Shared Task on Depression Detection from Malayalam and Tamil Speech Data

Jyothish Lal G[1], Premjith B[1], Bharathi Raja Chakravarthi[2], Saranya Rajiakodi,[3] Durairaj Thenmozhi[4], Prasanna Kumar Kumaresan[2]

[1]Amrita School of Artificial Intelligence, Coimbatore,
Amrita Vishwa Vidyapeetham, India,
[2]Unit for Inclusive AI, Data Science Institute, University of Galway, Ireland
[3]Central University of Tamil Nadu, Thiruvarur, Tamil Nadu, India,
[4]Department of Computer Science and Engineering,
Sri Sivasubramaniya Nadar College of Engineering,
Kalavakkam, Tamil Nadu, 603110, India,

Correspondence: g_jyothishlal@cb.amrita.edu

Abstract

Depression is one of the most common mental health problems in the world. It affects a person's emotions, thinking, energy levels, and daily life. Early detection of depression is very important to provide timely support and treatment. While many studies focus on identifying depression from text, speech also carries important emotional and psychological signals that are often not fully explored. This paper presents an overview of the shared task on Depression Detection in Dravidian Languages (DD- DL). The task focuses on identifying signs of depression from speech data in two low-resource Dravidian languages: Tamil and Malayalam. Participants were provided with curated training datasets and were asked to build systems to classify speech samples as Depressed or Non-Depressed. The shared task includes two subtasks: (1) Depression detection in Tamil and (2) Depression detection in Malayalam. Participants applied various machine learning and deep learning approaches to model the acoustic and linguistic characteristics of speech. All submissions were evaluated using the macro-F1 score, which ensures fair performance measurement across classes.

1 Introduction

Depressive disorder (often referred as Depression) is a common mental disorder (Goldwaser and Aaronson, 2023). It affects the mood and thinking capability of a person. Depression is basically different from the normal mood changes or feelings that is observable in daily life, which last for few hours only (Almaghrabi et al., 2023). Basic symptoms of depression include loss of interest in activities, poor concentration, sleep interruptions, suicidal tendency, and so on, for long periods of time. As per the findings by World Health Organization (WHO), around 332 million people in the world have depression (WHO, 2025). Hence, early diagnosis and assessment of depression are improving a person's quality of life (He and Cao, 2018).

From a clinical perspective, the diagnosis of depression mainly depends on clinician judgment (Kroenke et al., 2001). But it shows limited reliability due to variability in symptoms and subjective interpretation (Jiang et al., 2017; Hong et al., 2021). Also, limited resources and lack of trained professionals affect accurate diagnosis and monitoring for depression (Almaghrabi et al., 2023). This highlights the need for objective and reliable biomarkers for depression (Low et al., 2020). In the last decade, audio and text modality have been increasingly used as reliable biomarkers in depression detection, to complement the clinical assessments (Li et al., 2025). It is observed that these modalities provide promising cues for assessing the mental conditions and emotional states of a patient in connection with depression.

Studies have shown that depressive person exhibit distinct linguistic patterns and can therefore be extracted using suitable Natural language Processing (NLP) techniques for better assessment (Rathner et al., 2018; Trifu et al., 2017; Leis et al., 2019). Similarly, the speech data acquired from a speaker or patient carries rich information about the emotive/psychological and physiological state (Cummins et al., 2015). Since speech is the natural way of communication and can be collected non-invasively, automatic analysis of voice also offers a promising route for depression detection.

The acoustic analysis of depressed speech has

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 114–123

focused on several categories of bio-acoustic features. This include voice source measures, spectral features, prosodic parameters such as pitch, intonation, speaking rate, and articulatory formants (France et al., 2000). Voice source features such as jitter and shimmer have received considerable attention. Jitter measures cycle-to-cycle variability in fundamental frequency, while shimmer measures variability in amplitude. These parameters reflect the stability of vocal fold vibration. Several studies have reported increased jitter and shimmer values in depressed individuals, particularly in those with more severe symptoms (Low et al., 2020; Quatieri and Malyska, 2012; Silva et al., 2024). However, results are not fully consistent across studies. Some have reported negative correlations or task-dependent effects. Jitter and shimmer are more reliably measured in sustained vowels or steady phonation segments, and their reliability may decrease in spontaneous or highly variable speech (Franca, 2012). Spectral features have also been widely examined. Depression has been associated with shifts in spectral energy distribution. Some studies report increased energy in higher frequency bands, while others observe reduced energy in high-frequency regions and relatively stronger low-frequency components (Kiss and Vicsi, 2017). Mel-Frequency Cepstral Coefficients (MFCCs), which are widely used in speech processing, have also been explored in depression detection. Several studies have reported that specific MFCC coefficients differ between depressed and non-depressed speakers (Wang et al., 2019; Taguchi et al., 2018; Rejaibi et al., 2022). However, the particular coefficients identified as significant vary across studies. Differences in language, speech task type, recording conditions, and participant demographics may explain these inconsistencies. Prosodic features, including pitch, intensity, speaking rate, and pause structure, are among the most intuitively relevant markers. Many studies report reduced mean pitch and reduced pitch variability in depressed individuals (Almaghrabi et al., 2023).

Building on these acoustic observations, numerous studies have applied machine learning techniques to classify depressed versus non-depressed speech and to estimate depression severity. Early approaches used classical classifiers such as logistic regression, Gaussian Mixture Models (GMM), Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Random Forests (Long et al., 2017; Espinola et al., 2021; Jiang et al., 2017). Feature extraction is commonly performed using open-source tools such as openSMILE, Praat, COVAREP, and Voicebox. Reported classification accuracies generally range from 70 % to 85 %, depending on dataset size, feature selection, and evaluation methodology. More recent research has explored deep learning architectures, including Convolutional Neural Networks (CNNs) applied to spectrograms, CNN-LSTM models, and attention-based architectures (He and Cao, 2018; Kritika A, 2025; Abhinav et al., 2026). Despite encouraging results, several challenges remain. Demographic factors such as age and gender, as well as language differences and medication status, can significantly influence acoustic measures. Most research has focused on English datasets, with limited exploration of low-resource languages, including Dravidian languages. The scarcity of publicly available, well-annotated depression speech corpora in low-resource Dravidian languages remains as a challenge. Hence, the shared task on depression detection from Malayalam and Tamil speech data, held at DravidianLangTech@ACL 2026, is a step towards advancing the depression detection research in these low-resource languages.

2 Task Description

This shared task on depression detection using speech focuses on two major Dravidian languages: Malayalam and Tamil. The task aims to encourage research on identifying depressive indicators from speech signals in low-resource linguistic settings. Accordingly, the shared task is divided into two sub-tasks as follows:

Task 1: Depression detection in Malayalam

Task 2: Depression detection in Tamil

Participants are provided with curated speech datasets for both languages. Each audio sample is labeled as either Depressed (D) or Non-Depressed (ND). The objective is to develop automatic systems that can analyze the acoustic and prosodic characteristics of the speech recordings and accurately classify them into the appropriate category.

The dataset includes controlled recordings with varying articulation patterns, enabling participants to explore handcrafted acoustic features, deep learning-based speech representations, or hybrid approaches.

Model performance will be evaluated using the macro-F1 score, a standard metric for classification tasks that ensures balanced evaluation across

Parameter	Malayalam	Tamil
Total Samples	1,888	1,534
Depressed	888	534
Non-Depressed	1,000	1,000
Train Samples	1,688	1,374
Test Samples	200	160
Unique Speakers	8	9
Depressed	3	4
Non-Depressed	5	5

Table 1: Key corpus statistics by language and class.

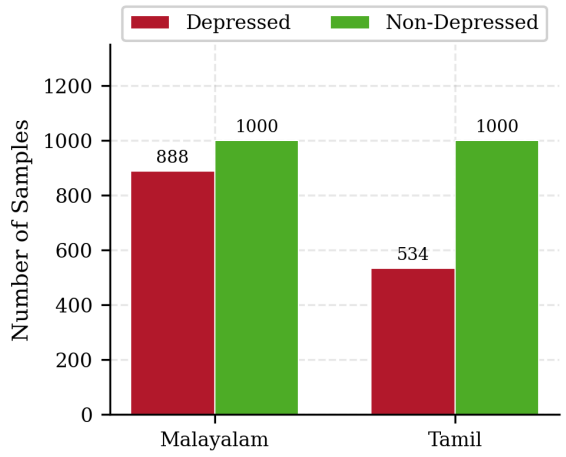

Figure 1: Sample counts by language and class (Depressed vs. Non-Depressed).

classes, particularly in scenarios with class distribution differences.

This shared task provides a benchmark platform to advance research in speech-based mental health assessment for Dravidian languages.

3 Dataset Description

Table 1 summarises the key statistics of the dataset broken down by language and class. This dataset was developed for research on automatic depression detection using speech recordings in two South Indian languages: Malayalam and Tamil. All depressed speech samples were collected in a controlled room environment with consistent acoustic settings to maintain low background noise and uniform recording quality. For the depression class, each sentence was recorded up to a maximum of four times by the same speaker with slight variations in articulation. These utterances have an average duration of 2–5 seconds and were recorded at a sampling rate of 16 kHz. In contrast, the non-depressed class consists of neutral utterances recorded at 48 kHz. In the file naming convention used, 'D' denotes a depressed sample, while any other alphabetic character (A, F, S, etc., with or without numbers) represents a unique speaker identity. Repeated versions of the same sentence are indicated at the end of the filename using numerical (1, 2, 3...) or alphabetical (a, b, c...) suffixes. Non-depressed speakers are labeled ND1, ND2, etc. The corpus comprises 3,422 total utterances across 17 unique speakers in a binary (Depressed / Non-Depressed) classification framework.

Figures 1–3 visualise the sample distribution across languages, classes, and data splits.

4 System Description

This section discusses the descriptions of the systems submitted by the participating teams. There are 62 registrations for the shared task. However, 6 teams submitted their predictions in the presceibed format through the provided Google form.

4.1 SERENE

In this work, the team tackle depression detection as a supervised binary classification problem using both text and audio approaches across Tamil and Malayalam. They employ a fine-tuned XLM-RoBERTa-base model for text analysis, trained on transcribed speech to identify depressive language patterns. For audio analysis, they utilize two methods: extracting 33 low-level acoustic features with Librosa for traditional classifiers and converting raw audio into Mel-spectrograms to train a CNN for learning time–frequency patterns linked to depression. The models are trained on labeled audio samples, predicting binary labels for unseen cases. This combination of textual and acoustic modeling effectively captures linguistic and paralinguistic indicators of depression.

4.2 TriVector

They explore speech-based depression detection through a binary classification approach that integrates handcrafted acoustic features with self-supervised speech representations. Audio recordings are standardized to 16 kHz prior to training two models in parallel: one utilizing Mel-Frequency Cepstral Coefficients (MFCCs) with a lightweight neural classifier, and the other using a fine-tuned Wav2Vec 2.0 architecture for capturing contextual speech representations. To ensure unbiased results, they implement a speaker-independent data splitting strategy and address class imbalance

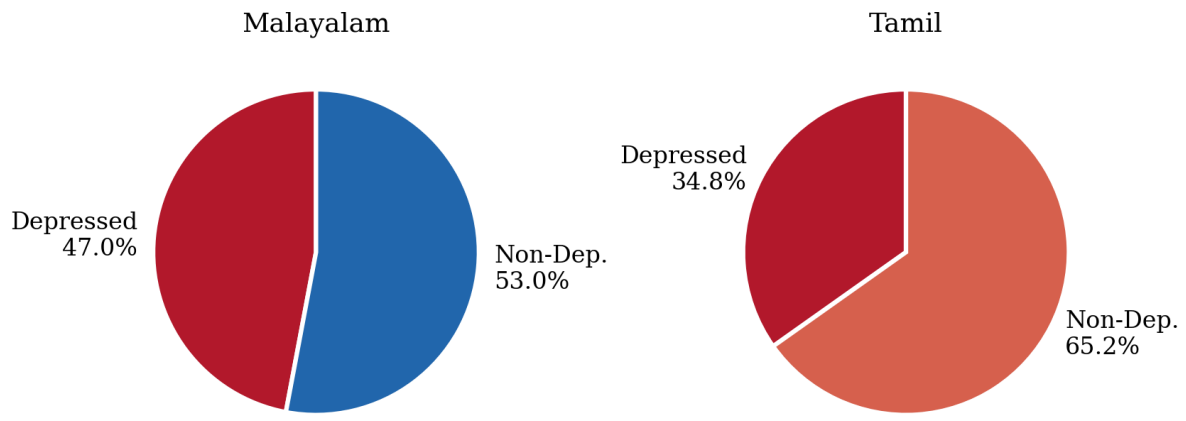

Figure 2: Class balance (Depressed / Non-Depressed) per language.

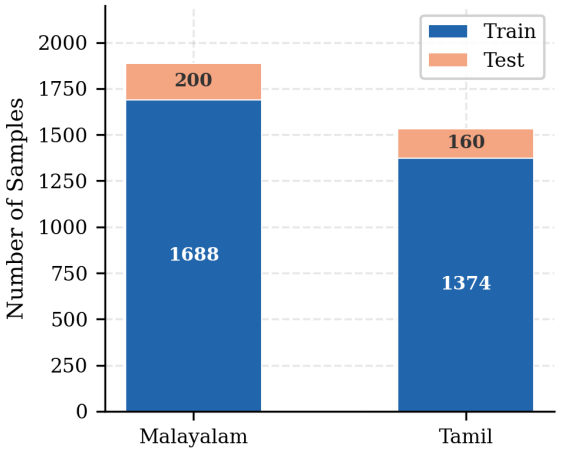

Figure 3: Train / test split per language.

via weighted loss functions. Overfitting is mitigated through techniques such as encoder freezing and early stopping. Their inference combines the outputs of both models via a weighted ensemble determined by validation F1 scores, with final predictions assessed using macro-F1 as the key metric.

4.3 Cuet_Neural_Navigators

The submission utilizes a structured machine learning pipeline for detecting depression from speech using transformer-based self-supervised audio representations. Speech recordings in Tamil and Malayalam were processed into compatible acoustic and prosodic features. Pretrained multilingual models such as Wav2Vec2-XLSR, HuBERT, and Whisper were fine-tuned for binary depression classification with speaker-aware data splits to minimize leakage. To counter overfitting, only upper encoder layers and classification heads were fine-tuned. Model performance was assessed using the macro F1-score. During inference, predictions were generated from unseen samples, and outputs from multiple models were combined through an ensemble strategy for enhanced robustness. The resulting system provides reliable depression predictions for each test utterance.

4.4 CREST

They present a language-agnostic speech-based framework for depression detection utilizing the pretrained Wav2Vec2 XLS-R (300M) model, processing raw audio waveforms to facilitate high-level speech representations without handcrafted features. Audio samples were resampled to 16 kHz and normalized, with a strict speaker-independent train-validation split to prevent overlap. The XLS-R encoder was frozen while a lightweight classification head was trained for binary classification, employing weighted cross-entropy loss to address class imbalance. The classifier was optimized with the Adam optimizer for enhanced convergence and to mitigate overfitting. Model evaluation focused on accuracy and F1-score, prioritizing balanced classification. The trained model was directly applied to unseen audio for predictions, outputting standardized CSV files. This methodology was effectively implemented across Tamil and Malayalam datasets, underscoring its scalability and cross-lingual applicability in speech-based mental health assessments.

4.5 KEC Launchpad

This system uses MFCC feature extraction combined with CNN-LSTM hybrid architecture. They

extract 40 MFCC coefficients plus delta features to capture vocal characteristics, then process them through convolutional layers for pattern detection and LSTM layers for temporal analysis. The model demonstrates strong cross-language depression detection capabilities, performing effectively on both Tamil and Malayalam speech datasets without language-specific modifications.

4.6 CODEX

In this work, the team implemented a speech-based classification approach. Audio recordings were preprocessed into mono at 16 kHz for uniformity, and a pretrained wav2vec 2.0 model extracted feature representations from the audio. These features were aggregated to create fixed-length vectors, which were used to train a Logistic Regression classifier on labeled depressed and non-depressed speech samples. The models, specific to each language, were then applied to unlabeled test audio for predictions, which were formatted into CSV files as per submission guidelines. System performance was evaluated using macro-averaged precision, recall, and F1-score.

5 Submission Analysis

Table 2 summarizes the approached used by the teams participated in this task.

Tables 3 and 4 present the official results of the Malayalam and Tamil tracks. Team SERENE achieves perfect scores across both languages, which shows the flawless classification and complete precision–recall balance. In contrast, CODEX shows substantial precision–recall imbalance, especially in Malayalam, where high precision but very low recall indicates a conservative classifier that misses many depressed cases.

SERENE achieved perfect F1 on both tracks, which demonstrates complete language-agnosticism in its system design. This likely reflects the use of multilingual pre-trained speech representations (e.g., MMS, XLS-R, or multilingual HuBERT) combined with minimal language-specific fine-tuning.

Three systems showed meaningfully higher performance on Tamil than on Malayalam. CREST exhibited the largest Tamil advantage: +0.1258 Macro F1 points (Tamil F1 = 0.9875 vs. Malayalam F1 = 0.8617). CODEX showed a +0.2324 Tamil advantage (0.5116 vs. 0.2792). KEC Launchpad achieved a marginal Tamil advantage.

TriVector and Cuet_Neural_Navigators both demonstrated higher performance on Malayalam. TriVector's Malayalam F1 (0.9950, Run 1) exceeded its Tamil F1 (0.8861, Run 2) by 0.1089 points - the largest Malayalam advantage among all teams. Cuet_Neural_Navigators showed a smaller but consistent Malayalam advantage of 0.0803 points (0.9345 vs. 0.8542).

Table 5 summarizes the descriptive statistics of Macro F1 scores for both language tracks. The mean Macro F1 for Malayalam is 0.8442 versus 0.8732 for Tamil, which indicates that the Tamil track, on average, was more tractable for participating systems by approximately 0.029 F1 points. The median Malayalam F1 is 0.9648 (between TriVector and Cuet_Neural_Navigators) versus 0.9368 for Tamil (between CREST and TriVector). The inversion of mean and median ordering between languages reflects different outlier structures: Malayalam has a single very low-performing outlier (CODEX at 0.2792) that pulls the mean down substantially, while Tamil has two outliers dragging it down but from a higher floor.

The standard deviation of F1 scores is significantly higher for Malayalam (0.2704) compared to Tamil (0.1900). This larger spread in Malayalam confirms greater inter-system variability and suggests that the Malayalam task created a harder differentiation challenge - some systems excelled while others failed substantially. The range (max-min) of 0.7208 in Malayalam versus 0.4884 in Tamil reinforces this finding.

Both tracks exhibit a ceiling behavior: the maximum F1 is 1.0 in both cases. However, the floor differs: 0.2792 for Malayalam versus 0.5116 for Tamil. The raised Tamil floor implies that even the weakest valid system managed above-chance classification on Tamil, whereas the weakest Malaysian system performed near or below chance-level for the positive class.

In addition to the metrics, discussed above, we computed the absolute F1–Accuracy divergence (|F1-Acc|) and Precision–Recall gap. The |F1-Acc| captures the extent to which a system's accuracy is being inflated by predicting the majority class more frequently. A large gap indicates systematic class bias in the prediction. The Precision–Recall gap reveals which type of error a system preferentially commits: a large positive gap means the system over-predicts the negative (non-depressed) class, achieving high precision at the cost of missing many depressed individuals (high false negatives);

Team	Ta/Ml	Fusion / Approach	XLM-R	W2V2	XLS-R	HuBERT	Whisper	MFCC	CNN-LSTM
SERENE	✓	Late fusion (text + audio)	✓	✗	✗	✗	✗	✗	✗
TriVector	✓	Parallel audio models; weighted ensemble	✗	✓	✗	✗	✗	✓	✗
Cuet_Neural_Navigators	✓	Multi-SSL fine-tuning; ensemble	✗	✗	✓	✓	✓	✗	✗
CREST	✓	Frozen XLS-R encoder + classifier head	✗	✗	✓	✗	✗	✗	✗
KEC Launchpad	✓	MFCC-based deep hybrid model	✗	✗	✗	✗	✗	✓	✓
CODEX	✓	W2V2 feature extraction + Logistic Regression	✗	✓	✗	✗	✗	✗	✗

Table 2: Comparison of submitted systems for depression detection.

Rank	Team	Run	Macro F1	Accuracy	Macro Prec.	Macro Recall
1	SERENE	Run 1	1.0000	1.0000	1.0000	1.0000
2	KEC Launchpad	Run 1	0.9950	0.9950	0.9951	0.9949
3	TriVector	Run 1	0.9950	0.9950	0.9949	0.9951
4	Cuet_Neural_Navigators	Run 2	0.9345	0.9350	0.9435	0.9337
5	CREST	Run 1	0.8617	0.8650	0.8953	0.8622
6	CODEX	Run 1	0.2792	0.2900	0.5621	0.1926

Table 3: Official results of Malayalam track

Rank	Team	Run	Macro F1	Accuracy	Macro Prec.	Macro Recall
1	SERENE	Run 1	1.0000	1.0000	1.0000	1.0000
1	KEC Launchpad	Run 1	1.0000	1.0000	1.0000	1.0000
2	CREST	Run 1	0.9875	0.9875	0.9878	0.9875
3	TriVector	Run 2	0.8861	0.8875	0.9082	0.8875
4	Cuet_Neural_Navigators	Run 2	0.8542	0.8562	0.8775	0.8562
5	CODEX	Run 1	0.5116	0.6312	0.6667	0.4208

Table 4: Official results of Tamil track

a negative gap means the system is more aggressive in flagging depression, achieving higher recall at the cost of some false alarms. CODEX team's submission in Tamil track is the most extreme case. Its accuracy of 0.6312 is 11.96 percentage points higher than its F1 of 0.5116. This divergence of 0.1196 represents an accuracy inflation of 23.4% relative to F1, which means accuracy overstates the true system quality by nearly a quarter.

The Figures 4 and 5 show radar charts for Malayalam and Tamil to plot four classification metrics (Macro F1, Accuracy, Precision, and Recall) for the top five systems. The radial scale is deliberately bounded between 0.80 and 1.00 to make differences between competitive systems visible. CODEX is excluded from both plots as a statistically significant outlier (Z = -2.09 on Malayalam, Z = -1.90 on Tamil), whose low scores would collapse the scale and obscure the performance profiles of the other teams. Each team appears as a coloured polygon: a perfectly balanced system produces a symmetric diamond pressing against the outer edge, while any inward recession on a particular axis reveals an imbalance in that metric.

In Malayalam track, SERENE and KEC Launchpad form near-perfect outer-boundary diamonds, reflecting their exceptional balance across all four metrics (F1 = 1.00 and 0.9950 respectively) with essentially no precision-recall gap. TriVector overlaps heavily with KEC Launchpad and is a statistical tie. Cuet_Neural_Navigators shows the first visible asymmetry - a slight pull toward the recall axis from mild precision dominance (gap = +0.0098), meaning it misses slightly more true cases than it raises false alarms. CREST has the most distorted polygon: precision (0.895) visibly exceeds recall (0.862), corresponding to a 13.78% false negative rate — nearly one in seven depressed speakers would be missed. Crucially, F1 and Accuracy are nearly perfectly aligned for all five systems, which confirms that none is gaming accuracy through class-imbalanced predictions.

In Tamil task, SERENE and KEC Launchpad again achieve perfect outer-boundary diamonds (all metrics = 1.00). The most striking change from the Malayalam plot is CREST: its asymmetric shape on Malayalam transforms into a near-perfect diamond on Tamil (F1 = 0.9875, P-R gap = 0.0003), which captures its dramatic rank jump from 5th on Malayalam to 2nd on Tamil. TriVector's polygon moves in the opposite direction - it becomes precision-leaning on Tamil (precision = 0.908, recall = 0.888, gap = +0.021), reversing its recall-dominant shape from Malayalam. This reversal is an architectural signal: the model makes qualitatively different errors depending on the language.

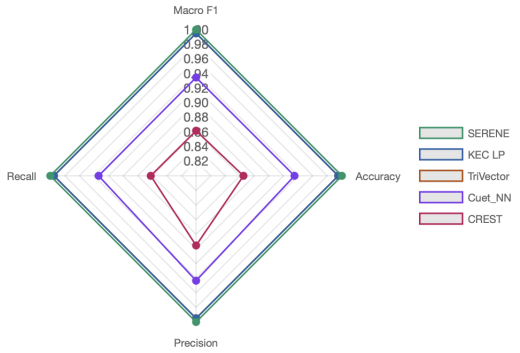

Figure 4: Radar (spider) charts showing the per-metric performance profiles of the top-five systems on the Malayalam track

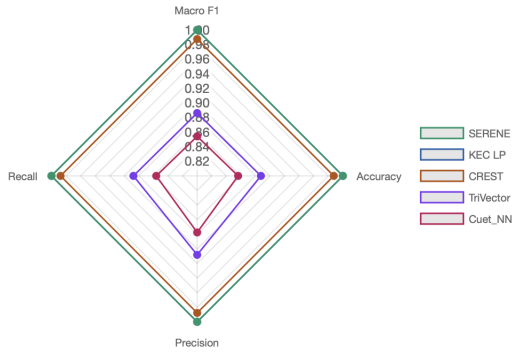

Figure 5: Radar (spider) charts showing the per-metric performance profiles of the top-five systems on the Tamil track

Table 5 shows the descriptive statistics over Macro F1-scores for all valid systems in each track reveal the overall difficulty level, performance spread, and central tendency of the competition. The mean Macro F1 for Malayalam is 0.8442 versus 0.8732 for Tamil, which indicates that the Tamil track, on average, was more tractable for participating systems by approximately 0.029 F1 points. The median Malayalam F1 is 0.9648 (between TriVector and Cuet_Neural_Navigators) versus 0.9368 for Tamil (between CREST and TriVector). The inversion of mean and median ordering between languages reflects different outlier structures: Malayalam has a single very low-performing outlier (CODEX at 0.2792) that pulls the mean down substantially, while Tamil has two outliers dragging it down but from a higher floor. The standard devia-

Statistic	Malayalam (F1)	Tamil (F1)
Mean	0.8442	0.8732
Median	0.9647	0.9368
Std Dev	0.2574	0.1715
Min	0.2792	0.5116
Max	1.0000	1.0000
Range	0.7208	0.4884

Table 5: Descriptive Statistics of Macro F1-scores by Language Track

tion of F1 scores is notably higher for Malayalam (0.2704) compared to Tamil (0.1900). This larger spread in Malayalam confirms greater inter-system variability and suggests that the Malayalam task created a harder differentiation challenge - some systems excelled while others failed substantially. The range (max-min) of 0.7208 on Malayalam versus 0.4884 on Tamil reinforces this finding. Both tracks exhibit ceiling behavior: the maximum F1 is 1.0 in both cases. However, the floor differs: 0.2792 for Malayalam versus 0.5116 for Tamil. The raised Tamil floor implies that even the weakest valid system managed above-chance classification on Tamil, whereas the weakest Malaysian system performed near or below chance-level for the positive class.

6 Conclusion

This paper presented an overview of the Shared Task on Depression Detection in Dravidian Languages at DravidianLangTech@ACL 2026. The task focused on binary classification of depression from speech data in two low-resource languages: Malayalam and Tamil. A total of 3,422 utterances from 17 speakers were released across both tracks, and system performance was evaluated using macro-F1 to ensure balanced assessment across classes. The shared task attracted 62 registrations, with six teams submitting valid runs. The participating systems explored a diverse range of approaches, including handcrafted acoustic features (MFCCs), CNN-LSTM hybrid models, pretrained self-supervised speech representations (Wav2Vec2, XLS-R, HuBERT, Whisper), and multimodal fusion combining text and audio. The results show that self-supervised and multilingual pretrained models are highly effective for this task. Both tracks exhibited strong top-end performance, with multiple systems achieving near-perfect or perfect macro-F1 scores. However, descriptive statistics reveal meaningful differences between languages. The Tamil track showed slightly higher mean performance and a higher performance floor, while the Malayalam track displayed greater inter-system variability and a wider performance range. These findings suggest that Malayalam posed a more sensitive differentiation challenge, where system design choices had a stronger impact on outcomes.

7 Limitations

This section discusses the limitations of this shared task. Here, the dataset size is relatively small, consisting of 3,422 utterances collected from only 17 speakers, which limits speaker diversity and may affect the generalizability of the developed models. Models trained on such a limited set of speakers may capture speaker-specific characteristics rather than robust depression-related speech patterns. Additionally, the task formulation is limited to binary classification (depressed vs. non-depressed) and does not account for different levels of depression severity that are relevant in clinical assessment.

8 Ethical Considerations

The dataset used in this shared task consists of elicited speech recordings rather than speech collected from clinically diagnosed depression patients. As a result, the data collection process does not involve sensitive clinical information, personal health records, or vulnerable patient groups. The recordings were obtained from speakers who produced predefined utterances in controlled settings, which minimizes potential privacy risks and ethical concerns typically associated with mental health datasets. Since no real patient data or diagnostic information was collected, the dataset avoids many ethical challenges related to informed consent, confidentiality, and the handling of sensitive medical data. However, it is important to acknowledge that systems developed using such elicited data should be interpreted carefully and should not be directly used for clinical diagnosis without proper validation using ethically collected clinical datasets.

Acknowledgments

This work was conducted with the financial support from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2), supported in part of Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

Vura Abhinav, Bhaswanth Reddy Indukuri, MS Karthik, Sai Praneeth Reddy Alavalapati, Ramisetty Lakshmi Venkat, and G Jyothish Lal. 2026. Vision transformer-based audio analysis for depression detection: A human factor in reliable cps. In *Reliability in Cyber-Physical Systems: The Human Factor Perspective*, pages 65–81. Springer.

Shaykhah A. Almaghrabi, Scott R. Clark, and Mathias Baumert. 2023. Bio-acoustic features of depression: A review. *Biomedical Signal Processing and Control*, 85:105020.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech communication*, 71:10–49.

Caroline Wanderley Espinola, Juliana Carneiro Gomes, Jessiane Mônica Silva Pereira, and Wellington Pinheiro Dos Santos. 2021. Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study. *Research on Biomedical Engineering*, 37(1):53–64.

Maria Claudia Franca. 2012. Acoustic comparison of vowel sounds among adult females. *Journal of Voice*, 26(5):671–e9.

Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and MJItoBE Wilkes. 2000. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):829–837.

Eric L. Goldwaser and Scott T. Aaronson. 2023. *Depressive Disorders*, pages 531–567. Springer International Publishing, Cham.

Lang He and Cui Cao. 2018. Automated depression analysis using convolutional neural networks from speech. *Journal of biomedical informatics*, 83:103–111.

Ran Ha Hong, Jill K Murphy, Erin E Michalak, Trisha Chakrabarty, Zuowei Wang, Sagar V Parikh, Larry Culpepper, Lakshmi N Yatham, Raymond W Lam, and Jun Chen. 2021. Implementing measurement-based care for depression: practical solutions for psychiatrists and primary care physicians. *Neuropsychiatric Disease and Treatment*, pages 79–90.

Haihua Jiang, Bin Hu, Zhenyu Liu, Lihua Yan, Tianyang Wang, Fei Liu, Huanyu Kang, and Xiaoyu Li. 2017. Investigation of different speech types and emotions for detecting depression using different classifiers. *Speech Communication*, 90:39–46.

Gábor Kiss and Klára Vicsi. 2017. Mono-and multilingual depression prediction based on speech processing. *International Journal of Speech Technology*, 20(4):919–935.

Arya Palackal Shijish Riya Rajeev Jyothish Lal G Kritika A, Meenakshy S. 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)*.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Angela Leis, Francesco Ronzano, Miguel A Mayer, Laura I Furlong, and Ferran Sanz. 2019. Detecting signs of depression in tweets in spanish: behavioral and linguistic analysis. *Journal of medical Internet research*, 21(6):e14199.

Yuxin Li, Sinchana Kumbale, Yanru Chen, Tanmay Surana, Eng Siong Chng, and Cuntai Guan. 2025. Automated depression detection from text and audio: A systematic review. *IEEE Journal of Biomedical and Health Informatics*.

Hailiang Long, Zhenghao Guo, Xia Wu, Bin Hu, Zhenyu Liu, and Hanshu Cai. 2017. Detecting depression in speech: Comparison and combination between different speech types. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1052–1058. IEEE.

Daniel M Low, Kate H Bentley, and Satrajit S Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope investigative otolaryngology*, 5(1):96–116.

Thomas F Quatieri and Nicolas Malyska. 2012. Vocal-source biomarkers for depression: A link to psychomotor activity. In *Interspeech*, volume 2, pages 1059–1062.

Eva-Maria Rathner, Julia Djamali, Yannik Terhorst, Björn Schuller, Nicholas Cummins, Gudrun Salamon, Christina Hunger-Schoppe, and Harald Baumeister. 2018. How did you like 2017? detection of language markers of depression and narcissism in personal narratives.

Emna Rejaibi, Ali Komaty, Fabrice Meriaudeau, Said Agrebi, and Alice Othmani. 2022. Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, 71:103107.

Wegina Jordana Silva, Leonardo Lopes, Melyssa Kellyane Cavalcanti Galdino, and Anna Alice Almeida. 2024. Voice acoustic parameters as predictors of depression. *Journal of Voice*, 38(1):77–85.

Takaya Taguchi, Hirokazu Tachikawa, Kiyotaka Nemoto, Masayuki Suzuki, Toru Nagano, Ryuki Tachibana, Masafumi Nishimura, and Tetsuaki Arai. 2018. Major depressive disorder discrimination using vocal acoustic features. *Journal of affective disorders*, 225:214–220.

Raluca Nicoleta Trifu, Bogdan Nemeş, Carolina Bodea-Haţegan, and Doina Cozman. 2017. Linguistic indicators of language in major depressive disorder (mdd). an evidence based research. *Journal of Evidence-Based Psychotherapies*, 17(1).

Jingying Wang, Lei Zhang, Tianli Liu, Wei Pan, Bin Hu, and Tingshao Zhu. 2019. Acoustic differences between healthy and depressed people: a cross-situation study. *BMC psychiatry*, 19(1):300.

WHO. 2025. Depressive disorder (depression). *Available in:< https://www.who.int/news-room/fact-sheets/detail/depression>(accessed 19.12.2025).*