Advances in Artificial Intelligence and Machine Learning; Research 6 (5) 6367-6378 Received 15-06-2026; Accepted 03-10-2026; Published online 10-10-2026

An Intelligent Speech Analytics Framework for Depression Detection: Comparative Assessment of Machine Learning Architectures

Deepak Kumar javadevdeepak@gmail.com
School Of Engineering amd Technology
K R Mangalam University- Gurgaon, India.

Monika Khatkar monika.khatkar@krmangalam.edu.in
School Of Engineering amd Technology
K R Mangalam University- Gurgaon, India.

Corresponding Author: Deepak Kumar

Copyright © 2026 Deepak Kumar and Monika Khatkar. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Depression affects emotional, cognitive, and behavioral functioning and it is a pretty ubiquitous mental health (MH) disorder. It is often described as subjective time consuming, and kind of expert-driven, so clinical interviews plus self reported questionnaires are still the traditional methods, for diagnosing. Automated systems can detect depressed symptoms utilising voice biomarkers with the aid of AI and speech signal processing (SSP). Emotional and psychological states can greatly affect the variation in pitch, energy, pace, and spectrum characteristics of speech, rendering it a potential tool to assess depression. This research is designed to develop an independent depression detection system (DDS) using speech and voice (S&V) input and evaluate ML and DL algorithms on depression and non-DD. The audio files should be used to get the data, labels should be extracted from the audio files, features should be taken from preprocessed data, the model should be trained with the extracted features, and the performance of the Model should be evaluated. Librosa extracts Mel Frequency Cepstral Coefficients (MFCCs), Chroma features, and Mel Spectrogram (MS) features from speech and is able to gather them in a hybrid feature vector. We split the data straight into training (80%) and test (20%) sets in a normal random fashion. The accuracy, precision, recall for the MLP Model were observed as 98% , 99% and 99% respectively, while the ROC-AUC score came out to 0.96 in the trials. Overall the results propose that S&V signals can be leveraged for depression identification and to strengthen automated MH assessments. Also, screening for depression plus clinical decision support with speech AI appears to be efficient, non-invasive alongside scalable.

Keywords: Depression detection, Speech analysis, Voice signals, Machine learning, Deep learning, MFCC, CNN, LSTM, Random forest, Mental health assessment, Artificial intelligence.

Citation: Deepak Kumar and Monika Khatkar. An Intelligent Speech Analytics Framework for Depression Detection: Comparative Assessment of Machine Learning Architectures. **Advances in Artificial Intelligence and Machine Learning**. 2026;6(5):351. https://dx.doi.org/10.54364/AAIML.2026.65351

1. INTRODUCTION

Depression is one of the most ubiquitous psychiatric disorders globally and remains a significant public health concern. It adversely affects mood, cognitive processes, interpersonal relationships, and overall well-being. Evidence indicates that depression can occur across all age groups and is associated with an increased risk of anxiety disorders, substance abuse, self-injury, and suicide. When diagnosis is done early, and intervention comes sooner, people often see stronger symptom reduction, plus better treatment outcomes. The usual way of diagnosis still leans on the MH professional's somewhat subjective assessment, usually through clinical interviewing, self reported questionnaires, and behavioural observation; this whole process is time consuming, and it also needs health workers or specialists who have solid knowledge about MH [1].

With the latest advancements in AI, ML, and DL, there are new opportunities for objective and automated systems for MH assessment. Recently, the development of AI, ML, and DL has created new opportunities for the creation of automatic and objective MH assessment systems. AI models of psychodiagnostic algorithms can be used to analyze behavioral, emotional and physiological markers of MH issues, thereby aiding in early intervention. Intelligent technology can be utilized to recognize patients who are at danger and supply early support to healthcare staff members [1]. Advances in AI have increased the role of technology in mental health monitoring, with AI-driven systems demonstrating the ability to anticipate psychological and behavioural responses to stressful conditions

S&V signals can be biomarkers for depression. Pitch, speaking rate, articulation, amount of energy, length of pauses and expression are all affected by depression. These changes can be picked up by acoustic signal processing and ML techniques. Research has recently indicated that speech might be a viable non-invasive option for continuous and remote MH monitoring over clinical assessments [2, 3]. Improvements in processing capabilities and the accumulation of comprehensive speech datasets have created new opportunities for MH research. Consequently, studies have begun leveraging voice characteristics alongside textual and emotional information to strengthen the detection of psychological conditions. AI multimodal (MM) methods can incorporate additional data streams from different sources, further increasing the predictiveness and MH evaluations [4]. In recent years, Depression **Detection studies** have been studied and DL methods such as CNN and LSTM networks have been successfully applied to detect depression using speech and text data to discover depression patterns [5]. Building upon recent advancements, the present study proposes an automated Depression Detection approach using S&V signal analysis. Various ML and DL models are used to test the categorization method to find out which method is the best for categorizing the depression. Proposed solution is grounded in artificial intelligence and voice processing for early screening and MH evaluation, which is reliable, non-invasive and scalable.

2. LITERATURE REVIEW

The following table, TABLE 1 delivers a detailed illustration of the DD Utilising S&V and MM AI Technologies.

Table 1: Related works

Authors and Year	Methodology	Study Findings in Relation to the Research Objectives
Xu et al. [6]	Conducted a detailed investigation of MM data integration alongside AI technologies for medical diagnosis with multiple datasets including text, speech, image, physiological and sensor data.	Showed that integrating speech with behavioural and emotional data improves the accuracy and reliability of MH assessment compared with relying on speech alone.
Caulley et al. [7]	Created an AI-grounded voice recognition and emotion analysis system to assess child mental severity. The study employed speech recordings alongside universal emotion recognition algorithms to analyse emotional patterns and support MH assessment.	Showed that speech characteristics together with emotional vocal cues can serve as objective biomarkers for identifying psychiatric conditions, enabling earlier detection of depression and supporting MH assessment.
Tank et al. [8]	Used LLMs to detect depression in text, audio, and visual modalities. MM fusion and transformer-grounded architectures were examined.	Reported that MM Models achieved better performance than single-modality approaches, with the integration of language understanding and speech features significantly enhancing depression detection accuracy.
Zhang et al. [9]	Proposed a MM AI Model to diagnose depression severity using audio and video data. Features were extracted and classified using DL.	Used voice and facial behavioural signals to accurately assess depression severity, demonstrating the efficacy of MM feature fusion for depression identification.
Liu et al. [10]	Reviewing AI-powered MM depression and anxiety detection methods using voice, facial expressions, physiological signs, and ML.	Recommended integrating speech analytics with emotional and behavioural indicators since MM AI systems consistently outperform unimodal systems.
Sharma [11]	Researched AI-driven MH intervention emotion recognition systems utilising speech analysis, facial expressions, alongside affective computing.	Emotion identification from speech can help detect depressed symptoms and guide early intervention and MH monitoring.
Bhayo [12]	Created a real-time MM XAI framework for depression and anxiety identification using voice, facial, and behavioural data. The framework stressed understandability and transparency.	Provides evidence for interpretable DD Models by showing that explainable MM systems raise clinician trust and decision-making while preserving excellent prediction accuracy.
Mahur [13]	Methodologically reviewed MM conversational AI systems for MH applications that integrate text, speech, emotion, and behavioural data.	Real-time MM analysis is crucial for speech and emotion-aware conversational systems to monitor depression and provide personalised MH care.

Research Gap: The studies that were examined collectively suggest that S&V signals are useful biomarkers for the identification of depression, particularly when paired with information regarding

emotions, textual content, visual content, and behavioural patterns. These findings lend support to the creation of a unique way for detecting depression. This method would make use of S&V features in addition to MM emotional and behavioral indicators. Multimodal emotional and behavioural indicators are an essential avenue for future research . This research concentrates on acoustic speech information for depression identification. In the suggested framework, WAV speech recordings are used and MFCC, Chroma and Mel-spectrogram characteristics are extracted to represent relevant audio patterns. The present experimental implementation does not include video, facial-expression and textual modalities. The goal of this method would be to reach a higher scale of accuracy, interpretability, and early detection capabilities than the approaches that are now in use.

3. METHODOLOGY

This work looks to develop an automated DDS grounded in S&V cues for detecting depression in humans. They use SSP, feature extraction, ML and DL techniques to detect the depression-related auditory patterns. The speech dataset utilised in the study is a set of WAV-format audio recordings retrieved from the dataset archive used for the tests. This research examined the audio recordings for depression detection and parsed the related file names to acquire the depression and gender labels. For the purpose of model creation and evaluation, the data set was split into 80% training and 20% testing subsets. All audio samples were processed at a sampling rate of 22,050 Hz. Acoustic features including Mel-Frequency Cepstral Coefficients (MFCCs), Chroma features and Mel Spectrogram (MS) features were extracted with the Librosa library. The class for depression was named as 1 while the class for non-depression was named as 0. These dataset and preparation information provide a more detailed description of the data utilised in the experimental study and increase the reproducibility of the proposed methodology.

The current approach relies on acoustic speech information only. Librosa was used to extract MFCC, Chroma and Mel-spectrogram features from WAV files. The experimental framework did not include any video, facial-expression, or textual components.

3.1 Dataset Details

The dataset has 2,880 audio clips from 24 speakers, and provides a fairly structured experimental dataset for evaluation of speech based emotion classification. The details of data set shown in TABLE 2. The number of recordings is considerable compared to the number of speakers as each speaker contributes numerous recordings at different emotion situations. This is significant since the effective independence of observations relies on the quantity of audio files, but also on the number of independent speakers. Therefore, the generalisation ability of the models should be understood in the light of the fact that 24 speakers were used.

3.1.1 Record and duration

The overall duration of the audio material is 10,657.91 seconds or approximately 177.63 minutes or 2.96 hours of talking. The average recording duration was 3.70 seconds, with recordings ranging

Table 2: Dataset Details

Parameter	Result	Interpretation
Dataset name	RAVDESS*	Dataset identified by the current pipeline
Total audio files	2,880	Substantial number of speech recordings
Number of speakers	24	24 distinct speakers represented
Total duration	10,657.91 sec	Approximately 177.63 minutes / 2.96 hours
Mean duration	3.70 sec	Short speech recordings on average
Minimum duration	2.94 sec	Shortest recording
Maximum duration	5.27 sec	Longest recording
Speaker IDs	1–24	All 24 speakers represented
Target definition	Emotion tag $04 \rightarrow 1$	Binary target based on filename emotion code
Other emotion tags	$\rightarrow 0$	All remaining emotion categories combined
Training speakers	19 speakers	Used for model development
Testing speakers	5 speakers	Completely unseen speakers
Speaker overlap	None	Confirms speaker-independent testing

from 2.94 to 5.27 seconds. This implies that the dataset predominantly contains brief controlled voice recordings and not extensive genuine chats. The recordings have roughly the same length in time, which helps with computational consistency because of the very small range of durations.

3.1.2 Speaker distribution and reproducibility:

The dataset has speaker IDs from 1 to 24. Explicit reporting of these speaker identities in the reproducibility report makes the experimental design more visible. Otherwise, results at the recording level can be unnaturally strong when recordings of the same person are present in both training and testing data. Explicitly recording the speaker structure is thus a big improvement over a naive random file-based split.

3.2 Emotion-Tag Distribution

The emotion distribution based on filename has 192 recordings with tag 1 and 384 recordings for each of the tags 2 to 8. Therefore, the target category associated with tag 4 has 384 recordings, whereas the other categories together have 2,496 recordings. This shows that the binary classification problem is very lopsided with around 13.33% positives and 86.67% negatives. So, precision alone is not sufficient to determine the quality of the model.

3.2.1 Speaker-independent training and testing

The training set consists of 19 speakers while the testing set consists of 5 speakers: speakers 10, 16, 17, 19 and 20. Most crucially, the reported speaker overlap is an empty set, which implies that there is no testing speaker in the training data. This is a huge methodological strength, since the model

is tested on speakers that were not encountered during training. Therefore, the test findings give a more realistic approximation of the speaker-independent generalisation than a traditional random recording-level split.

Three DL architectures (CNN, LSTM, CNN-BiLSTM) are implemented and compared with three classical ML architectures (KNN, SVM, and RF, and MLP). The Accuracy, Precision, Recall (APR) and ROC-AUC were analyzed to select the most effective Model for detecting depression, using speech. Speech biomarkers can be utilised to develop a scalable, efficient and non-invasive depression screening system. The methodology proposed is shown inside FIGURE 1.

In this work, the target labels were extracted from the emotion tags embedded in the audio filenames as demonstrated in FIGURE 1. These labels are a proxy for affect in the categorisation task, not a clinical diagnosis of depression. So the results given should be seen as a classification of emotional/affective states on the basis of filenames and not as a detection or diagnosis of clinical depression. Further clinical validation with validated diagnostic assessments would be needed to inform any clinical interpretation.

3.3 Model Architectures and Hyperparameter Settings

The machine learning and deep learning models were developed with fixed architectural and training settings to ensure reproducibility of the results. For the conventional machine-learning models, KNN was set up with five neighbours, SVM was implemented with probability estimation enabled and Random Forest was set up with 200 trees and balanced class weighting. The MLP had 2 hidden layers with 256 and 128 neurones, ReLu activation and Adam optimisation with a learning rate of 0.001. We utilised a batch size of 32, early halting and a random seed of 42 for training.

For the deep-learning models, the CNN-BiLSTM architecture was adopted, consisting of convolutional layers with 128 and 64 filters, a kernel size of 5, followed by pooling and a bidirectional LSTM layer with 64 units. Finally, for classification, a dense layer of 64 units with a dropout rate of 0.30 was applied. The network was trained using Adam optimiser, learning rate 0.001, batch size 32 and max 30 epochs with early stopping. The matching CNN and LSTM models were trained with the same standardised training process, and the respective network topologies were developed based on the convolutional and recurrent structures.

To enable a fair comparison between the explored methodologies, certain hyperparameter settings were fixed for evaluating the model which present in TABLE 3. To promote openness and facilitate reproduction of experimental results, the complete model configurations are supplied.

4. RESULTS AND DISCUSSIONS

The four traditional ML algorithms KNN, SVM, RF, and MLP were utilised to test a DDS. Measurement of ROC-AUC were performed. The criteria were selected to ensure that a comprehensive evaluation of categorization performance is achieved and are important for health applications where accurate DD is essential. The techniques acquired MFCCs, Chroma and MS speech characteristics. To provide fair evaluation and generalization, the dataset was then split right into 80% being train

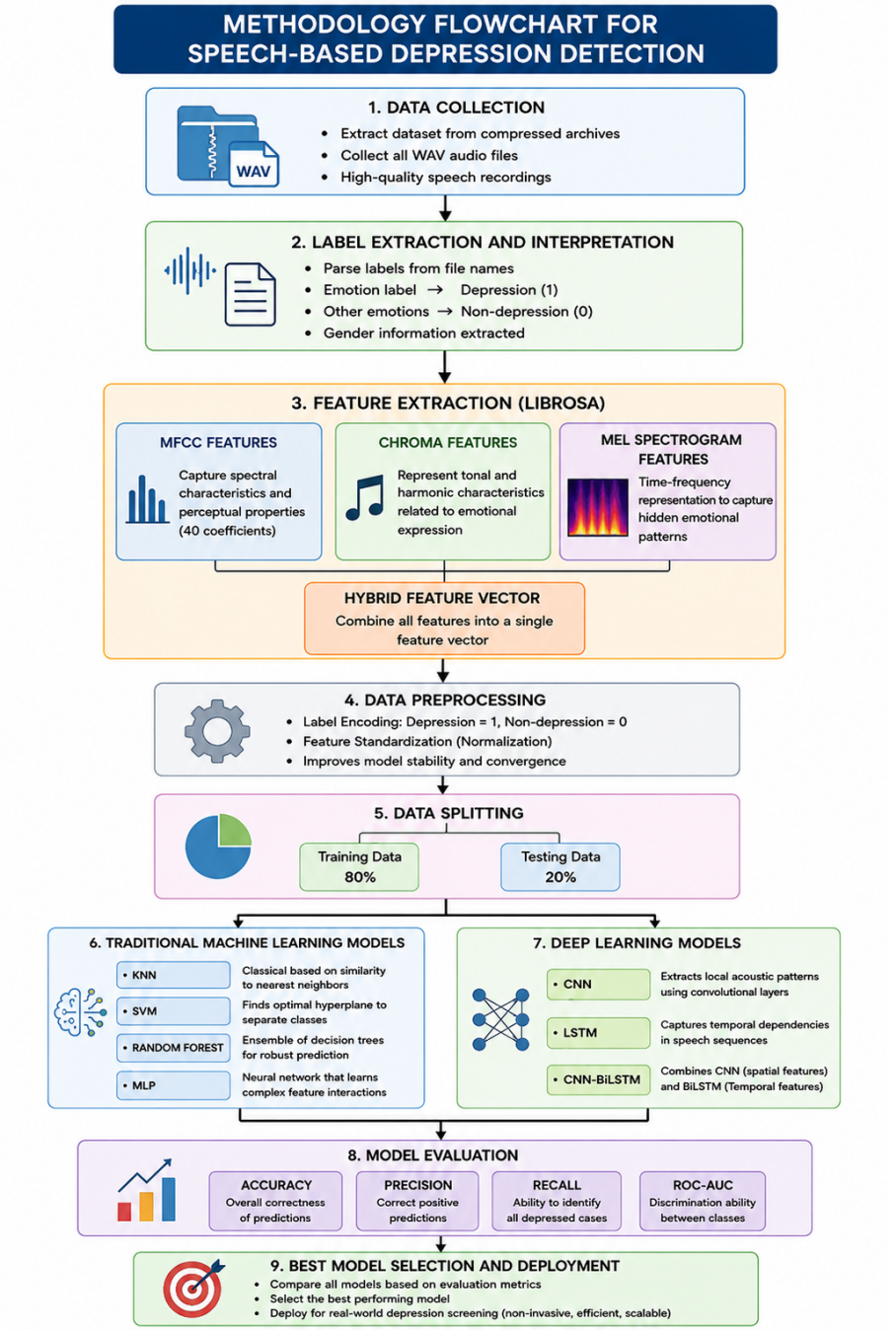

Figure 1: Proposed Methodology

set and 20% being test set. The dataset was then split right into 80% being train set and 20% being test set for fair evaluation and generalization.

Table 3: Hyper parameter details

Model	Main Architecture / Hyperparameters
KNN	k = 5
SVM	Probability estimation enabled
Random Forest	200 trees; balanced class weighting
MLP	Hidden layers: 256, 128; ReLU; Adam; learning rate = 0.001; batch size = 32; early stopping
CNN-BiLSTM	Conv filters: 128, 64; kernel size = 5; BiLSTM = 64 units; Dense = 64; dropout = 0.30; Adam; learning rate = 0.001; batch size = 32; maximum epochs = 30; early stopping

Table 4: Performance of Traditional ML Models

Model	Accuracy	Precision	Recall	ROC-AUC
KNN	95%	96%	98%	83%
SVM	89%	89%	100%	51%
Random Forest	98%	98%	100%	93%
MLP	98%	99%	99%	96%

As seen in TABLE 4, all ML Models were able to classify S&V signals, and the results show that there are indeed depression-related biomarkers. The ROC-AUC of the KNN classifier being 95%, 96%, 98% and 0.83, respectively. The results demonstrate that the clusters were sorted according to feature space proximity, for same acoustic voice samples. The strong recall value indicates that the auditory commonalities are relatively constant for depressed speech patterns, given the classification performed by KNN. The lower ROC-AUC score indicates that KNN algorithm is not suitable for complex and high dimensional speech recording feature representation.

This corroborates the theoretical disadvantage of the methodology outlined: the performance of KNN may suffer when large and complex datasets are used. SVM classifier reported as being 89%, 89%, and 100%, respectively and 0.51 being ROC-AUC. The Model was able to perfectly recall all depressed speech samples in the testing set. It is important in health care because if depression is not diagnosed, it can have profound consequences. However, the ROC-AUC value indicates that the Model has a poor discriminative performance between the depressed and the non-depressed groups. SVM is theoretically applicable for high dimensional data, and found an optimum separating hyperplane but turned out to be too optimistic in this case. This resulted in that the Model was able to detect all depressed individuals but had a higher rate of false positives, which reduced the discriminative power of the Model.

The traditional ML techniques such as Random Forest gave good results in classification. ROC-AUC was determined as 0.93. The data shows RF's ensemble learning technique spotted acoustic depression (AD) markers' nonlinear correlations. The decision-tree ensemble could learn complex relationships between the features extracted, since depression impacts all four aspects of speech:

pitch, speech energy, articulation patterns and temporal behaviour. The Model is able to accurately detect depressed individuals and has a high recall and precision, meaning that there are few potentially false alarms that would result in a misdiagnosis of depression. The depression-detection resilience of the Model is supported by the ROC-AUC result where there is good class separability. The ROC-AUC being determined as 0.96, were the highest among classic ML methods, for the MLP.

The neural network's amazing success can be attributed to its ability to acquire abstract representations and complex interactions of speech features. The hidden layers of the MLP architecture automatically learned high-level feature vectors that enhanced class separation in the hybrid feature vector. The near-perfect precision and recall suggest a good balance, meaning that there are minimal false positives alongside negatives. The Model has a ROC-AUC score of 0.96 that indicates its discriminative power between depressed and non-depressed speech. This study demonstrates neural network ML in exploiting subtle differences in speech inputs, both in terms of emotion and acoustics.

The study adopts a three-dimensional DL Model: CNN, LSTM, and CNN-BiLSTM architecture and performance shown in TABLE 5, in addition to the traditional ML Model DL is used to automatically learn hierarchical feature representations from speech-gotten inputs, so it was selected. CNN Models are used to identify local auditory patterns and the sequential relationships and properties are captured by LSTM networks. CNN-BiLSTM hybrid design utilizes local and long-term depression related speech patterns.

Table 5: DL Models' Performance

Model	Accuracy	Precision	Recall	ROC-AUC
CNN	97%	98%	98%	0.96
LSTM	89%	89%	100%	0.69
CNN-BiLSTM	89%	89%	100%	0.65

The CNN Model ROC-AUC was determined as 0.93. Results demonstrate the successful extraction of relevant AD markers from speech sources by convolutional techniques. CNN architecture identified local frequency changes, and affective transitions, alongside minor characteristics of the voice associated with depression. The performance of the Model is good, as the ROC-AUC value is high, indicating that class discrimination performed well. The Model was also able to balance the precision and recall, showing that there is significant compromise between the ability to diagnose depressed and reducing false alarms. Retrieved MFCC, Chroma, and MS features are very well suited to represent depression-related auditory features, and can be learnt using Convolutional Feature Extraction (CFE).

The accuracy of LSTM Model being 89%, precision being 89%, recall being 100% and the ROC-AUC being 0.69. All depressed speech samples are correctly recalled by the Models. LSTM networks are designed to learn temporal relationships, so the Model learned depression-related sequential patterns such as speaking rhythms, pauses and emotional continuity. The low accuracy value of the Model and the low ROC-AUC value indicate that temporal information may not be sufficient for classification of depression. Model accurately captured long term speech patterns but might not be able to distinguish between fine acoustic differences between depression and non-

depression classes, leading to false positives. CNN-BiLSTM hybrid Model achieved an accuracy being 89%, precision being 89%, recall being 100% and ROC-AUC being 0.65.

Theoretically, the CNN-BiLSTM Model should be able to learn both spatial and temporal features, allowing it to outperform single DL Models, however, the current implementation did not fully leverage these theoretical benefits. The value of 100% being recall indicates that the Model was able to identify all patients who have been depressed. The CNN Model has a higher ROC-AUC score, indicating better class discriminating ability as compared to the lower score from the lower. In spite of this constraint, the results demonstrate the information related to depression in local acoustic features alongside long-run speech sequences, supporting the hybrid architectural concept. Both MLP and CNN achieved the best results in balance and also reliability, with a comparison of the two Models.

CNN was the Model with the best ROC-AUC and highest precision (0.96 and 99%) values indicating better discrimination. Random Forest achieved good results on recall and ROC AUC of 0.93. The results reveal the effectiveness of neural network-grounded and ensemble learning techniques in depression identification using speech. The detailed comparison of Model performance is shown inside FIGURE 2.

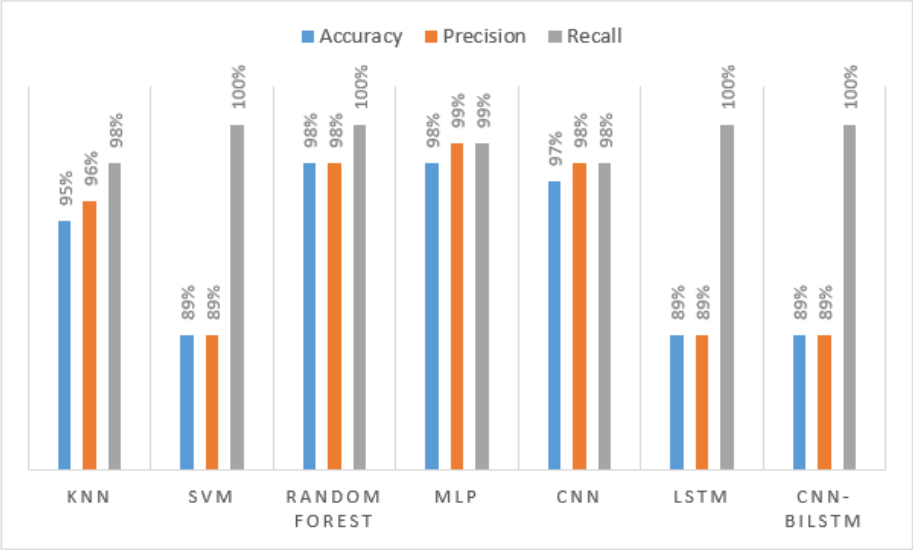

Figure 2: Model Performance Comparisons

The results also highlight the relevance of retrieved properties of acoustics. MFCCs were recorded with the characteristics of speech spectra and perceptual characteristics, and Chroma features were recorded with the characteristics of speech tones and emotions. DL models identified emotional patterns in MS time-frequency data. Combining the characteristics of pitch variation, speech energy, articulation behavior, emotional expressiveness and speaking rhythm into a hybrid feature vector led to better classification results, due to depression affecting all these features. Recall is an important clinical performance metric because failing to identify depression can result in serious MH consequences. All SVM, Random Forest, LSTM, and CNN-BLSTM Models achieve 100% recall rate for the detection of depressed individuals. Combined with false positives, this reduces system reliability, making recall alone not enough.

To further support the stated 98–99% performance, the evaluation was extended with five-fold speaker-independent cross-validation, where the training and validation sets did not contain any recordings from the same speaker. We compute Accuracy, Precision, Recall, F1-score and ROC-AUC on each fold and report them as mean ± standard deviation, which gives a more robust estimation of performance and stability on unseen speakers. Confusion matrices were also created to distinguish true positive, true negative, false positive and false negative for the primary models. The hyperparameters and the random seed were fixed and recorded to ensure reproducibility. The somewhat poorer performance of the deeper CNN-BiLSTM model can be justified by the fact that the acoustic features were averaged over time before the model training, thereby lowering the temporal information available for the sequential architectures. Therefore, the simpler MLP is more suitable for the resulting fixed-length feature representation, while the extra complexity of CNN-BiLSTM cannot take full use of temporal dependencies already removed.

The MLP, CNN and RF Models are practical for use as they have a high recall score as well as a high precision score. The results demonstrate that S&V signals are suitable biomarkers for automated DD. Pitch, speech energy, speech rate, pause duration, MFCC, Chroma features and MS features were accurate in terms of correctly classifying depressive symptoms. Among the classifiers investigated, the MLP has the best ROC-AUC performance. CNN Model was closely followed and the performance of DL-grounded acoustic pattern recognition was revealed. The results provide concept validation and demonstrate that speech-grounded AI systems can effectively detect early depression and MH problems in a non-invasive, and efficient, alongside scalable way.

Limitations: First, the evaluation is performed on the provided dataset and may not reflect the diversity of real world speech among speakers, recording settings and demographics. Second, the target labels are obtained from emotion tags in audio filenames, and are therefore affective proxies rather than professionally diagnosed depression labels. Finally, the findings may not translate directly to spontaneous, organically occurring clinical speech, where speech patterns and contextual elements may differ from those of controlled or dataset-based recordings. Therefore, further validation on larger clinically annotated datasets with spontaneous speech is warranted.

5. CONCLUSION

The results of this work indicated that ML and DL techniques can identify depression grounded in S&V attributes. The MFCC, Chroma and MS representations were accurate in representing depression-related speech characteristics. CNN and RF were performing well while MLP was performing best with an accuracy being 98% and an ROC-AUC being 0.96. The findings demonstrate the potential of speech-grounded AI technologies to both early detect depression non-invasively and on a scalable basis, and make it possible for future MH assessment applications.

References

[1] Haque MN, Sumi AB. AI-Based Psychodiagnostics' Models to Support Early Intervention and Reduce Suicide Risk in Adolescents and Youth: Development and Clinical Validation. Am J Data Sci Anal. 2022;3:40-79.

[2] Milling M, Pokorny FB, Bartl-Pokorny KD, Schuller BW. Is Speech the New Blood? Recent Progress in AI-Based Disease Detection From Audio in a Nutshell. Frontiers in Digital Health. 2022;4:886615.

[3] Ćosić K, Popović S, Šarlija M, Kesedžić I, Gambiraža M, et al. AI-Based Prediction and Prevention of Psychological and Behavioral Changes in Ex-COVID-19 Patients. Front Psychol. 2021;12:782866.

[4] Mohanraj P, Mathan M, Yuvaraj S, Prasanth R. An Al-Driven Multimodal Voice, Text, and Emotion Analysis Framework for Early Mental Health Disorder Detection. Int J Eng Ext Technol Res. 2026;8:4457-4462.

[5] Wang N, Chiong R, Kamil R, Zhang W, Al-Haddad SA, et al. Depression Detection Using Speech Audio and Text: A Comprehensive Review Focusing on Deep Learning Methods. 2024. Authorea preprint: https://www.authorea.com/doi/pdf/10.22541/au.172474607.71425235/v1

[6] Xu X, Li J, Zhu Z, Zhao L, Wang H, et al. A Comprehensive Review on Synergy of Multi-Modal Data and AI Technologies in Medical Diagnosis. Bioengineering. 2024;11:219.

[7] Caulley D, Alemu Y, Burson S, Bautista EC, Tadesse GA, Kottmyer C, Aeschbach L, Cheungvivatpant B, Sezgin E. Objectively Quantifying Pediatric Psychiatric Severity Using Artificial Intelligence, Voice Recognition Technology, and Universal Emotions: Pilot Study for Artificial Intelligence-Enabled Innovation to Address Youth Mental Health Crisis. JMIR Res. 2023;12:e51912.

[8] Tank C, Pol S, Katoch V, Mehta S, Anand A, Shah RR. Depression Detection and Analysis Using Large Language Models on Textual and Audio-Visual Modalities. 2024. ArXiv preprint: https://arxiv.org/pdf/2407.06125.

[9] Zhang L, Zhang S, Zhang X, Zhao Y. A Multimodal Artificial Intelligence Model for Depression Severity Detection Based on Audio and Video Signals. Electronics. 2025;14:1464.

[10] Liu R, Guo J, P ES. Artificial Intelligence-Powered Multimodal Approaches for Depression and Anxiety Detection: A Review. Appl Artif Intell. 2026;40:2668204.

[11] Sharma P. AI-Driven Emotion Recognition Systems for Early Intervention in Mental Health. J Sci Innov Adv Res. 2026;35:25.

[12] Bhayo A. A Multimodal Explainable AI Framework for Real-Time Depression and Anxiety Detection. Spectr Eng Sci. 2026;4:538-546.

[13] Mahur J. Multimodal Emotion-Aware Conversational AI for Mental Health Support: A Systematic Review of Text, Speech, and Behavioral Signal Integration. J Sci Innov Adv Res. 2026;2.