SERENE@DravidianLangTech 2026: Multimodal Approaches for Depression Detection in Dravidian Speech: Acoustic, Spectrogram, and Transformer-Based Models

TT Pranesh, KK Thamizhmathi, S Vigneshwaran, B Bharathi
Department of Computer Science and Engineering
Sri Sivasubramania Nadar College of Engineering
pranesh2370060@ssn.edu.in
thamizhmathi2370055@ssn.edu.in
vigneshwaran2370061@ssn.edu.in
bharathib@ssn.edu.in

Abstract

This paper presents our submission to the Depression Detection in Dravidian Languages shared task at DravidianLangTech 2026. We investigate three complementary approaches for speech-based depression detection in Tamil and Malayalam: (i) acoustic feature engineering using MFCC and prosodic features with a Support Vector Machine (SVM) classifier, (ii) a convolutional neural network (CNN) trained on Mel-spectrogram representations, and (iii) a transformer-based model using Whisper-generated transcripts fine-tuned with XLM-RoBERTa. Experimental results show that acoustic feature-based SVM and spectrogram-based CNN models achieve the strongest performance on both Tamil and Malayalam datasets, while the transformer-based approach also produces competitive results. We further discuss limitations and future research directions.

1 Introduction

Depression is a prevalent mental health condition characterized by persistent sadness, reduced interest, cognitive impairment, and emotional instability. Recent research in computational linguistics and speech processing has explored automatic detection of depression using textual and acoustic cues. While significant progress has been made in high-resource languages such as English, low-resource languages such as Tamil and Malayalam remain underexplored.

The Depression Detection in Dravidian Languages shared task (G et al., 2026) aims to address this gap by providing speech data in Tamil and Malayalam for automatic depression classification. Speech signals contain rich acoustic and linguistic cues including prosody, lexical choice, hesitation patterns, and emotional tone.

In this work, we explore three modeling paradigms of increasing representational complexity:

- Acoustic feature-based classification using MFCC and prosodic features with SVM.
- Spectrogram image-based classification using a Convolutional Neural Network (CNN).
- Transformer-based textual classification using Whisper-generated transcripts fine-tuned with XLM-RoBERTa.

Our goal is to systematically compare traditional machine learning, deep convolutional models, and transformer-based approaches within the shared-task setting.

The remainder of this paper is organized as follows. Section 2 reviews existing literature on speech-based depression detection and related approaches. Section 3 describes the dataset characteristics and distribution across Tamil and Malayalam. Section 4 presents the proposed approaches, including acoustic feature-based classification, spectrogram-based convolutional neural networks, and transformer-based text modeling. Section 5 reports the experimental results and comparative performance analysis of the proposed models. Section 6 discusses the limitations of the current study. Finally, Section 7 concludes the paper and outlines potential directions for future research.

2 Related Work

Automatic depression detection has attracted growing attention in speech processing and natural language processing communities. Early research primarily relied on handcrafted acoustic features such as MFCCs, pitch, and energy statistics to model depression-related vocal characteristics. With the

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 354–358
July 3, 2026 ©2026 Association for Computational Linguistics

advancement of deep learning, convolutional neural networks (CNNs) operating on spectrogram representations became increasingly popular. Ensemble CNN frameworks for spectrogram-based depression detection have demonstrated improved robustness over classical machine learning approaches (Vázquez-Romero and Gallardo-Antolín, 2020). Deep CNN models applied to smartphone-recorded speech signals have further shown that learned spectral representations effectively capture depression-related acoustic patterns (Kim et al., 2023). Hybrid approaches combining MFCC-based handcrafted features with CNN-generated spectrogram representations highlight the complementary strengths of statistical and learned features (Das and Naskar, 2024). Recent work in Dravidian languages has also explored speech-based depression classification using deep learning techniques (Kritika et al., 2025). Previous shared task initiatives in Dravidian languages have also promoted research on multimodal and speech-based analysis of low-resource languages (Premjith et al., 2024).

More recently, transformer-based models have significantly advanced depression detection research. Hybrid CNN and transformer-based pretrained language models have reported notable improvements compared to traditional methods (Ramalakshmi et al., 2024). The emergence of large-scale foundation models has further strengthened text-based analysis. A multilingual automatic speech recognition system trained with large-scale weak supervision enables reliable transcription of low-resource languages (Radford et al., 2023). Speech and text transformer architectures have demonstrated strong cross-lingual generalization capabilities for depression detection across tasks and languages (Gómez-Zaragozá et al., 2025).

Beyond unimodal approaches, multimodal fusion techniques have also been investigated to integrate acoustic, textual, and behavioral signals. Multimodal data fusion has been shown to enhance robustness and improve generalization in depression detection systems (Nykoniuk et al., 2025). Multimodal modeling in Dravidian languages has also been explored for related tasks (Anilkumar et al., 2026). While prior work has largely focused on high-resource languages and multimodal datasets, systematic evaluation of acoustic, spectrogram-based, and transformer-driven methods in low-resource Dravidian languages remains limited. Our work addresses this gap by comparatively analyzing these paradigms for Tamil and Malayalam speech.

Table 1: Dataset statistics for Malayalam and Tamil.

Statistic	Malayalam	Tamil
Total Samples	1,888	1,534
Depressed Samples	888	534
Non-Depressed Samples	1,000	1,000
Unique Speakers	8	9
Depressed Speakers	3	4
Non-Depressed Speakers	5	5
Training Samples	1,688	1,374
Test Samples	200	160
Train (%)	89.41	89.57
Test (%)	10.59	10.43

3 Dataset Description

The dataset consists of speech recordings in two Dravidian languages, Tamil and Malayalam, where each utterance is labeled as *Depressed* or *Non-depressed*, forming a binary classification task. Speech signals contain indicators of depression through acoustic and linguistic cues such as prosody, speech rate, and lexical patterns. The corpus includes 1,888 Malayalam and 1,534 Tamil samples collected from 8 Malayalam speakers (3 depressed, 5 non-depressed) and 9 Tamil speakers (4 depressed, 5 non-depressed). The data is split into training and test sets with approximately 90% used for training and 10% for testing, as summarized in Table 1. Since multiple utterances are recorded from the same speakers, intra-speaker similarities may influence model performance.

4 Methodology

4.1 Acoustic Feature-Based Classification

In the first approach, handcrafted acoustic features are extracted using the Librosa library. A 33-dimensional feature vector is constructed from MFCC coefficients, pitch statistics, energy statistics, and silence ratio, capturing spectral, prosodic, and temporal characteristics of speech associated with emotional and psychological states. The extracted features include 13 MFCC coefficients summarized using mean and standard deviation, along with pitch and energy statistics and silence ratio.

The feature vectors are used to train a Support Vector Machine (SVM) classifier with a radial basis function (RBF) kernel. The model is trained with $C = 10$, $\gamma =$ scale, and balanced class weights to address class imbalance. The overall acoustic feature-based pipeline is illustrated in Figure 1.

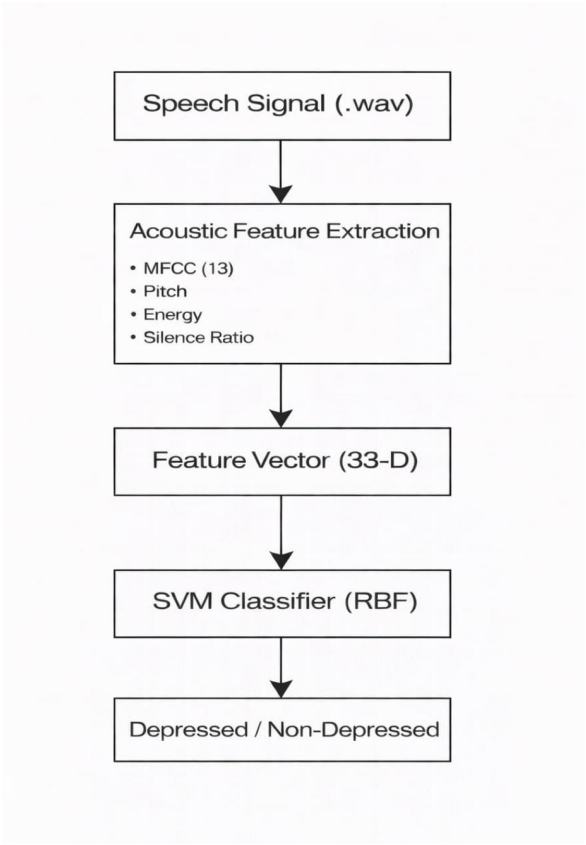

Figure 1: Acoustic feature-based depression detection pipeline

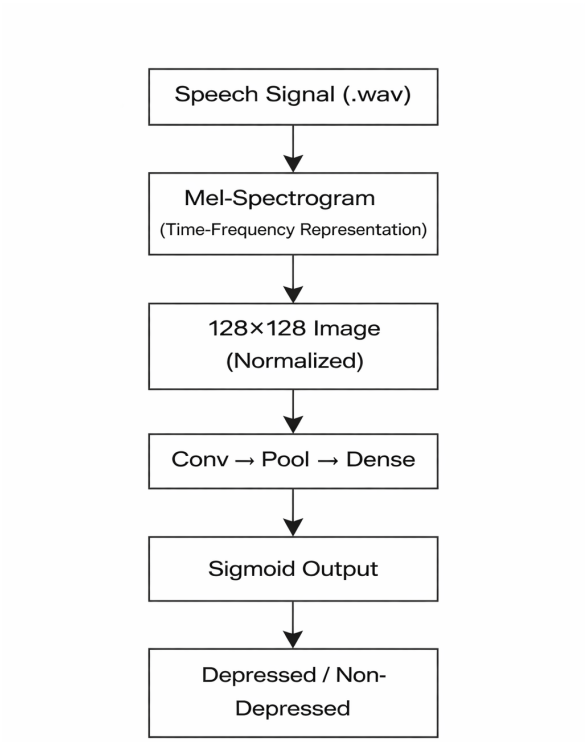

Figure 2: Spectrogram-based CNN pipeline.

4.2 Spectrogram-Based CNN Model

In the second approach, each speech signal is converted into a Mel-spectrogram representation using the Librosa library. Audio recordings are truncated to 5 seconds and transformed into Mel-frequency spectrograms in decibel (dB) scale to capture temporal and spectral speech characteristics.

The generated spectrograms are resized to 128×128 pixels and normalized before being used as input to a Convolutional Neural Network (CNN). The CNN architecture consists of two convolutional layers with 32 and 64 filters followed by max-pooling, a fully connected layer with dropout regularization, and a sigmoid output layer for binary classification.

The model is trained using binary cross-entropy loss and the Adam optimizer for 20 epochs. The overall spectrogram-based classification pipeline is illustrated in Figure 2.

4.3 Transformer-Based Text Classification

In the third approach, we incorporate linguistic information by converting speech signals into textual transcripts using the Whisper medium automatic speech recognition (ASR) model. Whisper is a multilingual transformer-based ASR system capable of transcribing low-resource languages such as Tamil.

The generated transcripts are cleaned by removing unnecessary punctuation and normalizing whitespace. The processed text is tokenized using the XLM-RoBERTa tokenizer and converted into subword embeddings.

We fine-tune XLM-RoBERTa-base, a multilingual transformer pretrained on 100 languages, for binary depression classification. All transformer encoder layers are updated during training, and the contextual representation of the classification token is passed through a linear classification head for prediction.

The model is trained using cross-entropy loss with a learning rate of 2×10^{-5}, batch size of 8, and 6 training epochs. The overall transformer-based pipeline is illustrated in Figure 3.

5 Results

We evaluate the proposed approaches using Accuracy(Acc), Macro-Precision(M-P), Macro-Recall(M-R), and Macro-F1(M-F1) score. Tables 2

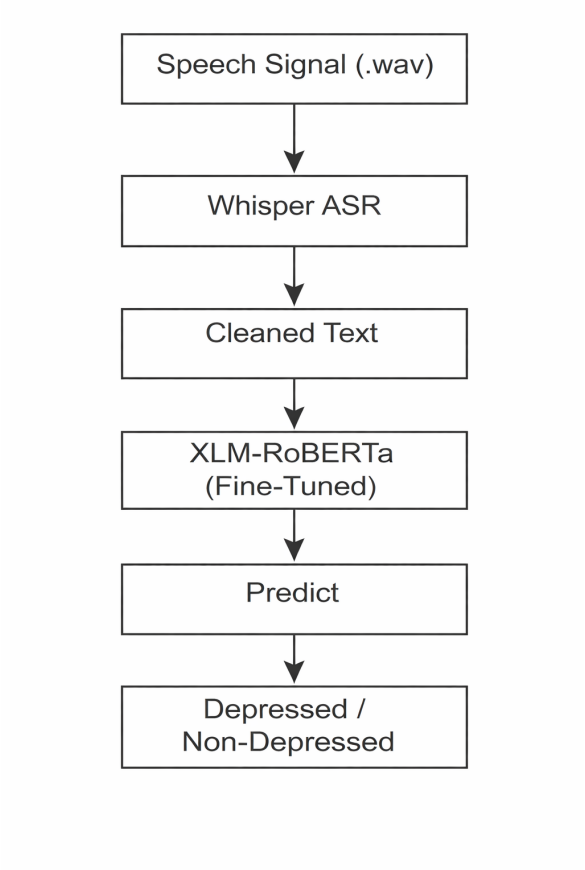

Figure 3: Transformer-based depression detection pipeline.

Table 2: Performance comparison on the Tamil dataset.

Model	Acc	M-P	M-R	M-F1
Acoustic + SVM	1.00	1.00	1.00	1.00
Spectrogram + CNN	1.00	1.00	1.00	1.00
Whisper + XLM-R	0.96	0.96	0.96	0.96

and 3 summarize the performance of the models on Tamil and Malayalam datasets. For Tamil, the acoustic feature-based SVM and spectrogram-based CNN models achieve near-perfect classification performance, while the transformer-based Whisper + XLM-RoBERTa approach obtains slightly lower scores. Similarly, on the Malayalam dataset, both acoustic and spectrogram-based models achieve near-perfect performance across all evaluation metrics. Our acoustic feature-based SVM system (Run 1) achieved **Rank 1** in both Tamil and Malayalam tracks of the shared task leaderboard. The implementation and experimental codes are publicly available on GitHub[1].

[1] https://github.com/Pranesh4950/Depression-Detection-Task

Table 3: Performance comparison on the Malayalam dataset.

Model	Acc	M-P	M-R	M-F1
Acoustic + SVM	1.00	1.00	1.00	1.00
Spectrogram + CNN	1.00	1.00	1.00	1.00

6 Limitations

The dataset size is relatively small and may limit generalization. Since multiple utterances are recorded from the same speakers, speaker-dependent patterns may influence model performance and lead to overestimated results. In addition, depressed and non-depressed recordings differ in recording conditions and sampling rates, which may introduce unintended acoustic cues that models can exploit. Whisper transcription errors may also affect transformer-based textual modeling. Future work will focus on speaker-independent evaluation and more robust validation settings.

7 Conclusion

We presented three approaches for depression detection in Tamil and Malayalam speech. Acoustic feature-based SVM and spectrogram-based CNN models achieved the strongest performance, while the transformer-based textual approach also produced competitive results. These findings highlight the effectiveness of acoustic and spectrogram representations for depression detection in speech. Due to the limited number of speakers and repeated utterances per speaker, speaker-independent evaluation remains an important direction for future work

8 Ethical Considerations

Depression detection systems should be used only as assistive tools and not as clinical diagnostic systems. Sensitive speech data must be handled responsibly with proper privacy protection and informed consent.

Acknowledgment of Generative AI Usage

Generative AI tools were used for language refinement and formatting assistance during manuscript preparation. All experiments and evaluations were conducted by the authors.

References

A. Anilkumar, Jyothish Lal, B. Premjith, and Bharathi Raja 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.

Arnab Kumar Das and Ruchira Naskar. 2024. A deep learning model for depression detection based on MFCC and CNN-generated spectrogram features. *Biomedical Signal Processing and Control*, 90:105898.

Jyothish Lal G, Premjith B, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Durairaj Thenmozhi, 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.

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 *Proceedings of Interspeech 2025*, pages 5253–5257.

Ah Young Kim, Eun Hye Jang, Seung-Hwan Lee, Kwang-Yeon Choi, Jeon Gue Park, and Hyun-Chool Shin. 2023. Automatic depression detection using smartphone-based text-dependent speech signals: A deep convolutional neural network approach. *Journal of Medical Internet Research*, 25:e34474.

A. Kritika, S. Meenakshy, Arya Palackal Shijish, Riya Rajeev, and 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)*.

Mariia Nykoniuk, Oleh Basystiuk, Nataliya Shakhovska, and Nataliia Melnykova. 2025. Multimodal data fusion for depression detection approach. *Computation*, 13(1):9.

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 International Conference on Machine Learning (ICML)*, pages 28492–28518. PMLR.

N. R. Ramalakshmi, Meghna Ganesh Kumar, and S. Raghavi. 2024. Automated depression detection using CNN and transformer-based pre-trained language models. In *Proceedings of the International Conference on Computer, Communication, and Signal Processing*, pages 42–53. Springer.

Adrián Vázquez-Romero and Ascensión Gallardo-Antolín. 2020. Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*, 22(6):688.