TriVector@DravidianLangTech 2026: Depression Detection from Tamil and Malayalam Speech with Speaker-Independent Evaluation using MFCC and Wav2Vec2

Tahmima Hoque Eid, Fawzia Tabassum, Oarisa Rebayet, Hasan Murad
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
{u2104114, u2104074, u2104129}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

Depression is a major mental health concern that can be reflected through subtle changes in speech patterns, prosody, and vocal characteristics. In low-resource and multilingual settings, depression detection from speech may become particularly more challenging. In this work, we present our system for the Shared Task on Depression Detection from Malayalam and Tamil. We explored both handcrafted acoustic features (MFCC) and pretrained speech representations (Wav2Vec2) for depression detection, along with a simple fusion strategy to examine their complementary strengths. Our observations showed that Wav2Vec2 generalized better for Malayalam, whereas for Tamil, a validation-tuned probability fusion performed best. The final system achieved macro-F1 scores of 99.5% for Malayalam and 88.6% for Tamil, securing 3rd place in both tasks.

1 Introduction

Depression is a common mental health condition characterized by persistent low mood and loss of interest in daily activities. It is also accompanied by fatigue, sleep disturbances, and cognitive difficulties (James et al., 2018). Despite its widespread impact, depression remains overlooked and undertreated even today. A cross-sectional study in India reported that although nearly two-thirds of adults with depression accessed health services, about 97% remained undiagnosed (Teufel et al., 2024). Speech can capture subtle cognitive and emotional changes. Hence, it is a promising non-invasive marker for detecting depressive states. However, the majority of this field's research focuses on high-resource languages, leaving many regional languages unexplored. Although Malayalam and Tamil are widely spoken, they have limited speech-based mental health resources. Multimodal systems can also be effective in this regard, but speech-based approaches are simpler for telehealth and low-resource contexts (Rezaee, 2026).

This study is conducted as part of the Depression Detection in Dravidian Languages (DD-DL) Shared Task (Lal et al., 2026). The organizers have provided a curated set of speech recordings collected under controlled conditions. The task comprises two subtasks[1]:

(i) Depression detection in Tamil.

(ii) Depression detection in Malayalam.

Similar dataset and task formulations are described in the DraviMood study, which explores speech-based depression classification in Dravidian languages using feature fusion and deep learning techniques (Kritika et al., 2025).

To solve this challenge, we developed a depression detection system that integrates handcrafted acoustic features with self-supervised speech representations. We used MFCC features to capture low-level acoustic cues, which provide compact spectral representations of speech and are computationally efficient. We also fine-tuned Wav2Vec 2.0 to learn higher-level patterns from raw audio. Finally, we explored a combined approach by fusing the predictions of both models at the probability level to leverage their complementary strengths. Our pipeline achieved macro-F1 scores of 99.5% for Malayalam and 94.4% for Tamil.

The main contributions of this work are as follows:

- We created speaker-aware, leakage-free splits using stratified grouping for more generalization to unseen speakers.
- A depression detection model was developed using MFCC features and fine-tuned Wav2Vec 2.0.
- We applied validation-tuned probability fusion and threshold calibration to enhance reliability and class balance.

[1] https://www.codabench.org/competitions/11331/

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 429–435

Detailed implementation information is available in our GitHub repository.[2]

2 Related Work

Early studies on speech-based depression detection relied on manually extracted speech features combined with traditional machine learning classifiers. Janardhan and Kumaresh (2022) utilized eGeMAPS features with Fisher score-based feature selection and dynamic ensemble strategies, which improved classification performance.

With the rise of deep learning, researchers shifted towards automatic representation learning from raw and spectral speech signals. Kim et al. (2023) demonstrated improved performance over conventional models by developing a smartphone-based CNN framework using log-Mel spectrograms. Gupta et al. (2024) proposed a deep convolutional attention cascaded LSTM framework for speech-based depression detection, combining feature extraction, optimization-based feature selection, and deep classification. To better capture hierarchical and long-range contextual information,Yin et al. (2023) further explored transformer-driven and parallel CNN designs.

More recent studies have focused on multilingual and low-resource scenarios. Mathew et al. (2024) introduced a bilingual spontaneous speech corpus to analyze emotional valence–arousal patterns across English and Malayalam, while related read-speech analyses were further examined by Daly and Olukoya (2025). Binu et al. (2024) investigated language-agnostic acoustic markers, suggesting that certain vocal biomarkers may generalize across linguistic settings. In low-resource environments, Zhang et al. (2024) utilized transfer learning with pretrained wav2vec 2.0, reporting notable improvements in F1 scores through contextual speech representations. Prior work in Dravidian language analysis has shown that transformer-based models such as multilingual BERT, RoBERTa, and MuRIL effectively capture linguistic patterns for sentiment and abusive content detection (Premjith et al., 2024)

Although advanced deep learning and transfer learning techniques have shown strong performance, most studies often overlook issues such as speaker leakage and data imbalance. In low-resource settings, resource scarcity remains a major hindrance to building stable and generalizable models (Anilkumar et al., 2026). Moreover, prior works rarely explore hybrid approaches combining acoustic and self-supervised features, highlighting the need for more adaptable frameworks.

3 Data Description

We used the dataset released for the Shared Task on Depression Detection from Malayalam and Tamil Speech Data – DravidianLangTech@ACL 2026 (Lal et al., 2026). The recordings were collected under controlled acoustic conditions to ensure consistent quality. Each utterance has an average duration of 2–5 seconds. All dataset statistics are summarized in Table 1.

Dataset Characteristics	Malayalam	Tamil
Training Samples	1688	1374
Test Samples	200	160
Depressed Samples	888	534
Depressed Speakers	3	4
Non-depressed Speakers	5	5
Recording Rate (Depressed)	16 kHz	16 kHz
Recording Rate (Non-depressed)	48 kHz	48 kHz

Table 1: Comprehensive dataset statistics for Malayalam and Tamil depression detection tasks.

4 Methodology

4.1 Audio Preprocessing

All recordings were standardized to 16 kHz mono for consistency. Silence trimming (top_db = 30) was applied in the Tamil pipeline, while near-silent Malayalam recordings were automatically detected and removed. To stabilize feature extraction, we applied amplitude normalization and clipping. We also discarded fully trimmed or empty Tamil recordings.

4.2 Speaker-Aware Leak-Safe Data Splitting

To prevent speaker leakage and reduce overfitting in limited-data settings, speaker identities were parsed from filenames and used for speaker-aware splitting. For Malayalam, StratifiedGroupKFold (5 folds) with speaker/group constraints was used, selecting the most class-balanced fold as validation. For Tamil, a speaker-stratified split ensured no speaker overlap between train and validation sets. Final splits were saved as leak-safe CSV files.

4.3 Overview of Experimental Models

4.3.1 MFCC-Based Acoustic Modeling

We used MFCC features to capture spectral patterns associated with depression-related speech

[2] https://github.com/tahmima114/DD-MT

changes. First, the audio was converted to 16 kHz and then transformed into fixed-size MFCC representations. Malayalam used 20 MFCCs (20×120 from 6-s audio) with train-only time–frequency masking and a 2-layer MLP classifier (serving as a computationally efficient baseline), whereas Tamil used 40 MFCCs (40×320) with silence trimming (top_db = 30) and a lightweight CNN to better capture time–frequency patterns. Finally, class imbalance was handled using weighted loss (and `WeightedRandomSampler` for Tamil). Training used AdamW optimization with early stopping based on validation macro-F1.

4.3.2 Wav2Vec2-Based Representation Learning

Wav2Vec2 was implemented for our model to learn high-level speech patterns directly from raw audio. For Malayalam, a Wav2Vec2-Base model processed 6-second segments with light training-time augmentation (speed, pitch, and noise). The classification layers were trained using class-weighted loss and gradient accumulation based on validation macro-F1. For Tamil, the Wav2Vec2-XLS-R-300M model was fine-tuned in two stages. In stage one, only the classifier was trained, whereas in stage two, the last two transformer layers were unfrozen to enable partial fine-tuning. Comparable hyperparameters are summarized in Table 2.

Hyperparameter	Malayalam	Tamil
Initial Encoder Setting	Frozen	Frozen
Training Strategy	Single-stage	Two-stage fine-tuning
Epochs (Stage 1)	15	8
Epochs (Stage 2)	–	8
Learning Rate (Stage 1)	3×10^{-4}	5×10^{-5}
Learning Rate (Stage 2)	–	1×10^{-5}
Mixed Precision (fp16)	Disabled	Enabled (Stage 1)
Early Stopping Patience	3	2

Table 2: Comparable Wav2Vec2 hyperparameters for Malayalam and Tamil.

4.3.3 Ensemble Decision Fusion

MFCC captures low-level acoustic features like spectral envelope and energy, while Wav2Vec2 learns higher-level phonetic representations from raw audio. Their combination integrates acoustic and phonetic representations for improved modeling. Accordingly, MFCC and Wav2Vec2 predictions were fused at the probability level using a weighted average of softmax outputs, with weights tuned exclusively on the validation set to avoid data leakage. The overall pipeline is illustrated in Fig. 1. For Malayalam, a small set of candidate weights was evaluated using validation macro-F1 to select the optimal fusion weight. For Tamil, a finer grid search was performed to jointly optimize the fusion weight and decision threshold using an F2-score objective, followed by a small threshold adjustment (+0.03) to improve the precision–recall balance.

5 Results and Analysis

This section presents the results of the depression detection task for Malayalam and Tamil speech. We evaluated model performance using precision, recall, and F1-score, emphasizing macro-averaged F1 (Macro-F1) to ensure balanced assessment across classes. The Macro-F1 score is computed as the average of the F1-scores of all classes:

$$F1_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \times P_i \times R_i}{P_i + R_i}$$

Here C denotes the total number of classes, and P_i and R_i represent the precision and recall of class i, respectively.

5.1 Task 1: Depression detection in Tamil

For the Tamil task, the MFCC–CNN model reached a macro-F1 of 77.8%, showing moderate ability to capture depression-related spectral cues. In contrast, the standalone Wav2Vec2 model performed much lower (macro-F1 = 48.1%) due to class imbalance effects and biased predictions. Fusion weights were optimized on the validation set using an F2 objective with threshold calibration. Since F2 emphasizes recall, the optimization favored probability distributions that preserved higher recall. On the validation set, incorporating MFCC probabilities did not increase recall beyond that achieved by Wav2Vec2 alone. Consequently, the final system relied on Wav2Vec2 posterior probabilities. But to improve class balance, we applied decision-threshold tuning to the depression-class, leading to a final macro-F1 of 94.4% as shown in Table 3.

Model	Prec. (%)	Rec. (%)	F1 (%)
MFCC–CNN	85.09	78.75	77.75
Wav2Vec2	77.03	57.50	48.13
Validation-tuned Ensemble	94.44	94.38	94.37

Table 3: Performance comparison of Tamil depression detection models on the test set.

5.2 Task 2: Depression detection in Malayalam

Initial experiments showed strong performance from both models, with the MFCC–MLP achiev-

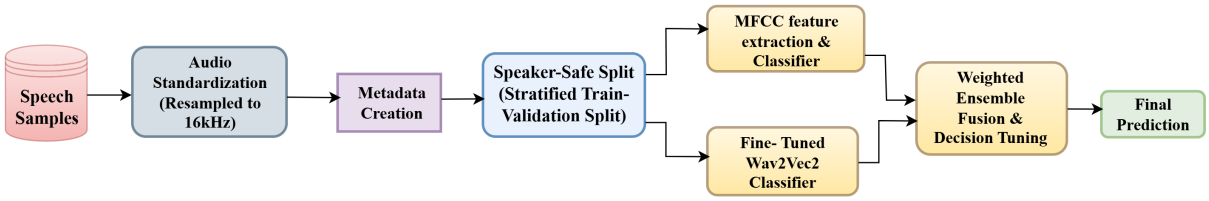
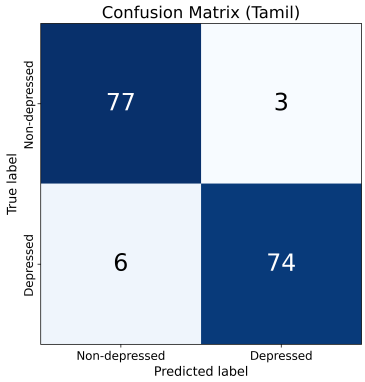


Figure 1: Overview of the proposed depression detection pipeline

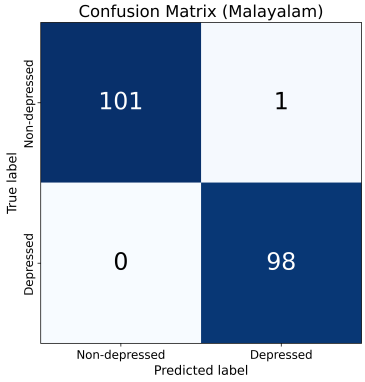
ing a macro-F1 of 99% and the Wav2Vec2 model obtaining 98.5% on the Malayalam test set (Table 4). Similar to the Tamil subtask, weight tuning for Malayalam also indicated that Wav2Vec2 representations generalized more reliably than MFCC-based predictions. Since fusion was implemented via a convex combination of posterior probabilities followed by argmax selection, introducing a non-zero MFCC weight modifies the resulting probability distribution and can shift class decisions. Validation results showed that increasing the MFCC contribution did not improve macro-F1 and slightly degraded performance. So, we retrained the Wav2Vec2 model on the combined training and validation data using light augmentation and class-weighted optimization. This slight refinement improved the macro-F1 score to 99.5% on the test set.

Model	Precision (%)	Recall (%)	Macro-F1 (%)
MFCC–MLP	99.04	98.98	99.00
Wav2Vec2	98.51	98.49	98.50
Refined Wav2Vec2	99.49	99.51	99.50

Table 4: Performance comparison of Malayalam depression detection models on the test set.



(a) Tamil



(b) Malayalam

Figure 2: Confusion matrices of the final models on the test sets.

5.3 Error Analysis

To further analyze model behavior, the confusion matrices of the final systems are shown in Fig. 2. The Malayalam model achieved almost perfect class separation with only one false positive. But the Tamil model showed slightly higher confusion between classes with 3 false positives and 6 false negatives.

Fusion weights indicated minimal complementarity between MFCC and Wav2Vec2 predictions in both languages. Malayalam benefited primarily from representation refinement, whereas Tamil required decision-threshold calibration to mitigate class bias and achieve balanced performance.

6 Conclusion

This work proposed a speech-based depression detection framework for Tamil and Malayalam by combining MFCC-based neural models and pre-trained Wav2Vec2 representations. MFCC features provided reliable spectral cues, while Wav2Vec2 captured higher-level speech patterns from raw audio. We also explored ensemble fusion to leverage complementary strengths but validation results indicated that Wav2Vec2 generalized better to unseen data, which led us to a refined training strategy. The final systems achieved strong performance on the test sets, reaching a macro-F1 of 99.5% for Malayalam and 94.4% for Tamil. Throughout the task, we tried to address major obstacles such as limited data, class imbalance, and speaker variability by implementing speaker-aware splitting, augmentation, and class-weighted optimization.

Limitations

Despite strong results, the system has several limitations. The datasets are relatively small and collected under controlled conditions with limited variability, which may restrict generalization to diverse speakers, environments, and demographic groups. Although speaker-aware splitting reduces leakage, some residual biases may still remain. A two-layer MLP is used as a lightweight baseline rather than a state-of-the-art model to evaluate MFCC effectiveness under low-resource settings. Similarly, most Wav2Vec2 encoder layers are frozen to ensure training stability, but this may limit task-specific adaptation. While recent foundation models such as HuBERT and Whisper offer stronger generalization through large-scale pretraining, their high computational and data requirements make them unsuitable for this setting. From a clinical perspective, the system is intended for preliminary screening or decision support rather than diagnosis, and therefore requires further validation on more diverse populations before real-world deployment.

Ethical Considerations

This work was conducted following the highest ethical research practices. The study aims to support research on depression detection while respecting privacy and cultural diversity. We acknowledge the sensitive nature of mental health analysis and have taken steps to reduce bias and ensure fair model behavior. The findings are intended to assist research and awareness efforts, not to replace professional clinical judgment.

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.

Sona Binu, Jismi Jose, Fathima Shimna K V, Alino Luke Hans, Reni K. Cherian, Starlet Ben Alex, Priyanka Srivastava, and Chiranjeevi Yarra. 2024. Language-agnostic analysis of speech depression detection. *Preprint*, arXiv:2409.14769.

Klara Daly and Oluwafemi Olukoya. 2025. Depression detection in read and spontaneous speech: A multimodal approach for lesser-resourced languages. *Biomedical Signal Processing and Control*, 108:107959.

S. Gupta, G. Agarwal, S. Agarwal, et al. 2024. Depression detection using cascaded attention based deep learning framework using speech data. *Multimedia Tools and Applications*, 83:66135–66173.

Spencer L James, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. 2018. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The lancet*, 392(10159):1789–1858.

N. Janardhan and N. Kumaresh. 2022. Improving depression prediction accuracy using fisher score-based feature selection and dynamic ensemble selection approach based on acoustic features of speech. *Traitement du Signal*, 39(1):87–107.

A. Kim, E. Jang, S. Lee, K. Choi, J. Park, and H. Shin. 2023. Automatic depression detection using smartphone-based text-dependent speech signals: Deep convolutional neural network approach. *Journal of Medical Internet Research*, 25:e34474.

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)*.

Jyothish G. 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.

Anjali Mathew, Raniya, Harsha Sanjan, Amjith S B, Reni K Cherian, Starlet Ben Alex, Priyanka Srivastava, and Chiranjeevi Yarra. 2024. Instant-emdb: A multi-model spontaneous english and malayalam speech corpora for depression detection. In *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6.

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.

Khosro Rezaee. 2026. Depression detection from speech data using deep learning–based optimized

temporal–frequency–channel attention with interpretable acoustic–prosodic mapping. *Journal of Affective Disorders*, 399:121077.

Felix Teufel, Aastha Aggarwal, Lydia Chwastiak, Vikram Patel, and Mohammed K. Ali. 2024. Depression diagnosis, treatment, and remission among adults in india. *JAMA Psychiatry*, 81(12):1265–1269.

Faming Yin, Jing Du, Xinzhou Xu, and Li Zhao. 2023. Depression detection in speech using transformer and parallel convolutional neural networks. *Electronics*, 12(2).

Xu Zhang, Xiangcheng Zhang, Weisi Chen, Chenlong Li, and Chengyuan Yu. 2024. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. *Scientific Reports*, 14:9543.

A Appendix

A.1 Bootstrap-Based Robustness Analysis

To assess the stability and statistical reliability of model performance, we applied bootstrap resampling on test set predictions. We estimated the distribution of macro-F1 scores over 1000 resamples and reported the 95% confidence interval (CI). This allowed evaluation beyond point estimates, capturing variability across different sampled subsets of the data.

A.1.1 Tamil Test Set

Model	F1 (%)	CI (L)	CI (U)
MFCC–CNN	78.00	72.00	83.00
Wav2Vec2	47.92	40.77	56.36
Validation-tuned Ensemble	**94.38**	**90.61**	**97.50**

Table 5: Bootstrap Macro-F1 with 95% CI (Tamil)

The Tamil results revealed clear differences in both performance and robustness. MFCC–CNN achieved moderate performance with relatively stable confidence bounds, whereas Wav2Vec2 exhibited low macro-F1 and high variability, indicating weak generalization. In contrast, the proposed ensemble substantially improved both accuracy and stability, achieving the highest score with a comparatively narrow CI.

A.1.2 Malayalam Test Set

In contrast to Tamil, all models achieved consistently high performance with narrow confidence intervals, suggesting stable behavior across resamples. The limited spread indicated low dataset variability, which may inflate performance estimates. Consequently, the ensemble provided only marginal improvement, while Wav2Vec2 showed slightly higher variability, reflecting its sensitivity to data distribution.

Model	F1 (%)	CI (L)	CI (U)
MFCC–MLP	98.51	97.48	99.00
Wav2Vec2	97.04	95.07	98.69
Refined Wav2Vec2	**99.01**	**95.90**	**99.80**

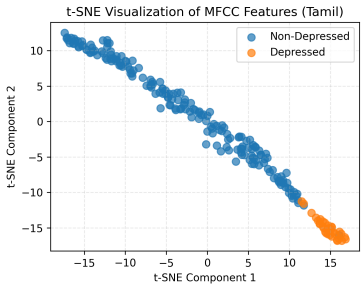
Table 6: Bootstrap Macro-F1 with 95% CI (Malayalam)

A.2 Representation-Level Analysis via t-SNE

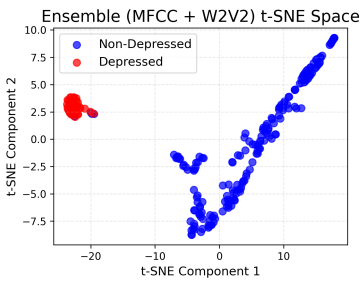
In this section, we conducted a representation-level analysis using t-SNE visualization. We extracted embeddings from intermediate layers prior to classification and applied t-SNE for dimensionality reduction to project them into a two-dimensional space. This allowed us to examine how effectively each model separates the Depressed and Non-Depressed classes in the learned feature space. Since the final systems for both languages were primarily driven by Wav2Vec2 representations, we showed only two t-SNE visualizations: one for the MFCC-based model and one for the final Wav2Vec2-driven system.

A.2.1 Tamil Dataset

We visualized the embeddings using t-SNE, as shown in Fig. 3



(a) MFCC embeddings



(b) Embeddings of Validation-tuned Ensemble

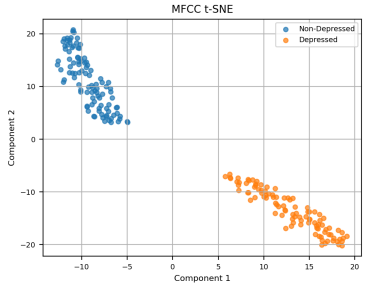
Figure 3: t-SNE visualization for the Tamil dataset.

While MFCC captured basic spectral features, it

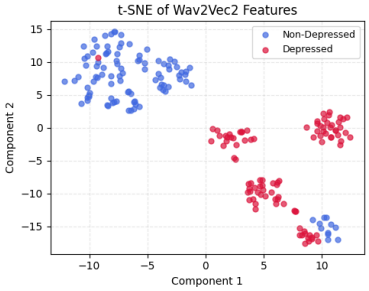
struggled with clear class separation, as evidenced by the linear distribution and overlap. In contrast, the Ensemble plot showed a much clearer separation between the classes. Wav2Vec2, through its self-supervised learning, captured high-level semantic and phonetic features from the Tamil dataset, allowing the model to isolate the Depressed samples into a dense, distinct cluster.

A.2.2 Malayalam Dataset

t-SNE visualizations of the Malayalam embeddings are shown in Fig. 4.



(a) MFCC Embeddings



(b) Refined Wav2Vec2 Embeddings

Figure 4: t-SNE visualization for Malayalam.

The t-SNE plots for MFCC and Wav2Vec2 confirmed high linear separability. While MFCC formed two isolated clusters, Wav2Vec2 displayed granular sub-clusters because it is a self-supervised deep neural network capturing complex, hierarchical acoustic patterns. The minor overlaps in Wav2Vec2 occurred due to inter-class similarity, where the model detected nuanced speech cues (like low energy) shared by both classes. Despite these overlaps, the transformer-based architecture provided a richer representation, ensuring reliable class discrimination.