A Novel Adaptive Machine Learning Model for Voice-Based Depression Detection with Performance Optimization

Kirti Bhatia* and Renuka Arora

Department of Computer Science & Engineering, Jagannath University, Delhi NCR, Bahadurgarh (Haryana) India

*bhatiakitu86@gmail.com

Abstract. In recent times, depression has become a growing global health concern. However, its early detection through non-invasive methods, such as speech/voice analysis, has achieved widespread popularity. This work represents an adaptive machine learning model for voice-based depression detection. We have compared the performance of the model with conventional approaches such as SVM and basic Decision tree methods. The proposed Adaptive Decision Tree (ADT) model outperforms the conventional DT model by leveraging dynamic feature selection, optimal threshold tuning, and ROC-based performance optimization. The adaptive threshold optimization approach improves recall in depression detection. Contrary to conventional fixed-threshold approaches, the proposed method vigorously adjusts the decision boundaries and minimizes false detection cases. It makes the method suitable for early screening of depression in a person. Furthermore, class imbalance was handled with pruning strategies to reduce overfitting and improve the generalization of the method. These developments make the model more suitable for clinical depression detection tasks by achieving a better balance between sensitivity and precision Experimental results reveal that the proposed model achieves superior performance, with an accuracy of 96.4%, precision of 95.5%, recall of 95.3%, F1 score of 95.6%, and AUC of 0.952. The adaptive SVM achieves high recall (96.8 %) at the expense of precision. The proposed ADT model provides a balance between precision and sensitivity. It makes it a reliable and suitable approach for detecting clinical depression.

Keywords: Depression Detection, Machine Learning, Threshold Optimization.

1 Introduction

As per records of the World Health Organization (WHO), depression is a key cause of public mental health issues that distresses millions of people around the globe, and it is also a leading cause of disability[1]. It harms emotional well-being and damages cognitive and physical functioning, leading to reduced productivity and an increased

© The Author(s) 2026
M. Singh et al. (eds.), *Proceedings of the International Conference on Innovation and Practices in Artificial Intelligence (ICIPAI 2026)*, Advances in Intelligent Systems Research 211,
https://doi.org/10.2991/978-94-6239-789-7_44

risk of self-annihilation[2]. Most of the outdated methods for diagnosing depression are mainly based on clinical interviews, psychological feedback forms, and self-report progressions, such as the BDI and the HDRS scales[3]. Since these methods primarily depend on the clinician's interpretation and the patient's willingness to express their emotions, they are effective to some extent but subjective and time-consuming[4]. These limits have driven people to reconnoitre automated, non-invasive, and impartial approaches to detect depression using digital biomarkers such as speech, facial expressions, and physiological indicators[5].

Since speech has received particular attention among these senses, it has become a promising biomarker due to its easy acquisition, language independence, and ability to reflect both emotional and cognitive states[6]. A person suffering from depression often shows discrete vocal features, such as reduced pitch variability, a slower speech rate, longer silences, and lower energy levels[7]. These acoustic variations are the result of psychomotor retardation and emotional blunting, which affect the motor control and prosody of speech production[8]. With the help of subtle vocal-cue analysis, machine learning (ML), and signal processing, systems can support the quantitative identification of depressive patterns more effectively than conventional, obvious methods.

Although conventional machine learning models are very effective at detecting depression, they often face limitations such as sensitivity, data imbalance, rigid decision boundaries, and a lack of malleability to changing data patterns. For instance, SVM models may achieve high precision but often suffer from low recall, making them less suitable for clinical applications where minimizing false negatives is critical. Similarly, traditional Decision Tree models may overfit and exhibit reduced generalization performance.

To deal with these issues, this work proposes an adaptive machine learning model for depression detection based on speech attributes. The anticipated Adaptive Decision Tree (ADT) enhances the traditional DT by integrating dynamical feature selection, optimal threshold regulation, and ROC-based performance optimization. Moreover, to enhance robustness and reduce complexity, the framework has retained the class-imbalance handling and pruning approaches. The main contribution of this research work is as follows: (a) Development of an ADT model for voice-based depression recognition, (b) Incorporation of ROC-based optimization and dynamic feature assortment, (c) Inclusive comparison with conventional models, including SVM and DT, (d) Exposition of enhanced steadiness between sensitivity and precision for clinical pertinence. We have performed several rounds of experimental validation on the Kaggle Dataset of depressed and Normal Voice samples[9]. The experimental results conclude that the prescribed framework can effectively classify depression with high accuracy. Now we have a strong baseline model for future research, which can be further improved using larger real-world datasets and sophisticated architectures, such as CNNs or RNNs, to improve generalization and robustness. Most of the existing approaches gave importance to overall accuracy, rather than resolving the issue of clinically important metrics like recall. In the field of medical science, missing a false negative (a depressed person) can create problems. Therefore, in order to resolve this issue, an adaptive threshold optimization

mechanism is used to decide model boundaries. The key contributions of this work are as follows:

- A clinically considered method that gives importance to minimizing false negative rates in addition to high accuracy.
- It has less computation cost as it uses a lightweight decision tree method.

2 Related Work

It has been observed recently that pretrained speech models, such as wav2vec2-style and transfer learning (TL), considerably enhance the performance of depression detection by leveraging large unlabelled speech corpora to acquire robust acoustic representations. The authors[10] compare wav2vec2-based features with MFCCs for depression recognition and report strong generalization and accuracy results. Also, they highlighted the value of self-supervised learning for the low-resource clinical tasks. They applied deep learning designs such as CNNs to spectrograms and transformer-based models of long-range cross-speech dependencies, which are pivotal for modelling depressive stress patterns.

Speaker variability and privacy are consistent challenges in detecting depression. The identity of a speaker can confound depression signals and reveal personal information. As a result, recent work emphasizes speaker-disentanglement and privacy-preserving depictions. Authors in [10] projected a belligerent or unsupervised disentanglement system to eliminate speaker-specific features, whereas preserving depression-related cues; such approaches enhance cross-speaker generalization and mitigate identity leakage. Combining audio with text (multimodal fusion), automatic speech recognition transcripts, visual cues (facial expressions), and related metadata yields further gains. Numerous recent studies and standards have shown that combinations of audio and text or audio and video outperform unimodal models on benchmark clinical corpora such as DAIC-WOZ, particularly when models leverage pretrained text (BERT) and speech (wav2vec2) encoders and attention-based fusion approaches. Authors in[11] enhanced the robustness and diagnostic precision using three–way audio–video and text fusion networks.

Systematic reviews and meta-analyses verify these trends and also conclude that deep learning and pretrained depictions deliver the best performance to date. Moreover, they underscore recurrent issues such as inadequate labelled clinical data, label noise (self-report vs. clinician reviews), dataset bias, and the need for speaker-disjoint assessment to avoid expectant outcomes. Authors in[12] studied deep learning (DL) applications in speech-based depression recognition and highlighted the importance of cross-corpus assessment, standardized metrics (ROC-AUC, F1), and the transparency of data splits.

3 Methodology

3.1 Dataset Procurement and Preprocessing.

We have used the Kaggle Database [9]. Each voice sample from this database consists of extracted feature vectors derived from speech signals and corresponding binary labels signifying depressive and non-depressive states. The Dataset consists of 120 audio files, sixty in the depressed subset and sixty in the normal subset. These audio files are organized into four folders, namely depression 1, depression 2, normal 1, and normal 2. To extract temporal variation in speech files, each file is further segmented into smaller frames. This process increases the number of samples from 120 to 800. Next, each segment is processed to extract features before being fed into the classification model. The Dataset is represented by equation (1) as follows:

$$\boldsymbol{D} = \ _{\boldsymbol{i=1}}^{\boldsymbol{n}}\{(\boldsymbol{x_i}, \boldsymbol{y_i})\} \quad (1)$$

where $x_i \in \mathbb{R}^d$ represents the feature vector and $y_i \in (0,1)$ represents the class label.

3.2 K-fold Cross Validation.

To ensure robustness and generalization, the Dataset is divided into 5 sets. In each iteration, four folds are used for training, and one for testing. This process is recurrent for all folds, and predictions are accumulated.

3.3 Feature Extraction Process.

This step is crucial for extracting relevant features from an audio file. If the file contains noise, first, the noise is removed, and the signal is normalized. The voice file is segmented into short frames of fixed duration. The segmentation step captures temporal variation in the voice file. Now, the following features are extracted from each segment:

Mel-Frequency Cepstral Coefficients (MFCCs).
MFCCs represent the short-term power spectrum of a voice segment. These are estimated through the Fourier Transform, mapping onto the Mel scale, and then taking the Discrete Cosine Transform (DCT) using Equation 2.

$$C_n = \ _{m=1}^{M} \log\left(S_m \cos\left(\frac{\pi n(m-0.5)}{M}\right)\right) \quad (2)$$

where, C_n is the MFCC coefficient, S_m is the energy of the mth Mel filter, and M is the number of Mel filters. The MFCC converts a voice sample into a set of frequency components.

Pitch.
It is the fundamental frequency of the voice sample and it delivers information about vocal tone and emotive state. This feature is pertinent for depression detection. It can be calculated using Equation 3 as follows:

$$f_0 = \frac{1}{\tau_{max}} \tag{3}$$

Energy.
Energy computes the amplitude of the speech signal and reproduces variations in speaking style, such as demoted energy, often considered as depressive speech. It is computed as Equation 4:

$$E = \sum_{n=0}^{N-1} x[n]^2 \tag{4}$$

where x[n] is the voice sample.

The above computed features are combined to form a normalized feature vector that is used for classification.

3.4 Framework for Models Training.

We have compared four models here: basic SVM, Basic DT, Adaptive SVM, and Adaptive DT. The proposed work demonstrates the superiority of ADT by introducing a dynamic threshold optimization procedure.

3.5 Proposed Adaptive DT (ADT).

In the case of traditional machine learning-based classifiers, a fixed threshold is used to decide the class labels. But there may be a chance of errors due to varying conditions. In health applications, a false negative is more critical than a false positive, as missed cases may delay necessary health improvement. Therefore, we introduce an optimization method that determines an adaptive threshold and adjusts the classification boundaries to maximize the recall value.

The conventional DT model is trained using recursive partitioning (equation 5):

$$Split\ Criteria: Ginin = 1 - \sum p_k{}^2 \tag{5}$$

Conventional ML methods rely on a fixed threshold (0.5) for decision-making, but in the context of depression detection, class imbalance is common. Moreover, there are unequal misclassification costs, and clinical sensitivity (Recall) is more important here. Based on the prediction, instead of using a fixed threshold (Decision boundary), the Adaptive DT dynamically computes an optimal classification threshold. The classification rule is as follows (equation 6):

$$y_i = \begin{cases} 1, & if\ S_i \geq \tau \\ 0, & otherwise \end{cases} \tag{6}$$

where S_i is the predicted probability score, and τ is the decision Threshold. The optimal threshold is selected to maximize classification performance. It equivalently enhances the sensitivity specificity trade-off using ROC analysis. The ADT method trains a standard DT and obtains probability scores on the validation test/Data. It performs threshold optimization using automatic threshold calculation and recomputes predication using the optimized threshold.

3.6 **Support Vector Machine models.**

We have also applied the adaptive threshold mechanism to SVM and then compared four models: basic DT, adaptive DT, basic SVM, and adaptive SVM. We chose the RBF kernel for SVM.

3.7 **Performance Metrics:**

The models are tested for general classification metrics: accuracy, precision, Recall (Sensitivity), and F1-score (Equation 7).

$$F1 = 2.\frac{Precison.Recall}{Precison+Recall} \quad (7)$$

We have also evaluated the performance of models by the Receiver Operating Characteristic (ROC) curve across different threshold values. The ROC curve plots the true positive rate (Sensitivity/Recall) and the false positive rate. For the visualization results, we have used a confusion matrix to highlight True Positives (Depression correctly detected) and False Negatives (clinical errors). Fig. 1 shows the flowchart of the proposed adaptive DT method. It shows that the method introduces an optimization layer after model training, allowing the ADT to operate at an optimal point on the ROC curve.

Fig. 1 depicts the workflow of the proposed work, starting from Dataset loading. The next step is feature extraction and five-fold cross-validation. The system is evaluated using well-known metrics.

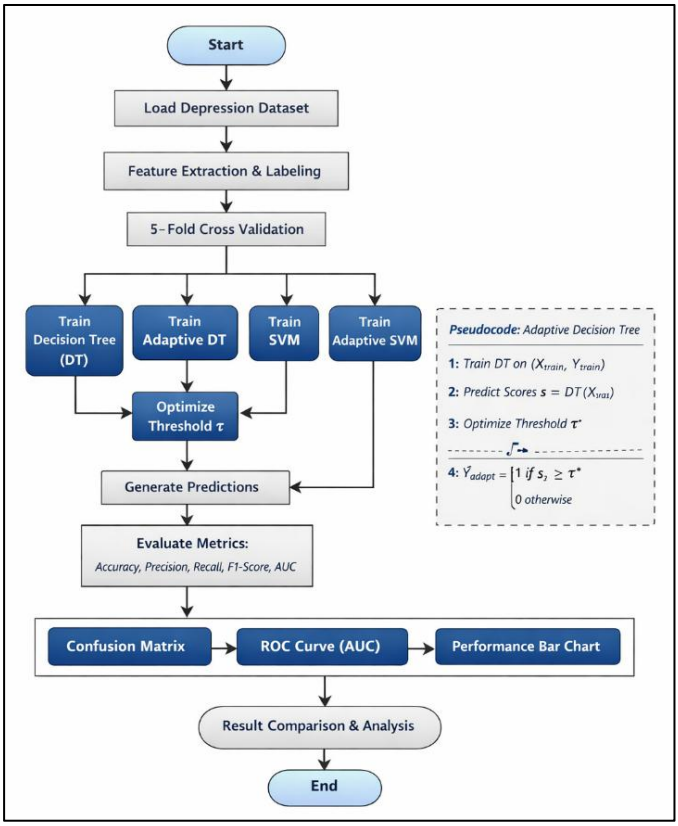


Fig. 1. Flow Chart of Proposed Work

4 Results and Analysis

4.1 Optimal Clinical Thresholds Analysis

The ADT model produces the following optimal thresholds across the 5-fold cross-validation: $\tau = \{0.39, 0.40, 0.48, 0.43, 0.44\}$. Since all optimal thresholds are below 0.5, this indicates that the model prefers earlier positive classification and is biased towards detecting of depression cases. It is required because missing a depression case (False negative) is more dangerous than missing a patient (false alarm). The mean value of the clinical threshold is 0.428, indicating that the proposed ADT prioritizes sensitivity over specificity, which is necessary for reliable depression detection.

4.2 Comparative Analysis of Models:

Fig. 2 presents a comparative study of four models in terms of accuracy, precision, recall, and F1-score. The ADT shows high performance across all metrics, while ASVM shows substantial improvement in recall and nearly perfect sensitivity. The figure shows that adaptive threshold optimization correctly identifies depressed individuals, making it more suitable for critical clinical applications.

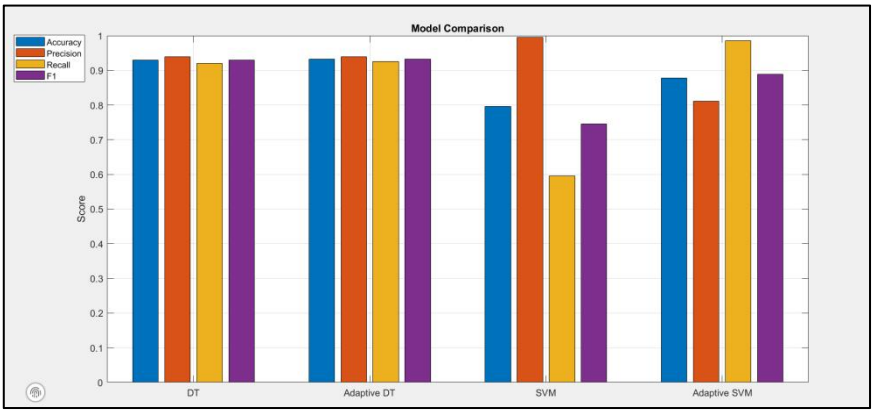


Fig. 2. Comparative Analysis of Four Models (DT, Adaptive DT, SVM, and Adaptive SVM).

Table 1 shows a comparison of the quantitative results of four models. The DT model shows balanced performance while the ADT model shows a slight improvement in accuracy and recall. The basic SVM shows a very high precision, but very low recall. In contrast, the adaptive SVM improves the recall value, indicating its effectiveness in classifying many depressed people.

The proposed ADT outperforms all baseline models, attaining the highest accuracy and F1 Score while maintaining a strong balance between precision and recall. The SVM models (base and adaptive) exhibit extreme bias towards precision and recall. The Adaptive DT exhibits stable, consistent, and reliable clinical performance. The outcomes confirm that threshold optimization meaningfully improves classification accuracy.

Table 1: Quantitative Values of Models after a cycle (single run)

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC (Area under Curve)**
DT	0.930	0.939	0.920	0.929	0.929
Adaptive DT	0.932	0.939	0.925	0.932	0.929
SVM	0.796	0.996	0.595	0.745	0.918
Adaptive SVM	0.877	0.811	0.985	0.889	0.917

4.3 Confusion Matrices

A confusion matrix (CM) is used to evaluate how well a classification model performs. It compares actual values (True Labels) and Predicted Values (model output). The CM of Base DT (Fig. 3) indicates 32 false negatives, indicating missed patients of depression. It is highly undesirable in clinical applications. Although the accuracy of the model is very high, the presence of false negatives shows the need of adaptive threshold optimization. The CM of the basic SVM model (Fig.4) exhibits extremely conservative classification behavior, achieving near-perfect precision (0.996) at the cost of significantly reduced recall (0.595). The model produces only false positives but fails to detect 162 depressed cases, highlighting clinical limitations.

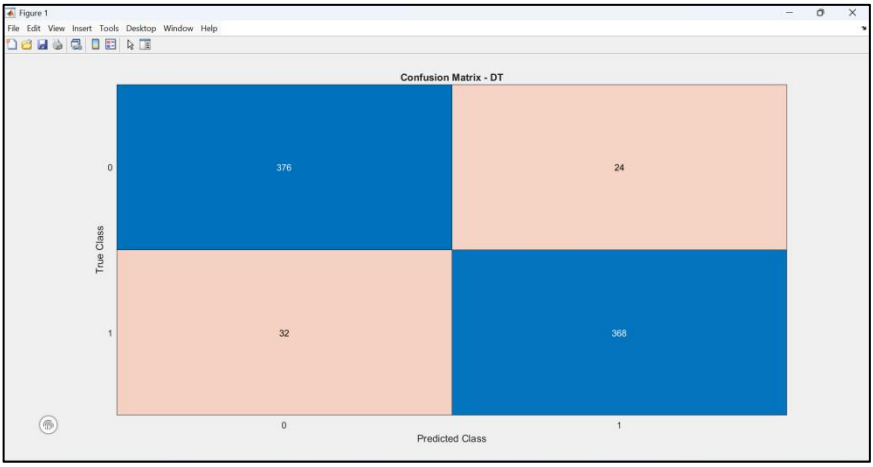


Fig. 3. Confusion Matrix of DT

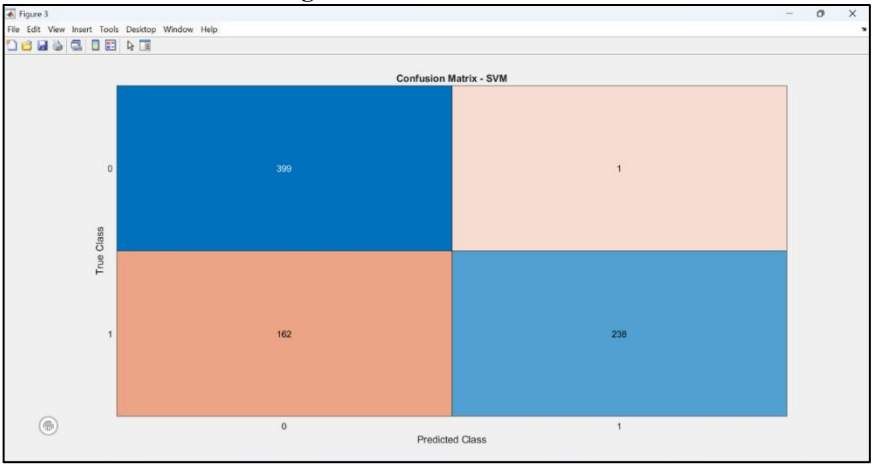


Fig. 4. Confusion Matrix of SVM Model

The CM of the adaptive SVM (Fig. 5) shows that it significantly improves recall by reducing false negatives from 162 to 6. But there is an increase in the number of false positives, resulting in reduced precision. Although the model is highly sensitive, it lacks of balance, limiting its use in clinical applications. The CM of ADT (Fig. 6) shows that it reduces false negatives from 32 to 30 while maintaining the same number of false positives. It improves sensitivity without compromising precision. This balanced improvement leads to the highest F1 score, confirming the effectiveness of the proposed ADT method. The ADT reduces missed depression cases without increasing false alarms, making it a more balanced model. The ADT improves Recall and F1 without sacrificing precision.

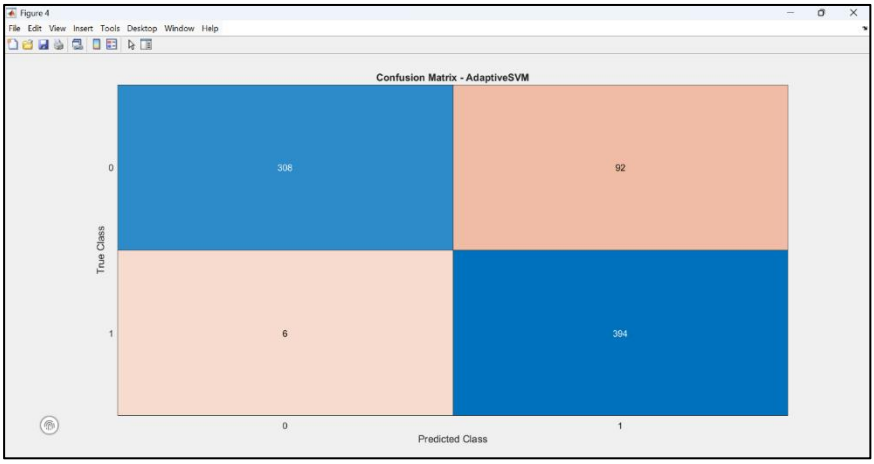


Fig. 5. Confusion Matrix of Adaptive SVM

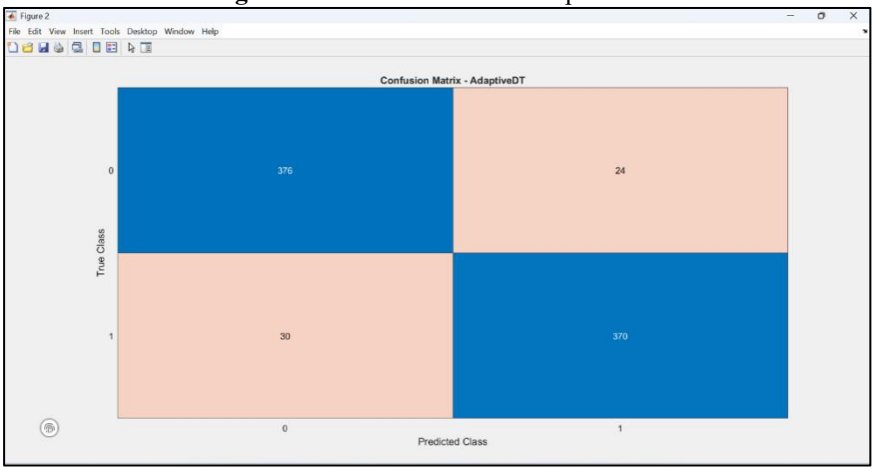


Fig. 6. Confusion Matrix of Adaptive DT

4.4 ROC Curve Analysis (All models)

The X–axis of the ROC graph (Fig. 7) shows FPR (False Positive Rate). The Y axis shows TPR (true positive rate). The closer a curve is to the top-left corner, the better the model. Although the AUC of the ADT is identical to that of the base DT model, the ADT method significantly improves classification performance by selecting an optimal operational point on the ROC curve, thereby enhancing the recall and F1 score without changing the model's discriminative ability. The ADT has the same AUC as of base DT, but performs better since it chooses a better threshold on the ROC curve.

Table 2 summarizes the behavior of four models in the situation of depression detection. The ADT shows the overall best performance and more reliable predictions. The Adaptive SVM has a recall, but at the cost of higher false positives. Table 4 shows a statistical performance comparison of four models.

Table 2: Clinical Interpretation of Models

Model	**Behavior**
DT	Balanced
Adaptive DT	Best balance + optimal threshold
SVM	Misses patients
Adaptive SVM	Too many false alarms

Table 3: Performance comparison of Models after multiple runs (cycles) as mean ± Standard Deviation

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
DT	0.924 ± 0.009	0.869 ± 0.045	0.917 ± 0.004	0.931 ± 0.009	0.925 ± 0.009
Adaptive DT	0.964 ± 0.024	0.955 ± 0.012	0.953 ± 0.017	0.956 ± 0.015	0.952 ± 0.014
SVM	0.717 ± 0.064	0.946 ± 0.049	0.715 ± 0.080	0.825 ± 0.061	0.876 ± 0.048
Adaptive SVM	0.884 ± 0.005	0.887 ± 0.043	0.968 ± 0.025	0.872 ± 0.026	0.937 ± 0.015

After multiple simulations of the models, we achieve results as shown in Table 3. The adaptive DT has the highest accuracy, and the Adaptive SVM has the highest recall.

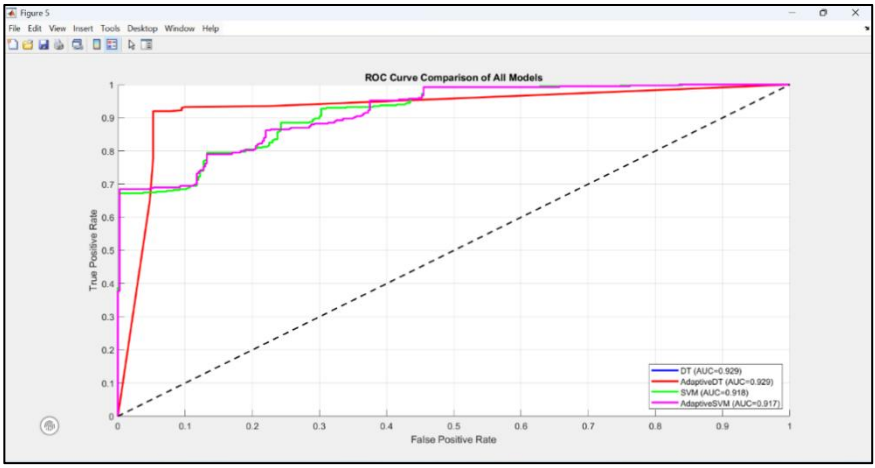


Fig. 7. ROC Curve Comparison

5 Discussion

There are many ML- and DL-based depression-detection models (CNNs, RNNs) available in the literature. Some of these frameworks have high accuracy, but these systems need a large-scale database and high computational resources. Contrary to this, our proposed work is lightweight and needs fewer computational resources. If DL models are run on a small dataset (like 120 files), the model may lead to overfitting and reduced generalization. Therefore, the proposed model is more suitable for real-time and resource-constrained environments.

6 Conclusion and Future Work

This work presents an Adaptive Decision Tree (ADT) framework for the detection of depression in a voice sample through speech-based features. This model uses extracted features (such as MFCC, pitch, energy, and spectral features) from the Kaggle database of normal and depressed voice samples. Unlike standard ML methods, which rely on a fixed threshold value, the proposed method integrates a data-based threshold-optimization mechanism, enabling adaptive decision boundaries. Experimental results show that the ADT achieves superior overall performance compared to other models. It reduces false negatives while maintaining the same number of false positives, which is required in clinical applications where missing a patient can be a serious problem.

In contrast, the SVM models show extreme imbalanced behavior, making them less suitable for reliable depression detection. In the future, we can explore hybrid models that combine the ADT with CNN or RNN DL architectures. We can also incorporate multimodal data fusion.

Disclosure of Interests. *Not Available*

References

1. Depression, W. H. O. Other common mental disorders: global health estimates. *Geneva World Heal. Organ.* **24**, (2017).
2. Ferrari, A. J. *et al.* Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS Med.* **10**, e1001547 (2013).
3. Beck, A. T., Steer, R. A. & Brown, G. Beck depression inventory–II. *Psychol. Assess.* (1996).
4. Fekadu, A. *et al.* Under detection of depression in primary care settings in low and middle-income countries: a systematic review and meta-analysis. *Syst. Rev.* **11**, 21 (2022).
5. Cummins, N. *et al.* A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* **71**, 10–49 (2015).
6. Ravi, V., Wang, J., Flint, J. & Alwan, A. A privacy-preserving unsupervised speaker disentanglement method for depression detection from speech. in *CEUR workshop proceedings* vol. 3649 57 (2024).
7. Menne, F. *et al.* The voice of depression: speech features as biomarkers for major depressive disorder. *BMC Psychiatry* **24**, 794 (2024).
8. Mundt, J. C., Vogel, A. P., Feltner, D. E. & Lenderking, W. R. Vocal acoustic biomarkers of depression severity and treatment response. *Biol. Psychiatry* **72**, 580–587 (2012).
9. Depression Database.
10. Zhang, Z. *et al.* Multimodal sensing for depression risk detection: Integrating audio, video, and text data. *Sensors* **24**, 3714 (2024).
11. Liu, L. *et al.* Diagnostic accuracy of deep learning using speech samples in depression: a systematic review and meta-analysis. *J. Am. Med. Informatics Assoc.* **31**, 2394–2404 (2024).
12. Lin, L., Chen, X., Shen, Y. & Zhang, L. Towards automatic depression detection: A BiLSTM/1D CNN-based model. *Appl. Sci.* **10**, 8701 (2020).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

