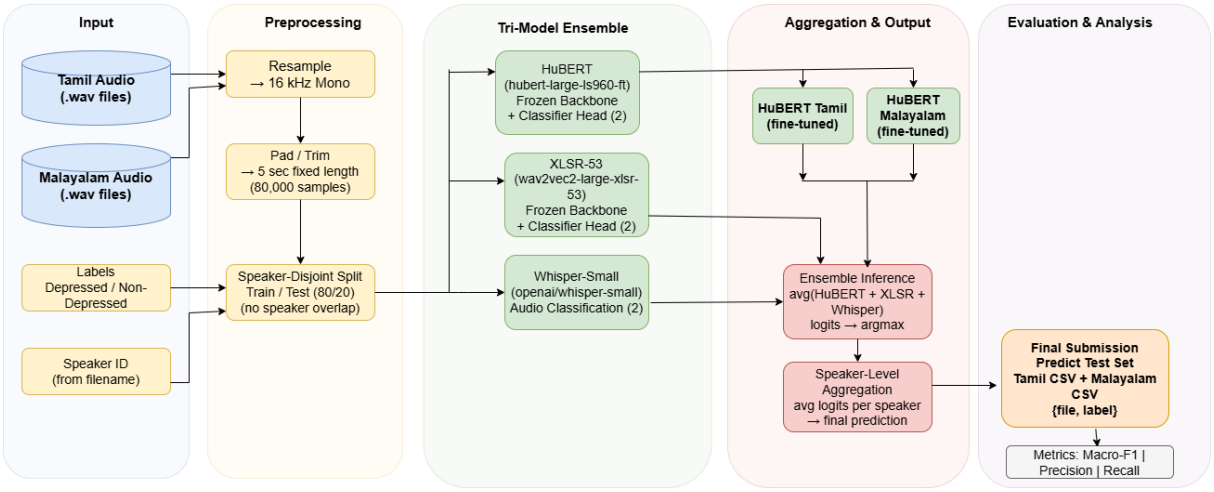

Figure 1: Overview of the proposed depression detection framework.

$$\hat{y} = \frac{1}{3}(z_{\text{HuBERT}} + z_{\text{XLS-R}} + z_{\text{Whisper}})$$

where z denotes model logits. Final predictions were obtained by applying the softmax function to the averaged logits.

4.5 Evaluation Protocol

Experiments were conducted under two evaluation settings:

- **Stratified Split**: Standard 80/20 train–test split preserving label distribution. No fixed random seed was used during splitting, which may cause minor variations across runs.
- **Speaker-Disjoint Split**: Train and test sets were constructed with non-overlapping speakers to evaluate speaker generalization.

Cross-lingual experiments were also performed by training on one language and evaluating on the other. Macro-F1 was used as the primary evaluation metric due to class imbalance.

4.6 Speaker-Level Aggregation

Since depression is a speaker-level condition, utterance-level predictions were aggregated by averaging logits across all utterances from the same speaker before computing final predictions.

4.7 Implementation Details

Models were implemented using `PyTorch` and the `HuggingFace Transformers` library. Training used the AdamW optimizer with a learning rate of 1×10^{-4} for 20 epochs and a per-device batch size of 1. Gradient accumulation with 8 steps produced an effective batch size of 8. Mixed-precision (FP16) training was enabled when supported. Model checkpoints were evaluated each epoch, and early stopping based on validation Macro-F1 retained the best checkpoint. Encoder backbones were frozen and only the classification head was updated.

5 Results and Analysis

In this section, we evaluate our models under monolingual stratified, speaker-disjoint, and cross-lingual settings. Performance is measured using Accuracy, Precision, Recall, and Macro-F1, with Macro-F1 used as the primary metric due to potential class imbalance.

The official shared-task submission achieved Macro-F1 scores of 0.8542 for Tamil and 0.9345 for Malayalam on the test set, ranking **4th** on the shared task leaderboard.

All results reported in the following tables correspond to our experimental runs, which may show slight variations due to stochastic training and data splitting.

5.1 Model Comparison

Table 3 presents the performance of the HuBERT baseline and the proposed ensemble model for Tamil and Malayalam under stratified evaluation. The ensemble model consistently outperforms the single-model baseline in both languages.

For Tamil, the ensemble achieves a Macro-F1 score of 0.9058, improving over the HuBERT baseline by approximately 1.3%. Similarly, for Malayalam, the ensemble obtains the best performance

Experiment	Accuracy	Macro-F1	Precision	Recall
Tamil HuBERT (Stratified GT)	0.8938	0.8930	0.9044	0.8938
Tamil Ensemble (Stratified GT)	**0.9063**	**0.9058**	**0.9141**	**0.9063**
Malayalam HuBERT (Stratified GT)	0.9250	0.9243	0.9359	0.9235
Malayalam Ensemble (Stratified GT)	**0.9400**	**0.9396**	**0.9474**	**0.9388**
Tamil (Speaker-Disjoint)	–	0.8288	0.8202	0.8839
Malayalam (Speaker-Disjoint)	–	0.8157	0.7848	0.9137
Tamil → Malayalam (Cross-lingual)	0.6150	0.6142	0.6149	0.6143
Malayalam → Tamil (Cross-lingual)	0.7688	0.7557	0.8419	0.7688

Table 3: Performance comparison across monolingual, speaker-disjoint, and cross-lingual evaluation settings.

with a Macro-F1 score of 0.9396. These results indicate that combining multiple acoustic models helps capture complementary speech representations, improving depression detection performance.

5.2 Generalization Analysis

To assess robustness across unseen speakers, we conduct speaker-disjoint experiments where training and testing speakers do not overlap. Performance drops compared to stratified evaluation, with Macro-F1 scores of 0.8288 for Tamil and 0.8157 for Malayalam, highlighting challenges in generalizing across speakers. Nevertheless, recall remains relatively high, especially for Malayalam (0.9137), indicating strong sensitivity to depressed speech.

We also evaluate cross-lingual transfer between Tamil and Malayalam. Training on Malayalam and testing on Tamil achieves a Macro-F1 of 0.7557, while the reverse direction achieves 0.6142. This asymmetry suggests that depression-related acoustic patterns learned from Malayalam transfer more effectively to Tamil.

5.3 Error Analysis

Figures 2 and 3 present the confusion matrices for the stratified evaluation setting. For Malayalam, the model correctly classifies 102 non-depressed and 83 depressed samples, while for Tamil it correctly predicts 78 non-depressed and 67 depressed samples, indicating strong class separation.

Most misclassifications occur near the boundary between depressed and non-depressed speech, likely due to subtle acoustic similarities between certain utterances.

6 Conclusion

We presented a speech-based depression detection system for Tamil and Malayalam in the DravidianLangTech@ACL 2026 Shared Task. Using pretrained self-supervised speech encoders (HuBERT, XLS-R, and Whisper), our approach learns

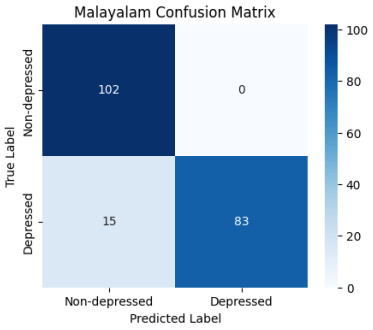

Figure 2: Confusion matrix for Malayalam under stratified evaluation.

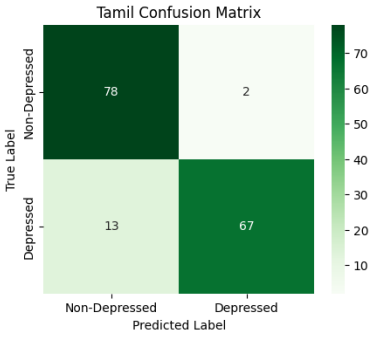

Figure 3: Confusion matrix for Tamil under stratified evaluation.

depression-related acoustic patterns directly from raw speech.

The ensemble model achieved strong monolingual performance, reaching Macro-F1 scores of 0.9058 for Tamil and 0.9396 for Malayalam. The evaluation results show that both speaker-disjoint and cross-lingual tests produced lower scores, which demonstrate the system's difficulties with speaker generalization and its ability to handle different languages.

The study demonstrates that multilingual speech models pre-trained on multiple languages work effectively for depression detection in low-resource environments while showing that existing systems need better performance with various speakers and languages.

Limitations

This work is limited by the relatively small dataset and the small number of unique speakers, which affects generalization, particularly in speaker-disjoint settings. The collected recordings used short utterances which were produced in controlled environments and do not represent the full range of real-world conversational variations.

The training process used frozen pretrained encoders to prevent overfitting, which created a problem for adapting to specific language acoustic features. The performance gaps between Tamil and Malayalam indicate existing transferability problems between the two languages.

Ethics Statement

Depression detection from speech involves sensitive mental health data. The proposed system is not designed for clinical diagnosis but functions as a professional supervision screening tool.

The actual deployment requires organizations to evaluate three main factors including the risks of misclassifying data and the need to protect user privacy and the possibility of demographic biasing. Responsible use requires informed consent from participants, clear communication of the system's limitations, and qualified professional oversight at every stage of deployment.

References

A. Anilkumar, G. Jyothish Lal, B. Premjith, and B. R. Chakravarthi. 2026. Dravlangguard: A multimodal approach for hate speech detection in dravidian social media. In *Speech and Language Technologies for Low-Resource Languages*, volume 2656 of *Communications in Computer and Information Science*, Cham. Springer.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Gober, Romi Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Proceedings of Interspeech*.

Nirmal Varghese Babu and E Grace Mary Kanaga. 2021. Sentiment analysis in social media data for depression detection using artificial intelligence: A review. *Journal of Healthcare Engineering*, 2021.

Sona Binu, Jismi Jose, Fathima Shimna K V, and 1 others. 2024. Language-agnostic analysis of speech depression detection. *arXiv preprint arXiv:2409.14769*.

M. Hemalatha, R. Varshini, Devananda Anil, and E. Janani. 2025. Towards inclusive AI: Deep learning based depression detection with transformers RNN-LSTM. *Journal of Emerging Technologies and Innovative Research*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Benber, and Mohamed Abdelrahman. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

G. Jyothish Lal, B. Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Thenmozhi Durairaj, and Prasanna Kumar Kumaresan. 2026. Shared task on depression detection from malayalam and tamil speech data. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

A. Kritika, S. Meenakshy, Arya Palackal Shijish, Riya Rajeev, and G. Jyothish Lal. 2025. Dravimood: Speech-based depression classification in dravidian languages using feature fusion and deep learning. In *Proceedings of the Fourth International Conference on Speech and Language Technologies for Low-Resource Languages (SPELLL 2025)*.

Anjali Mathew, Raniya, Harsha Sanjan, Amjith S B, and 1 others. 2024. Instant-emdb: A multi model spontaneous english and malayalam speech corpora for depression detection. In *Proceedings of the IEEE Conference*.

B. Premjith, G. Jyothish, V. Sowmya, and B. Bharathi. 2024. Findings of the shared task on multimodal social media data analysis in dravidian languages (msmda-dl)@dravidianlangtech 2024. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 56–61.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

B Surekha Reddy, Jishitha Kondaveti, V Akshaya Bhavani, and P Aishwarya. 2024. Development of a depression detection system using speech and text data. EasyChair Preprint.

Kayalvizhi Sampath, Durairaj Thenmozhi, Bharathi Raja Chakravarthi, and 1 others. 2023. Overview of the shared task on detecting signs of depression from social media text. In *Proceedings of the Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.