

SonoMind: deep learning-based voice analysis for mental health monitoring

JACOB JITHIN¹, K S KANNAN¹

¹Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankoil, Srivilliputhur, Tamil Nadu.

Summary. Depressive disorder is a major global health concern and is often underdiagnosed due to stigma, unreliable self-reporting, and limited access to proper mental health screening tools. Despite recent advances, most existing automated approaches depend on questionnaires or multi-modal data, while efficient and reliable voice-only clinical detection frameworks remain limited, creating a clear research gap. Motivated by the need for a non-invasive, objective, and privacy-preserving diagnostic alternative, this study proposes SonoMind, an adaptive deep learning framework for early depression detection using voice signals. The methodology incorporates Adaptable Spectral Pairing for effective noise reduction, SynchroSonic Learning for synchronized feature extraction, and Adaptive Krill-Wolf Optimization for optimal feature selection, followed by a neural classification stage. The framework was evaluated using the publicly available DAIC-WOZ clinical interview dataset. Experimental results show that SonoMind achieves 97.22% accuracy, 100% precision, 95.92% recall, 97.92% F1-score, MAE of 0.027, and RMSE of 0.1666. These results confirm the robustness and reliability of the system, demonstrating its potential as a scalable and supportive tool for mental health professionals in voice-based depression screening.

Key words. Adaptable Spectral Pairing technique, depression, mental health prediction, SynchroSonic Learning, Voice analysis.

SonoMind: analisi vocale basata sull'apprendimento profondo per il monitoraggio della salute mentale.

Riassunto. Il disturbo depressivo è una delle principali preoccupazioni sanitarie a livello globale ed è spesso sottodiagnosticato a causa dello stigma, dell'inaffidabilità dell'auto-segnalazione e dell'accesso limitato a strumenti di screening adeguati per la salute mentale. Nonostante i recenti progressi, la maggior parte degli approcci automatizzati esistenti si basa su questionari o dati multimodali, mentre i framework di rilevamento clinico basati esclusivamente sulla voce, efficienti e affidabili, rimangono limitati, creando un'evidente lacuna nella ricerca. Motivato dalla necessità di un'alternativa diagnostica non invasiva, oggettiva e rispettosa della privacy, questo studio propone SonoMind, un framework di apprendimento profondo adattivo per la diagnosi precoce della depressione mediante segnali vocali. La metodologia incorpora l'accoppiamento spettrale adattabile per un'efficace riduzione del rumore, l'apprendimento sincronizzato SynchroSonic per l'estrazione sincronizzata delle caratteristiche e l'ottimizzazione adattiva Krill-Wolf per la selezione ottimale delle caratteristiche, seguita da una fase di classificazione neurale. Il framework è stato valutato utilizzando il dataset di interviste cliniche DAIC-WOZ, disponibile al pubblico. I risultati sperimentali mostrano che SonoMind raggiunge un'accuratezza del 97,22%, una precisione del 100%, un recall del 95,92%, un punteggio F1 del 97,92%, un MAE di 0,027 e un RMSE di 0,1666. Questi risultati confermano la robustezza e l'affidabilità del sistema, dimostrandone il potenziale come strumento scalabile e di supporto per i professionisti della salute mentale nello screening della depressione basato sulla voce.

Parole chiave. Analisi vocale, apprendimento sincronico sonico, depressione, previsione della salute mentale, tecnica di accoppiamento spettrale adattabile.

Introduction

Depression is a major public health concern worldwide, significantly impairing emotional, cognitive, and social functioning, and contributing substantially to disability across populations^{1,2}. The disease burden is especially pronounced among children and young adults, where depressive disorders account for nearly 20% of disease-related disability and affect over 293 million individuals globally, resulting in more than 44 million disability-adjusted life years (DALYs)^{3,4}. Despite the availability of effective interventions, early diagnosis remains difficult

due to the reliance on subjective screening tools, inadequate clinical resources, and the absence of scalable, objective diagnostic systems⁵. Consequently, a significant proportion of individuals remain undiagnosed or receive delayed treatment, increasing the risk of chronic disability and suicide.

Recent technological advancements have enabled the development of automated mental health screening systems, among which speech-based depression detection has emerged as a promising and non-invasive approach^{6,7}. Emotional and cognitive disturbances associated with depression manifest in vocal attributes such as pitch, speaking rate, and prosodic variation, making speech a valuable behavioral biomarker for

mental state assessment^{8,9}. Compared with conventional diagnostic methods and physiological sensing, voice-based approaches offer advantages in terms of cost, privacy, and ease of deployment, enabling large-scale screening in real-world environments.

Machine learning and deep learning techniques have significantly improved affective computing systems by extracting meaningful representations from complex speech signals and enabling accurate classification^{10,11}. While multimodal analysis using speech, facial expressions, and text has further enhanced detection performance^{12,13}, voice analysis remains particularly attractive due to its minimal intrusiveness and scalability. In addition, wearable sensors and smartphone-based monitoring systems facilitate continuous behavioral tracking^{14,15}; however, such methods often introduce privacy concerns and operational constraints that limit real-world adoption. Despite these advances, several challenges remain unresolved, including acoustic variability, language and cultural differences, background noise, and inconsistent recording quality, all of which affect model generalization. Furthermore, many existing approaches demonstrate limited robustness and lack clinical reliability in practical deployment scenarios. Therefore, a need exists for an improved deep learning-driven framework that can model subtle emotional patterns in speech while maintaining stability across diverse conditions. In response to these challenges, this work proposes a novel voice-based depression prediction framework designed to improve feature representation, enhance classification accuracy, and increase robustness across datasets. The goal is to contribute a scalable and objective solution to support early diagnosis and improve accessibility to mental healthcare. The primary contributions of this research are as follows:

- A novel noise-aware modeling approach is introduced to enhance speech signal quality while preserving psychologically meaningful vocal patterns, enabling more reliable extraction of emotional and cognitive cues from audio recordings.
- A learning strategy is proposed to improve generalization in speech-based mental health prediction by reducing overfitting and balancing representation learning with computational efficiency, ensuring stable performance across diverse datasets.
- An adaptive learning mechanism is developed that allows the model to dynamically adjust its internal representations based on task complexity, thereby improving classification accuracy in real-world mental health assessment scenarios.
- A feature relevance modeling framework is presented to automatically prioritize informative acoustic cues associated with mental states, leading to improved prediction precision and reduced dependency on redundant or noisy inputs.
- A robust classification pipeline is designed to enable consistent and reliable differentiation among mental health conditions from speech data, offering improved sensitivity to subtle emotional variations in voice patterns.

Following is the arrangement of the remaining portion of the manuscript: a summary of the general findings is provided in the conclusion. The second section examines the body of existing literature. The third section provides a detailed explanation of the research technique. The fourth section addresses the application of the suggested method and gives the results.

Literature review

Recent studies highlight the growing role of artificial intelligence in digital mental health as a decision-support tool rather than a replacement for clinical diagnosis. EEG-based methods proposed by Chen et al.¹⁶ and Zhang et al.¹⁷ offer neurophysiological insights but depend on specialized equipment, limiting everyday clinical use. Speech-based approaches, such as the multilingual framework by Guo et al.¹⁸ and the interpretable model by Ntalampiras¹⁹, demonstrate the value of vocal biomarkers, though challenges remain in training complexity and robustness.

Text-based models developed by Karamat et al.²⁰ and Daru et al.²¹ effectively analyze linguistic content but may not fully reflect emotional or physiological states. Multimodal systems introduced by Zhou et al.²² and Jin et al.²³ improve detection performance but increase technical complexity and privacy concerns. Survey-based studies by Fu et al.²⁴ and Shukla et al.²⁵ contribute population-level insights but lack real-time personal assessment capability. In contrast, the proposed SonoMind framework complements clinical workflows by providing a scalable, speech-based, and non-invasive screening approach that supports early detection, continuous monitoring, and clinician decision-making. Although AI-based mental health assessment methods have shown promising results, their real-world adoption remains limited due to several challenges. Reliance on EEG systems restricts scalability, complex models increase computational cost, and small or biased datasets limit generalization across populations. Additionally, dependence on surveys and handcrafted features reduces diagnostic reliability, while weak noise-handling degrades performance in real-world conditions. Lack of interpretability further reduces clinical trust, keeping many systems misaligned with practical healthcare needs.

Proposed work

Conventional approaches to mental health prediction face major limitations, including stigma, privacy concerns, bias, inaccurate predictions, intrusive data collection, and reliance on subjective reporting, which often lead to misdiagnosis and inequitable care. To overcome these challenges, the Adaptive Sonic Synergy Mental Discovery Architecture is proposed as an ethical and effective alternative. The framework begins with advanced noise removal using Adaptable Spectral Pairing to preserve critical vocal cues while reducing background interference. Feature extraction is then enhanced through SynchroSonic Learning, which reduces overfitting, improves generalization, and captures meaningful representations from speech data. Adaptive Krill-Wolf Optimization is used to select the most informative features, improving prediction accuracy. Finally, classification is performed using dense layers and a sigmoid function to generate reliable outputs. Overall, the framework provides a more robust, accurate, and practical solution for voice-based mental health assessment. Figure 1 represents the block diagram of the proposed model.

DATASET DESCRIPTION

This study is entirely based on a secondary analysis of the publicly available DAIC-WOZ dataset. The DAIC-WOZ (Distress Analysis Interview Corpus - Wizard-of-Oz) dataset serves as the primary data source for this research. It is a widely recognized, publicly available multimodal corpus used for automatic mental health assessment, particularly for detecting depression. The dataset comprises 359 semi-structured clinical interviews, capturing audio, visual, and textual modalities, along with rich metadata such as participant demographics, emotional states, clinical assessments, and diagnostic labels. Each interview includes high-quality audio recordings and transcribed dialogues, facilitating the development and evaluation

of models that leverage multimodal data for psychological distress analysis.

For this study, audio data from the DAIC-WOZ dataset were analyzed, with depression severity measured using the PHQ-8 (Patient Health Questionnaire-8) scale. Audio features were extracted using the open-source software COVAREP, which computes 74 low-level descriptors at 10-millisecond intervals. Nine features exhibiting near-zero variance were excluded, as they could negatively affect the computation of correlation coefficients necessary for constructing reliable similarity graphs²⁶.

PREPROCESSING USING ADAPTABLE SPECTRAL PAIRING TECHNIQUE

ASP combines Spectral Subtraction and the Kalman Filter to reduce non-stationary background noise while preserving diagnostically relevant speech patterns.

The noisy speech is transformed using FFT:

Noise is estimated using VAD-based non-speech frames:

$$N^{\wedge}(k,t)=VAD(Y) \tag{2}$$

Spectral subtraction is applied:

$$|S^{\wedge}(k,t)|=|Y(k,t)|- \alpha|N^{\wedge}(k,t)| \tag{3}$$

To reduce musical noise and distortion, the Kalman Filter refines the spectral estimate:

$$S^{\wedge}_t=AS^{\wedge}_{(t-1)}+K_t(Y_t-HS^{\wedge}_{(t-1)}) \tag{4}$$

Finally, the clean signal is reconstructed using IFFT:

$$S^{\wedge}(t)=IFFT(S^{\wedge}(k,t)) \tag{5}$$

This process results in a denoised waveform and a cleaner spectrogram in figure 2.

FEATURE EXTRACTION USING SYNCHROSONIC LEARNING

The SynchroSonic Learning (SSL) module employs a transfer-learning-based deep CNN combined with a bidirectional layer, as illustrated in figure 3. After preprocessing, the denoised audio is transformed into time-frequency representations and passed through a sequence of stacked convolutional layers and max-pooling operations, which capture localized

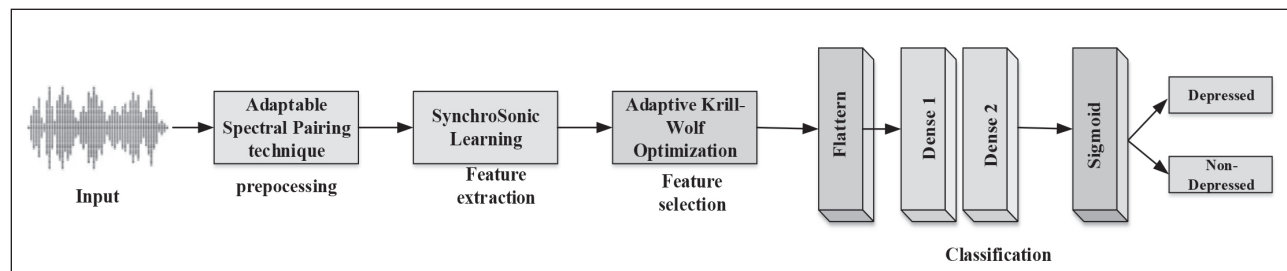


Figure 1. Block diagram of the proposed methodology.

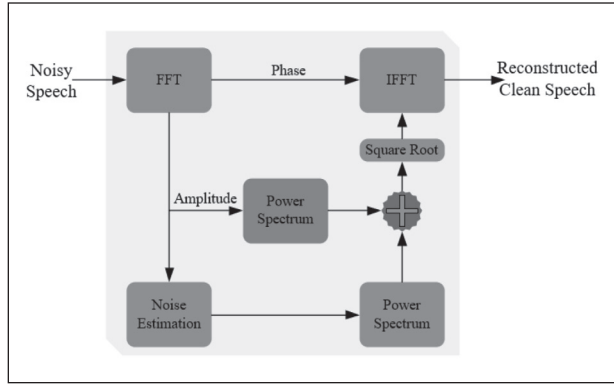


Figure 2. Block diagram illustrating Spectral Subtraction Method.

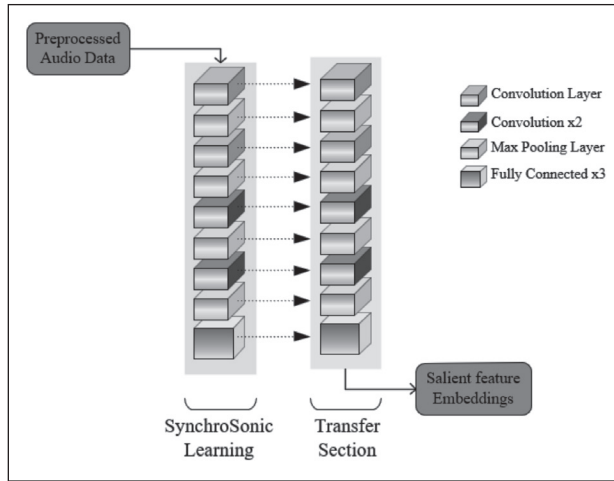


Figure 3. SSL architecture using transfer learning strategy.

spectral and temporal patterns. A bidirectional sequence-learning layer is then incorporated to model contextual dependencies within the speech signal, enhancing the detection of subtle variations associated with depressive states. To maintain stability and prevent overfitting, the pre-trained weights inherited from the source model are partially frozen during fine-tuning. Additionally, BL-SMOTE is applied to address class imbalance by generating boundary-focused synthetic samples for the minority class. Through this structured feature extraction process, SSL produces rich, hierarchical embeddings that preserve essential acoustic and prosodic markers relevant for accurate depression classification.

FEATURE SELECTION PHASE

The model employs the AKWO algorithm to improve accuracy by removing redundant and irrelevant features while reducing computational complexity.

By combining Krill Herd and Grey Wolf Optimization, AKWO dynamically balances exploration and exploitation for efficient global search. This adaptive mechanism reduces overfitting, handles non-linearities, and produces robust feature subsets for reliable voice-based depression prediction.

AKWO algorithm

The algorithm aims to select the most relevant features that enhance classification accuracy while minimizing redundancy and noise. It begins by initializing a population of candidate feature subsets, which are iteratively optimized to identify the most effective selection.

$$X_i = [x_{i1}, x_{i2}, \dots, x_{id}] \text{ where } x_{ij} \in \{0,1\} \quad (8)$$

The fitness function is defined as:

$$F(x_i) = \alpha * Err(x_i) + (1 - \alpha) * \frac{\sum_{j=1}^d x_{ij}}{d} \quad (9)$$

where $Err(x_i)$ is the classification error (i.e., 1 minus the accuracy) of a classifier trained on the selected features x_i and $\alpha \in [0,1]$ balances the trade-off between classification error and subset size.

The AKWO algorithm alternates between two optimization strategies in each iteration, the GWO algorithm is applied. It identifies the leading wolves (α, β, δ) based on fitness and updating the positions of solutions accordingly to exploit the search space effectively. The best solution based on fitness is:

$$X_\alpha = \arg \min (X_i) \quad (10)$$

The position update equations, simulating the wolves encircling prey, are given by:

$$\left. \begin{aligned} D_\alpha &= |C_1 \cdot X_\alpha - X| \\ D_\beta &= |C_2 \cdot X_\beta - X| \\ D_\delta &= |C_3 \cdot X_\delta - X| \end{aligned} \right\} \quad (11)$$

$$\left. \begin{aligned} X_1 &= X_\alpha - A_1 D_\alpha \\ X_2 &= X_\beta - A_2 D_\beta \\ X_3 &= X_\delta - A_3 D_\delta \end{aligned} \right\} \quad (12)$$

$$X^{new} = \frac{X_1 + X_2 + X_3}{3} \quad (13)$$

From equation (11) and (12), A_i and C_i where ($i=1,2,3$) are coefficient vectors updated dynamically to control the balance between exploration and exploitation.

For odd-numbered iterations, the KH algorithm is applied, which updates solutions based on local movements inspired by krill foraging behavior. The update rule is:

$$X^{new} = X + F_{local} \quad (14)$$

The local effect of neighboring krill on agent i is defined as:

$$\alpha_{i,local} = \sum_{j=1}^{N_N} \hat{R}_{i,j} \hat{X}_{i,j} \quad (15)$$

The average distance between krill and their neighbors is calculated by:

$$d_{s,i} = \frac{1}{5N} \sum_{j=1}^N \|X_i - X_j\| \quad (16)$$

This process ensures the selection of highly relevant features that lead to more accurate and robust classification results. The pseudocode of the AKWO algorithm has been moved to table 1 for clarity and to avoid interrupting the flow of the main methodology section.

CLASSIFICATION PHASE

The classification stage transforms the selected feature embeddings into final depression predictions using a neural network composed of flattening, dense layers, dropout, and a sigmoid output. The flattening layer converts multidimensional feature maps into a 1-D vector, which is then processed by fully connected layers to learn complex decision boundaries. A sigmoid activation generates the final probability score, enabling accurate binary classification between depressed and non-depressed states.

Result

This section presents a comprehensive analysis of the performance metrics and results obtained using the proposed model.

IMPLEMENTATION SETUP

Table 1. AKWO algorithm.

Pseudocode 1: Adaptive krill-wolf optimization

- Initialize the population of solutions (each solution is a subset of features)
- Evaluate the initial fitness of the population
- Define parameters for both GWO and KH
- While stopping criterion not met:
 - if iteration % 2 == 0; // Even iterations: Apply GWO
 - Identify α , β , δ wolves based on fitness
 - For each solution in population:
 - Update position based on α , β , δ
 - Else: Apply KH
 - For each krill in the population:
 - Compute local
 - Update krill position based on these movements
 - Evaluate the fitness of the updated solutions
 - Select top-performing solutions to form the next generation
 - if iteration % 2 == 0;
 - update α , β , δ based on new fitness values
 - Increment iteration counter
 - Return the best feature subset found

The experimental setup was conducted on a machine running Windows 10 (64-bit) as the operating system, equipped with an Intel Core i5 processor and 16 GB of RAM to ensure smooth computational performance. The implementation was carried out in the Python 3.10 environment, utilizing the PyTorch deep learning framework for model development and training.

Hyperparameter setting

Table 2 shows the detailed hyperparameter settings assigned for this proposed model.

IMPLEMENTATION OUTPUT

Figure 4 illustrates the effectiveness of the ASP technique in removing background noise from voice signals. The noisy waveform shows heavy amplitude fluctuations, while the denoised version reveals a cleaner signal with clearer speech peaks. This improvement is achieved through spectral subtraction and Kalman filtering, which suppress noise without distorting speech. The spectrogram comparison further confirms that ASP preserves key speech frequencies, enhancing the quality of features used for mental health prediction.

Figure 5 shows the class imbalance in the dataset, with fewer non-depressed samples compared to depressed ones. This imbalance can bias the model toward the majority class. After applying BL-SMOTE, synthetic samples are generated for the minority class, resulting in a balanced distribution. This improves generalization and reduces prediction bias, making the model more robust for medical and psychological classification tasks.

Table 2. Hyperparameter configuration.

Hyperparameters	Values
Number of epochs	150
Learning rate	0.001
Batch size	32
Optimizer	Adam
Dropout rate	0.3
Kernel_size	(3,3)
Stride	(1,1)
Padding	(1,1)
Population size (AKWO)	10
Number of iterations (AKWO)	50
Number of layers	18
Activation functions	ReLU
Loss	Binary cross entropy
Thresholds (for final classification)	0.5

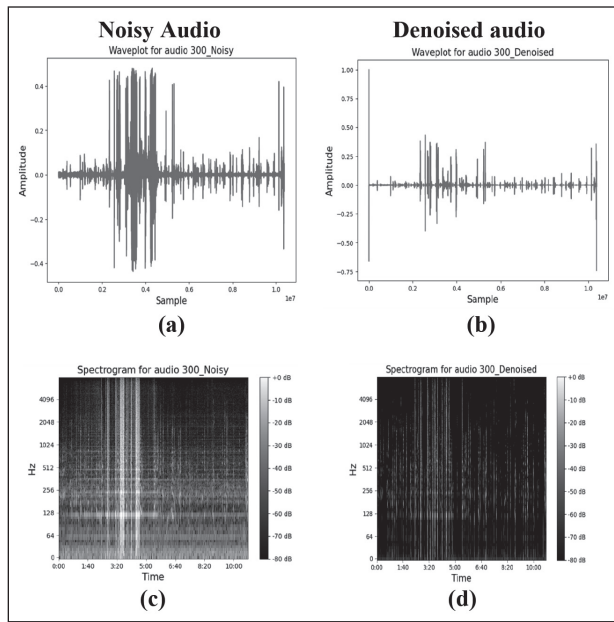


Figure 4. Impact of ASP on audio signal analysis and its spectrogram comparison.

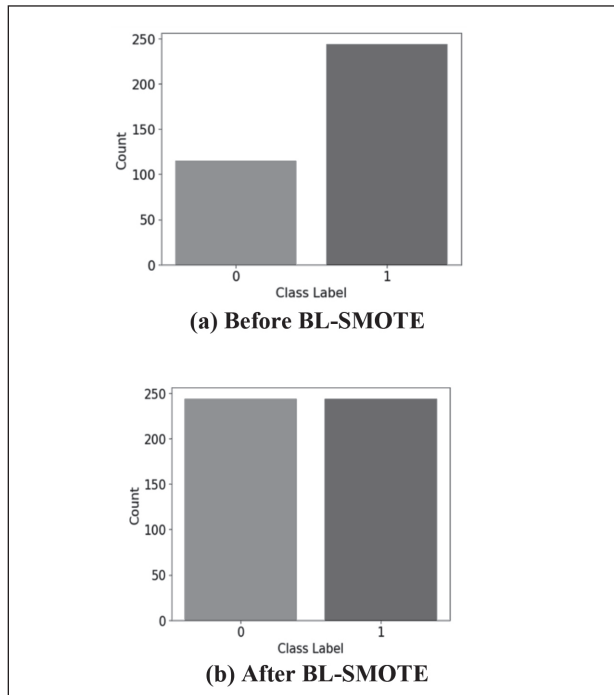


Figure 5. Addressing class imbalance with BL-SMOTE.

PERFORMANCE EVALUATION

The proposed model is evaluated using classification metrics (accuracy, precision, recall, specificity, F1-score), error metrics (MAE, RMSE), and the Kappa coefficient to provide a comprehensive assess-

ment of its effectiveness, reliability, and suitability for voice-based mental health evaluation.

Figure 6 shows that the AKWO algorithm rapidly converges during feature selection, quickly reducing fitness values and stabilizing around 0.021. This demonstrates AKWO’s effective balance of exploration and exploitation, highlighting its robustness in selecting relevant features efficiently for voice-based depression detection.

Figure 7 shows that over 150 training epochs, the proposed model achieves closely aligned training and validation accuracy and loss curves, with rapid early learning and steady improvement. This indicates efficient learning, strong generalization, minimal overfitting, and robust predictive performance for voice-based depression detection.

The confusion matrix shows that the proposed model reliably distinguishes depressed and non-depressed individuals, with 235 true positives, 115 true negatives, 10 false positives, and 0 false negatives, reflecting high accuracy and precision in depression detection (figure 8).

Figure 9 illustrates the ROC curves, showing the model’s exceptional diagnostic capability across all five cross-validation folds. The AUC values range from 0.98 to 1.00, indicating outstanding discriminatory performance. The curves remain close to the top-left corner, reflecting high true positive rates and minimal false positives. This consistent performance across folds confirms the model’s robustness and accuracy in distinguishing between depressed and non-depressed classes. Table 3 further summarizes the overall performance of the proposed model.

5-FOLD CROSS VALIDATION

To ensure the robustness and generalizability of the proposed model, a 5-fold cross-validation stra-

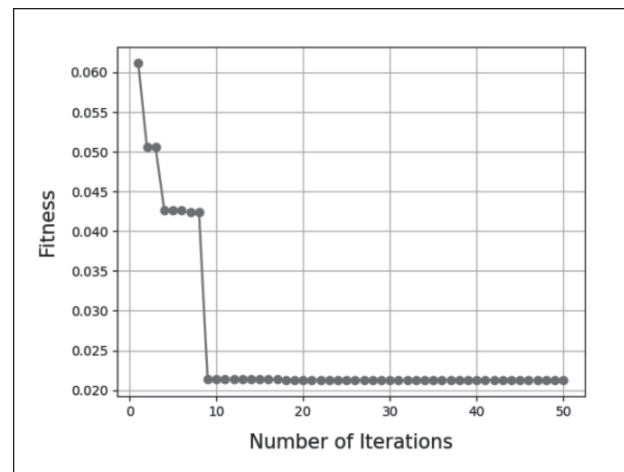


Figure 6. Fitness convergence analysis of AKWO algorithm.

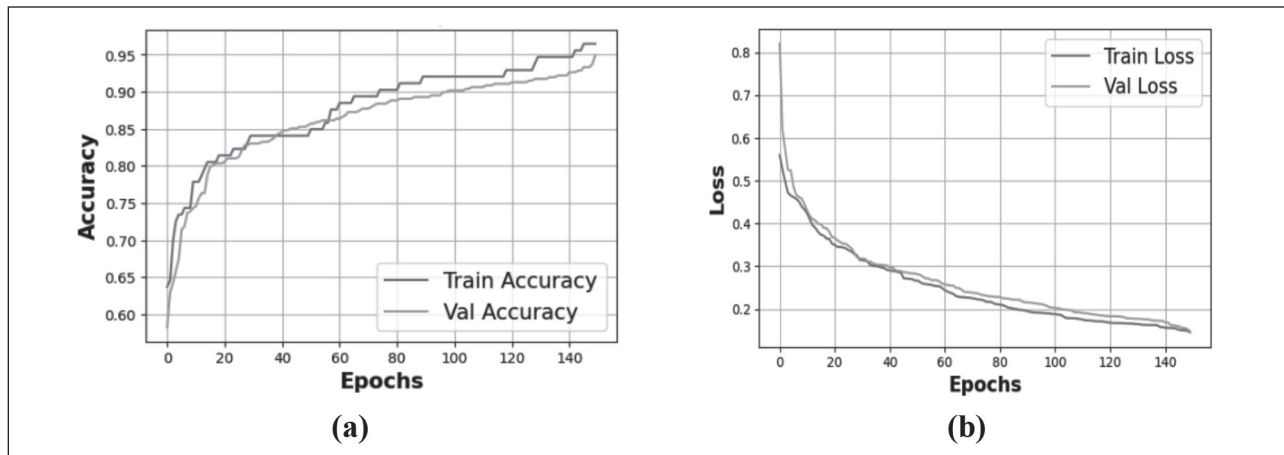


Figure 7. Model learning and overfitting analysis using (a) Accuracy Curves (b) and Loss Curves.

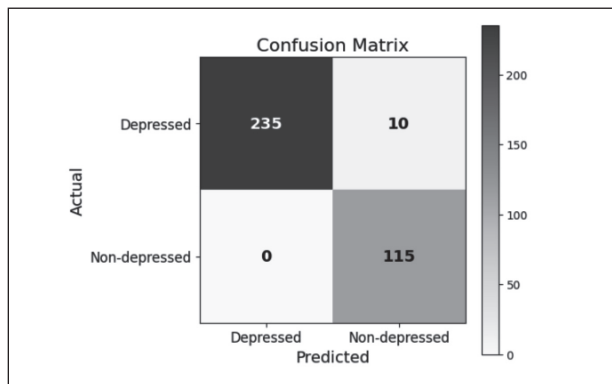


Figure 8. Predictive accuracy evaluation using confusion matrix.

Table 3. Performance evaluation of the proposed Depression Classification Model.

Parameters	Performance (%)
Accuracy	97.22
Precision	100
Recall	95.92
F1-score	97.92
MAE	0.027
RMSE	0.1666
Kappa	93.76

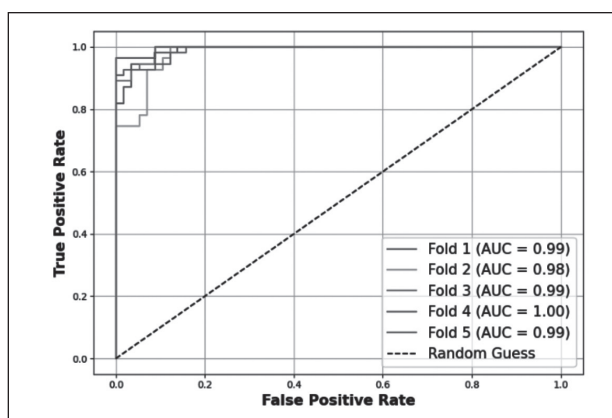


Figure 9. Proposed models robustness evaluation using ROC analysis

tegy was employed, and the performance metrics were averaged across all folds as presented in Table 4.

Table 4 and Figure 10 show that the proposed model performs consistently across 5-fold cross-validation, achieving high accuracy (96.98-

97.62%), strong precision, recall, F1-score, and specificity, low MSE, and substantial Kappa agreement, demonstrating its robust and reliable depression classification.

COMPARATIVE ANALYSIS

To evaluate the effectiveness of the proposed model, a comparative analysis was conducted against existing multimodal and transfer learning approaches, with the results summarized in table 5²⁷.

The proposed model significantly outperforms benchmark multimodal and transfer learning methods on the DAIC-WOZ dataset, achieving higher accuracy, precision, recall, and F1-score (figure 11). Wilcoxon signed-rank tests ($p < 0.05$) confirm that these improvements are statistically significant, demonstrating robust, reliable, and clinically relevant performance for depression detection. Table 6²⁸ shows error analysis; figure 12 shows error analysis comparison with existing models.

Table 4. Model performance analysis using 5-Fold Cross-Validation.

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MSE	RMSE	Kappa (%)
1	97.01	100	95.41	97.65	0.028	0.171	93.22
2	97.44	100	96.33	98.11	0.026	0.162	94.15
3	96.98	100	95.10	97.53	0.029	0.178	92.84
4	97.62	100	96.50	98.22	0.025	0.159	94.62
5	97.05	100	95.26	97.60	0.027	0.163	93.91
Mean	97.22	100	95.92	97.92	0.027	0.167	93.76

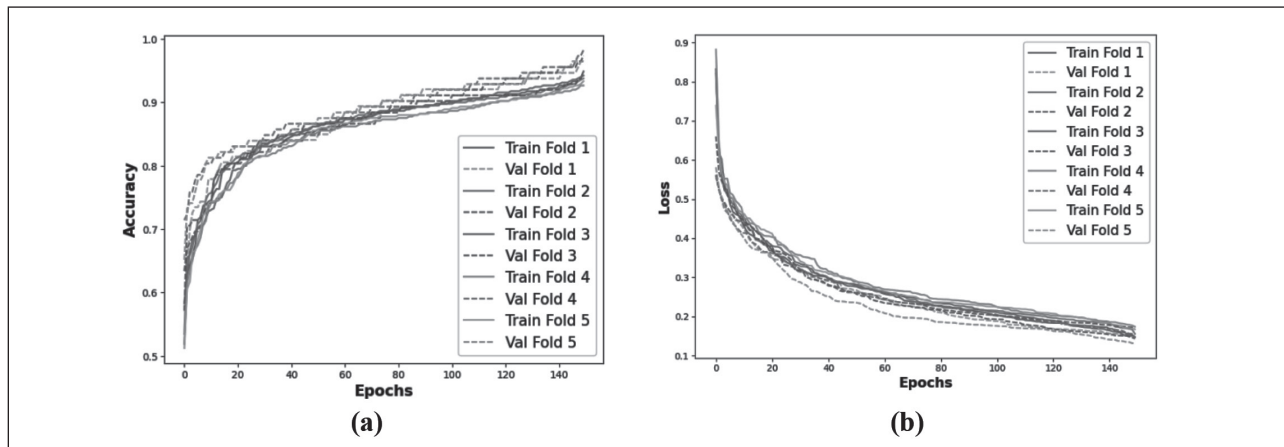


Figure 10. 5-Fold-Cross-Validation analysis of our proposed model.

Table 5. Comparative analysis of multimodal and transfer Learning Models.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
UE ²⁷	80.54	80.96	78.13	79.52
MMIT ²⁷	81.22	82.01	81.22	81.61
MMSDTL ²⁷	87.55	89.26	89.71	89.21
MMSDntTL-MCD ²⁷	93.81	94.82	92.11	93.45
MMSDntTL-SNGP ²⁷	95.07	95.35	92.86	94.09
Proposed	97.22	100	95.92	97.92

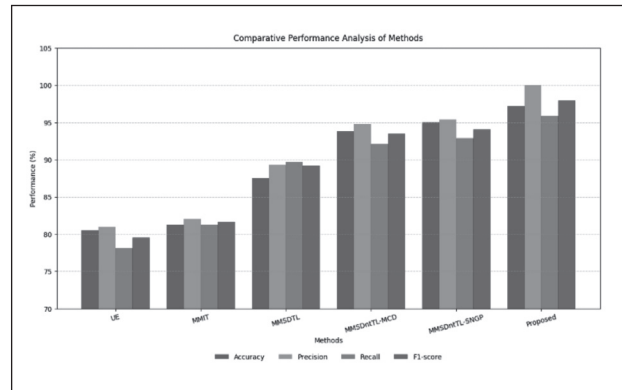


Figure 11. Comparison of our proposed model with other existing models.

Discussion

The proposed voice-based depression detection model achieves strong and consistent performance, with high accuracy, perfect precision, and strong recall across cross-validation, confirming reliable classification. High Kappa values, strong ROC-AUC scores, and well-regulated training and validation curves indicate robust discrimination, good generalization,

Table 6. Error analysis of our proposed model with other existing methods.

Methods	MAE	RMSE
DCNN-DNN28	5.16	5.97
DCNN28	4.63	5.52
TE-CNN28	4.48	5.37
Bi-LSTM-ATT28	3.18	3.68
GCNN28	1.25	2.15
Proposed	0.027	0.1666

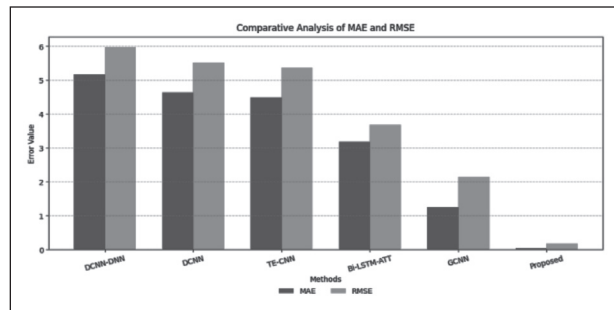


Figure 12. Error analysis comparison with existing models.

and minimal overfitting. Clinically, the model shows promise for early screening and remote mental health monitoring through speech analysis, especially in resource-limited settings, and could be integrated into telehealth systems for continuous monitoring and automated risk alerts. Comparative results demonstrate that the model outperforms existing methods, achieving higher predictive accuracy and lower error rates. However, practical deployment is limited by language variability, possible algorithmic bias, limited interpretability, absence of external validation, and lack of real-world clinical testing. Future work should focus on diverse datasets, clinical trials, and explainable AI techniques to improve reliability, fairness, and real-world applicability, supporting the model's potential role in digital psychiatry and telemedicine.

Conclusions

The proposed voice-based depression detection framework demonstrates strong predictive performance on the DAIC-WOZ dataset, achieving high accuracy and low error rates, which highlights its potential for early mental health screening and digital psychiatry applications. However, these findings are currently limited to a single dataset, and independent external validation is required before clinical deployment. Future work will focus on validating the model on diverse populations, improving interpretability through explainable AI, and optimizing the framework for real-time use within telehealth systems.

Conflict of interests: the authors have no conflict of interests to declare.

References

- Elliott M, Ragsdale JM. Nature and/or nurture: causal attributions of mental illness and stigma. *Soc Psychol Q* 2023; 87: 175-96.
- Pullen E, Ekl EA, Felix E, Turner C, Perry BL, Pescosolido BA. Labeling, causal attributions, and social network ties to people with mental illness. *Soc Sci Med* 2021; 293:114646.
- Kieling C, Buchweitz C, Caye A, et al. Worldwide prevalence and disability from mental disorders across childhood and adolescence. *JAMA Psychiatry* 2024; 81: 347.
- Liu X, Yang F, Huang N, Zhang S, Guo J. Thirty-year trends of anxiety disorders among adolescents based on the 2019 Global Burden of Disease Study. *Gen Psychiatr* 2024; 37: e101288.
- Jo AH, Kwak KC. Diagnosis of depression based on four-stream model of bi-LSTM and CNN from audio and text information. *IEEE Access* 2022; 10: 134113-35.
- Schultebrucks K, Khan Z, Chang J, Chang B. Digital biomarkers for diagnostic assessment of stress pathologies and neurocognitive performance. *Psychoneuroendocrinology* 2024; 160: 106754.
- Rentoumi V, Vassiliou E, Sali D, Demiraj A, Paliouras G. Towards Automatic early detection: assessing LANGaware's language and speech biomarkers in neurocognitive and affective disorders. *Alzheimers Dement* 2024; 20(S2): e088700.
- Jo H, Park C, Lee E, et al. Neural effects of one's own voice on Self-Talk for emotion regulation. *Brain Sci* 2024; 14: 637.
- Rybner A, Jessen ET, Mortensen MD, et al. Vocal markers of autism: assessing the generalizability of machine learning models. *Autism Res* 2022; 15: 1018-30.
- Mai N, Lee B, Chung W. Affective computing on Machine Learning-Based emotion recognition using a Self-Made EEG device. *Sensors* 2021; 21: 5135.
- Zhou Y, Han W, Yao X, Xue J, Li Z, Li Y. Developing a machine learning model for detecting depression, anxiety, and apathy in older adults with mild cognitive impairment using speech and facial expressions: a cross-sectional observational study. *Int J Nurs Stud* 2023; 146: 104562.
- Provost EM, Sperry SH, Tavernor J, Anderau S, Yocum A, McInnis MG. Emotion recognition in the real world: passively collecting and estimating emotions from natural speech data of individuals with bipolar disorder. *IEEE Transactions on Affective Computing* 2025; 16: 28-40.
- Yadav G, Bokhari MU, Alzahrani SI, Alam S, Shuaib M. Emotion-Aware Ensemble Learning (EAEL): revolutionizing mental health diagnosis of corporate professionals via intelligent integration of multi-modal data sources and ensemble techniques. *IEEE Access* 2025; 1.
- Sideri K, Van Dijk N. The techno-politics of computing the mind: opening the black box of digital psychiatry. *Soc Stud Sci* 2025; 55: 382-403.
- Liu Z, Zhao J. Leveraging deep learning for robust EEG analysis in mental health monitoring. *Front Neuroinform* 2025; 18: 1494970.
- Chen Q, Xia M, Li J, et al. MDD-SSTNet: detecting major depressive disorder by exploring spectral-spatial-temporal information on resting-state electroencephalography data based on deep neural network. *Cereb Cortex* 2025; 35: bhae505.
- Zhang F, Yang C, You L, et al. WS-BiLSTM-MA: wavelet scattering-based BiLSTM with mixed attention block for MDD recognition using multi-channel EEG signals. *IEEE Transactions on Instrumentation and Measurement* 2025; 74: 1-13.
- Guo W, He Q, Lin Z, et al. Enhancing depression recognition through a mixed expert model by integrating speaker-related and emotion-related features. *Sci Rep* 2025; 15: 4064.
- Ntalampiras S. Interpretable probabilistic identification of depression in speech. *Sensors* 2025; 25: 270.
- Karamat A, Imran M, Yaseen MU, Bukhsh R, Aslam S, Ashraf N. A hybrid transformer architecture for multi-

- class mental illness prediction using social media text. *IEEE Access* 2025; 13: 12148-67.
21. Daru D, Surani H, Koladia H, Parmar K, Srivastava K. Depression detection using hybrid transformer networks. In: Sharma N, Goje A, Chakrabarti A, Bruckstein AM (eds). *Data Management, Analytics and Innovation. Lecture Notes in Networks and Systems*, vol 662. Singapore: Springer, 2023.
 22. Zhou Y, Yu X, Huang Z, et al. Multi-modal Fused-attention network for depression level recognition based on enhanced audiovisual cues. *IEEE Access* 2025; 13: 37913-23.
 23. Jin N, Ye R, Li P. Diagnosis of depression based on facial multimodal data. *Front Psychiatry* 2025; 16: 1508772.
 24. Fu Y, Ren F, Lin J. Apriori algorithm based prediction of students' mental health risks in the context of artificial intelligence. *Front Public Health* 2025; 13: 1533934.
 25. Shukla A, Tandel BN, Kajaliya PP. Auditory and mental well-being of teachers in urban noise environment: a partial least square structural equation model approach. *Applied Acoustics* 2024; 230: 110417.
 26. Gratch J, Artstein R, Lucas G, et al. The Distress Analysis Interview Corpus of human and computer interviews. In: Calzolari N, Choukri K, Declerck T, et al. (eds). *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014.
 27. Ahmed S, Yousuf MA, Monowar MM, Hamid A, Alassafi MO. Taking all the factors we need: a multimodal depression classification with uncertainty approximation. *IEEE Access* 2023; 11: 99847-61.
 28. Ishimaru M, Okada Y, Uchiyama R, Horiguchi R, Toyoshima I. A new regression model for depression severity prediction based on correlation among audio features using a graph convolutional neural network. *Diagnostics* 2023; 13: 27.