An Adaptive Multi-Scale Feature Fusion Framework for Detecting Depression and Assessing Its Severity from Speech Signals

Raminder Kaur Nagra and Vikram Kulkarni *

Department of Information Technology, Mukesh Patel School of Technology Management & Engineering, Shri Vile Parle Kelvani Mandal's Narsee Monjee Institute of Management Studies, Mumbai, India
Email: raminderkaur.nagra@nmims.edu (R.K.N.); vikram.kulkarni@nmims.edu (V.K.)
*Corresponding author

Abstract*—Depression is a chronic mental health disorder characterized by psychological, physical, and social impairments, often resulting in diminished interest in daily activities and, in severe cases, suicidal ideation. Recent physiological studies have revealed measurable differences in vocal attributes between depressed and non-depressed individuals, motivating the use of speech signal analysis for automatic depression detection. However, conventional approaches relying on Mel-frequency or Fourier-based spectrograms often fail to capture discriminative features for effective emotional classification. To address this limitation, this study proposes an optimized deep learning framework, Adaptive Multi-scale Fused Feature-based Deep Network (AMF2-DNet) for both depression detection and severity assessment. Speech signals from the Distress Analysis Interview Corpus with Wizard-of-Oz (DAIC-WOZ) benchmark dataset undergo noise reduction and normalization, followed by feature extraction, including Mel-Frequency Cepstral Coefficients (MFCCs), spectral features, and deep features learned through a Sparse Autoencoder (SAE). These multimodal features are integrated within a residual Bidirectional Recurrent Neural Network (Bi-RNN) architecture enhanced with a novel loss function for improved convergence. Further, the Refined Attacking Strategy of Giant Armadillo Optimization (RASGAO) dynamically optimizes network parameters and scale weights. Experimental results demonstrate that AMF2-DNet achieves 94.88% accuracy, an F1-Score of 84.50%, and a Mathews Correlation Coefficient (MCC) of 0.82, surpassing state-of-the-art baselines by up to 7.44% in accuracy. The proposed framework also exhibits strong robustness across noise levels and speaker variability. Future work will focus on real-time deployment using lightweight architectures, cross-lingual validation, and multimodal fusion with facial and physiological cues for enhanced clinical applicability.

***Keywords*—feature extraction, speech recognition, speech analysis, recurrent neural network, affective computing**

I. Introduction

Currently, depression has become an inevitable mental disorder that cannot be avoided. The people affected by depression mostly suffer from negative feelings for a long time. Additionally, individuals with depression often find it difficult to lead normal lives and may engage in self-mutilation, as well as experience suicidal thoughts and behaviors [1]. Based on the World Health Organization (WHO) report in 2020, nearly 280 million people (3.8% of the world population) are troubled by depression. In addition, the WHO forecasts that depression may become the second leading cause by 2030 [2]. Therefore, if depression is left uncured, it can negatively impact the quality of life and result in suicide. But, if it is detected properly, depression is a curable disease, and its signs can be relieved [3]. Nowadays, the diagnosis outcomes for depression are mostly obtained via joint consultation with specialized doctors. During the entire diagnosis process, doctors need to expend considerable energy and time; as a consequence, individuals with an affected condition may not be aware of their situation at the correct time, resulting in treatment delays [4]. There is a requirement to implement a technique for reliable diagnosis and screening of depression automatically to support remote validations and highly accurate treatment with personalization [5]. Based on the severity, depression is treated differently; therefore, rapid and appropriate recognition of its severity is significant for planning proper treatment procedures [6].

Nowadays, various experiments have been conducted to forecast the severity of depression by employing deep learning mechanisms in order to assist physicians in diagnosis. These experiments employed modality information, including text data, facial expression data, and audio data, as training details and developed regression approaches on the basis of neural networks for forecasting the severity of the depression [7]. Notably, specific audio features can be detected in the voices of depressed patients; based on these features, various audio-aided approaches for predicting depression severity have been developed. [8]. Moreover, audio features that capture the correlation between different acoustic patterns in depressed speech have been identified as critical for accurate severity prediction.

Hence, the depressed patient's voices can be categorized based on the combinatorial relationships among distinct audio features that are critical factors for

Manuscript received June 25, 2025; revised July 14, 2025; accepted August 28, 2025; published December 18, 2025.

accurately forecasting depression severity. Nevertheless, the conventional studies have treated individual features as independent; hence, have not developed approaches that leverage the relationships between these features. It is well established that, over the past decades, considerable efforts have been made by experts to develop automatic methods for identifying depression that bypass the conventional stages of response collection and manual assessment of participants' mental states [9]. Nevertheless, significant challenges pose a complexity to the progress of depression diagnosis experiments. Initially, limited sample sizes abound in this sector as a result of the relatively high expense of data acquisition, often demands human intervention and poses privacy risks [10]. Although several machine learning approaches can achieve high accuracy even when trained on limited, the same cannot be said for deep learning strategies. Next, certain depressed patients may conceal their symptoms due to social stigma [11]. Some may provide wrong responses regarding their condition, which can lead to incorrect diagnosis, and adversely affect their treatment options [12].

The automatic detection of depression has gained attention with the availability of existing data sources and the capabilities of machine learning approaches for understanding complex relationships [13]. Among speech-aided approaches, conventional experiments have focused on employing handcrafted acoustic features such as cepstral, formant, and prosody features, and subsequently categorizing the patterns using machine learning techniques, including "random forest, logistic regression, and Support Vector Machine (SVM)". These experiments have suggested that acoustic features are strongly associated with depression [14]. However, handcrafted acoustic features require significant time and effort, and because the retrieved features rely on the domain knowledge of experts, several significant data points related to depression may be lost [15]. Identifying reliable acoustic features for detecting depression remains an open experimental challenge. Over the past decades, deep learning has demonstrated significant advancements across nearly all research domains, including application-oriented fields such as healthcare [16]. Despite progress in both modeling techniques, deep learning models for speech-based depression identification still face substantial challenges. Universal acoustic features for validating depression have not yet been established, rendering current methods less effective for detecting depression through speech. Therefore, this study introduces a novel framework for depression and severity detection, supported by advanced deep neural networks.

The developed depression and its severity level identification framework includes the following contributions.

- While the individual components employed in this study—such as Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction, Sparse Autoencoder (SAE) for deep feature learning, Bidirectional Recurrent Neural Network (Bi-RNN) for sequence modeling, and metaheuristic optimization—are well-established, the novelty lies in their task-specific integration and coordinated adaptation for depression and severity level detection from speech. In particular, we introduce an adaptive multi-scale feature fusion framework that combines MFCC, spectral, and deep features using weight optimization via the Refined Attacking Strategy of Giant Armadillo Optimization (RASGAO). Furthermore, a customized loss function is incorporated to enhance the model's robustness to outliers and imbalanced classes. Unlike prior approaches, our framework jointly optimizes feature fusion weights and model parameters, resulting in improved detection performance. These enhancements are validated through comparative analysis, demonstrating the practical impact of our integrated design without overstating its originality.
- To perform the automatic feature extraction for improving the efficiency of the detection mechanism. The speech signals are initially pre-processed by reducing the noise and performing normalization. This operation enhances the clarity of the signal and allows better feature extraction. Further, from these pre-processed signals, the MFCC, spectral, and deep features are retrieved. The features obtained from the MFCC attain the capacity to capture the speech's perceptual qualities, making it robust to channel and noise distortions. The retrieved spectral features help to capture the comprehensive details across the spectrum and also support recognizing the subtle changes in the data. Finally, the deep features obtained from the SAE enhance the robustness to noise and overfitting and also retrieve the meaningful features.
- To develop a RASGAO for the developed detection model's performance. The developed RASGAO properly balances the exploitation phase with exploration by refining the attacking mechanism (exploration) of conventional GAO. Thus, it achieves relatively better convergence rates and also reduces the computational complexities. In the developed AMF2-DNet, the RASGAO is incorporated for optimally selecting the network weights and learning rate, thus enhancing the Mathews Correlation Coefficient (MCC) and precision values of the developed depression and its severity identification process.
- To design an AMF2-DNet for identifying the depression and its severity level. This developed AMF2-DNet utilizes the extracted MFCC, spectral, and deep features as input, and these three features are passed to the multi-scale task. Further, these features are fused by integrating the details from distinct feature sets, thus eliminating the redundant data and improving the robustness of the approach. Moreover, the fused feature is subjected to the deep network, where the residual Bi-RNN model is utilized. The residual Bi-RNN obtains the

complex and related features by improving the gradient flow, thus providing highly promising and timely depression and its severity level detection outcomes. Here, the weights supported in the feature fusion and the parameters of Bi-RNN are optimized by RASGAO. Finally, for minimizing the error rates, the novel loss function is included, thus the developed AMF2-DNet model produces highly accurate detection results.

The proposed framework introduces a well-coordinated integration of MFCC, spectral, and SAE-derived deep features into a residual Bi-RNN architecture, optimized using a refined GAO-based metaheuristic (RASGAO). While each component is well established in the literature, the novelty of our approach lies in the adaptive multi-scale fusion mechanism, that enables effective combination of heterogeneous features, and in the use of RASGAO for dynamic optimization of both feature-level weights and network hyperparameters, addressing convergence and generalization challenges specific to depression detection from speech. Furthermore, the customized loss function enhances robustness to outliers and imbalanced class distributions, as demonstrated through comparative evaluations against standard loss functions. The method achieves superior performance in both detection and severity prediction tasks on the Distress Analysis Interview Corpus with Wizard-of-Oz (DAIC-WOZ) dataset, offering a meaningful advancement over prior work.

The structure of the developed depression and its severity level identification framework is outlined as follows. Section II reviews existing work on depression identification. Section III presents the structural description of the implemented depression and severity detection framework with speech signal collection. Section IV details the noise removal process and prominent feature extraction from speech signals for the developed DeepDepressionNet. Section V describes the implementation of an adaptive multi-scale fused feature-based deep network for effective depression and severity detection. Section VI reports the performance evaluation of the proposed framework, and Section VII provides the conclusion.

II. Existing Works

A. Foundational Background

The early research on depression detection from speech and Speech Emotion Recognition (SER) primarily focused on handcrafted acoustic features such as MFCCs, pitch, formants, and prosodic attributes, combined with machine learning classifiers like SVM and Random Forest. Eyben *et al.* [17] introduced the widely adopted open-source Speech and Music Interpretation by Large-space Extraction (openSMILE) toolkit, enabling standardized extraction of acoustic features for emotion and affective state analysis. Furthermore, the Audio/Visual Emotion Challenge (AVEC) series [18], initiated in 2011, provided benchmark datasets and baseline methods for affective computing and depression detection, influencing subsequent advancements. Although these approaches were effective in capturing basic acoustic cues, they required extensive manual feature engineering and were sensitive to noise and speaker variability, limiting their robustness in real-world scenarios.

B. Related Works

Zhou *et al.* [19] proposed a model that is a fusion of Self-Attention Networks (SAN) and Deep Convolutional Neural Networks (DCNN) to classify individuals who are depressed or not based on their speech. SAN uses long-term connections of acoustic features, and DCNN assumes that the spatial interrelationships in spectrograms are used. When joined and pooled together, these features enhance the accuracy of detecting depression, and it has been confirmed through the experimentation conducted on AVEC datasets, which support the use of SAN as an alternative to recurrent models, and hold potential in the future applications in speech analysis.

Seneviratne and Espy-Wilson [20] presented a multi-stage dilated Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model for classifying depression severity from speech into normal, moderate, and severe levels. It used Articulatory Coordination Features (ACFs) derived directly from vocal tract variables, capturing neuromotor slowing effects. The CNN processed short speech segments to extract embeddings, which an LSTM then aggregated for session-level classification. This approach outperformed traditional features like MFCCs, achieving a 27.47% relative improvement in unweighted average recall. The model improved granularity and robustness in depression severity assessment, addressing data scarcity and overfitting in clinical speech analysis.

Niu *et al.* [21] explored a WavDepressionNet for designing original speech signals to improve the accuracy of the prediction. In this mechanism, a representation module was suggested for discovering a group of basis vectors for building the optimal transformation region. The assessment module was developed to showcase the useful factors. The research outcomes obtained from the different depression data sources illustrated the efficacy of this model over conventional works.

Iyortsuun *et al.* [22] have concentrated on discerning the telltale signs from the patient statements. Both text and audio data were leveraged, with an attention mechanism to learn the significant weights. This designed model showed its potential in the identification of depression while leveraging the speech and textual modalities without employing the preset queries for efficient identification.

Kim *et al.* [23] designed a model for recognizing depression automatically employing the large acoustic features on the basis of the Korean language. The smartphones were supported for recording the voices of smartphones while conducting the operation of reading predefined text-aided sentences. Distinct models were estimated and contrasted to identify the depression. The CNN model offered high accuracy.

Ravi *et al.* [24] designed different disentanglement approaches for retrieving the speaker-identification data from the depression-based details. The experiments, which included distinct model architectures and input features,

have shown enhanced F1-Score values. By minimizing the reliance on speaker-identity-based features, this model provided an effective path for speech-aided depression identification that ensured the privacy of the patient.

Rejaibi *et al.* [25] suggested a deep RNN model for identifying depression and for forecasting its level of severity from speech. High and low-level audio features were retrieved from the audio recordings. The functionality of the model was estimated under multi-feature and multi-modal experiments. The suggested model supported real-time applications.

Yang *et al.* [26] developed a joint learning model for classifying depression. This developed model generated biologically significant acoustic features by utilizing the time-domain filters. The developed model gathered a new data source for supporting the research in the evaluation of depression. This approach showed a promising solution for recognizing depression.

Ishimaru *et al.* [27] designed a new deep learning-aided regression approach that enabled the forecasting of depression severity according to the correlation between the audio features. The suggested approach was implemented employing the graph CNN. This approach trained the voice features employing the graph-structured data created to express the correlation between the audio features. The prediction research was conducted on depression severity employing standard data sources in some existing works. From the outcomes, the suggested approach could offer a promising solution for diagnosing depression.

Lau *et al.* [28] presented an approach for resolving the data scarcity issue for depression validation. It employed pre-trained large models. This mechanism was developed by employing a tiny data source of tunable attributes. The approach provided optimal functionality.

TABLE I. FEATURES AND CHALLENGES OF EXISTING DEPRESSION DETECTION MODELS

Author	Methodology	Features	Challenges	Key Findings/Results	Identified Gaps
Zhou *et al.* [19]	Hybrid Network	• Low-level acoustic descriptors (eGeMAPS) for Self-Attention Networks (SAN) • 3D log-Mel spectrograms for Deep Convolutional Neural Networks (DCNN)	• Extracting robust, depression-salient features from speech • Capturing local and long-term dependencies • Efficient pooling strategies and feature fusion	Outperforms prior methods (including RNN/CNN hybrids and multimodal baselines) on Audio/Visual Emotion Challenge (AVEC) 2013/2014.	May overlook high-level semantic and linguistic information; lacks cross-corpus generalization analysis.
Seneviratne and Espy-Wilson [20]	Multi-stage Dilated CNN-LSTM	• Articulatory Coordination Features (ACFs) derived from vocal Tract Variables (TVs) • Compared with Mel-Frequency Cepstral Coefficients (MFCCs), Formants, and eGeMAPS	• Scarcity of labeled data for deep models • Sensitivity to speaker variation • Segment-wise predictions may vary in confidence	ACFs improve depression severity, classification accuracy significantly. Session-level predictions enhance segment-level classifier strengths.	Lacks integration of linguistic/textual features; limited to acoustic features only
Niu *et al.* [21]	WavDepressionNet	• Captures depression cues by modeling variations among individuals with different depression levels. • Improves prediction accuracy.	• High training time and computational cost. • Suffers from the dimensionality curse.	Demonstrated improved depression prediction accuracy using raw speech signals.	Lacks feature fusion and optimization strategies for handling high-dimensional data.
Iyortsuun *et al.* [22]	BiLSTM	• Handles variable-length speech segments effectively. • Ignores artifacts and noise for robustness.	• Requires extensive hyperparameter tuning • Struggles with heterogeneous voice data.	Enhanced robustness against artifacts and variable speech lengths.	Does not integrate multi-scale features or optimization techniques.
Kim *et al.* [23]	Convolutional Neural Networks (CNN)	• Processes voice data rapidly for real-time applications. • Automatically extracts features, reducing manual effort.	• Sensitive to data quality. • High computational cost.	Achieved high accuracy with smartphone-based speech recordings.	The model is language-specific and lacks cross-lingual adaptability.
Ravi *et al.* [24]	LSTM	• Detects early depression signs • Captures nonlinear and complex speech patterns.	• Computationally expensive. • Poor generalization on unseen data.	Improved detection accuracy by disentangling speaker identity.	Does not address severity prediction or integrate multimodal features.
Rejaibi *et al.* [25]	Recurrent Neural Network (RNN)	• Handles sequential data and temporal patterns effectively • Suitable for large-scale datasets.	• Suffers from the exploding gradient problem. • High computational complexity.	Provides accurate severity assessment under multi-feature conditions.	Limited optimization for convergence and generalization.
Yang *et al.* [26]	Joint Learning Framework	• Handles missing data effectively • Scalable and flexible for new tasks.	• Risk of overfitting. • High implementation cost for large datasets.	Generated biologically significant acoustic features.	Does not incorporate multi-modal signals such as facial or physiological data.
Ishimaru *et al.* [27]	Graph Convolutional Neural Network (GCNN)	• Provides interpretable solutions. • Captures complex relationships among audio features.	• Requires strong prior domain knowledge • Suffers from over-smoothing issues.	Offers correlation-based feature interpretation for severity prediction.	High complexity limits real-time applicability.
Lau *et al.* [28]	BiLSTM	• Captures temporal dependencies effectively. • Recognizes depression-related features accurately.	• Requires large, labeled datasets. • Model decisions are hard to interpret.	Demonstrated high accuracy through parameter-efficient tuning.	Limited scalability; lacks a lightweight design for real-time deployment.

C. Research Gaps and Challenges

Depression is a serious medical disorder that highly impacts an individual's behaviors, thoughts, and emotions, and in severe cases, can lead to suicide. Timely recognition of depression is therefore critical. When detected accurately, depression is treatable, and its symptoms can be alleviated. However, accurate detection of depression is challenging due to the clinical and biological heterogeneity of the disorder. In addition, timely access to experienced clinicians can be difficult, and the diagnostic process itself is complex. Consequently, numerous automatic depression identification frameworks have been proposed in recent years. Nevertheless, existing depression detection models continue to face the following challenges.

- Most depression detection models utilize video and text as input. Though these modalities are superior, speech has the benefit of being remotely collected, inexpensively, non-invasively, and non-intrusively, which makes it an effective modality to identify depression. Hence, using voice data for depression detection is important in order to attain maximum results.
- Most traditional depression detection approaches have not considered severity prediction, which can limit the effectiveness of treatment outcomes. Therefore, incorporating severity prediction into depression detection is important.
- Traditional models undertake the feature extraction step before detecting depression. Nevertheless, a few methods are based on manual analysis, whereas others ignore the features of interest when dealing with large datasets. Consequently, following a correct and automated feature extraction process is imperative.
- Current deep learning-supported depression detection models tend to have difficulties identifying new and unknown patterns in the input, which may degrade detection performance. Hence, suggesting a deep model with a new loss function is significant in distinguishing such patterns efficiently.
- One of the major issues of traditional deep models is fine-tuning. Despite the many abilities of deep models to handle data well, training can be unstable due to the intensive use of parameters of the network. Thus, it is necessary to use an efficient algorithmic fine-tuning technique.

Hence, a new mechanism for depression and severity detection is implemented with standard models. Table I shows the traditional depression identification model's advantages and issues.

III. Structural Description of the Implemented Depression and Severity Detection Framework with Speech Signal Collection

A. Overview of Implemented DeepDepressionNet

In contemporary times, life in a fast-moving world puts a lot of pressure on people, which can cause mental illness. The symptoms run from chronic low mood and loss of interest to sleeping and eating disorders. Depression can cause suicide in serious cases. Thus, depression has become an extremely common disorder, badly influencing the quality as well as overall health of human life. Proper diagnosis of depression is thus vital. Unfortunately, since objective diagnostic guidelines have not been developed, diagnosis usually depends on the doctor's personal judgment. Depression is thus commonly missed or diagnosed belatedly. Treatment is varied according to the severity of the disorder, so quick assessment of severity is important for planning the right intervention. In recent years, many attempts have been made to predict depression severity with the help of deep learning algorithms. However, traditional models tend to lack good accuracy and are unable to reproduce some of the most important characteristics. To overcome the above shortcomings, this research puts forward an efficient framework to detect depression and its severity. A diagrammatic representation of the suggested framework is shown in Fig. 1, which makes use of deep learning and speech signals.

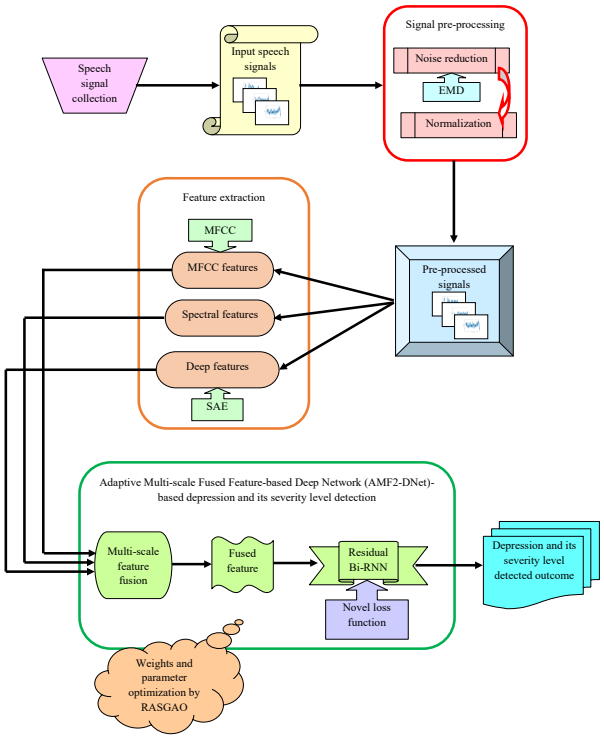


Fig. 1. Diagrammatic illustration of developed depression and its severity level identification with the help of speech signals and deep learning.

This developed work introduces a promising depression and severity identification mechanism by utilizing speech signals. Initially, the necessary speech signals for implementation are obtained from the available dataset. Further, from these speech signals, the unwanted noise and outliers are eliminated with the support of noise reduction and normalization approaches for minimizing the training time and improving the signal quality. Next, from the pre-processed signals, prominent features such as MFCC, spectral, and deep features are extracted. For retrieving the deep features, the proposed framework utilizes the SAE technique. This allows the required features to be obtained, making it possible for the model to concentrate on

depression-related features. After that, these features are given to the AMF2-DNet technique. Here, the deep network is considered as residual Bi-RNN with a novel loss for identifying the depression and its severity levels. Moreover, in this detection phase, a new algorithm named RASGAO is introduced for optimizing the parameters and weights of AMF2-DNet, which enhances the network's efficacy. Lastly, the performance estimations of the introduced depression identification mechanism are done with several evaluation measures. The outcomes are further contrasted with other techniques for guaranteeing the implemented mechanism's strategy.

B. Speech Signal Collection

The designed depression and its severity identification mechanism utilizes the speech signals, and these signals are gathered from the following dataset.

DAIC-WOZ dataset [29]: It includes speech signals and is accessed utilizing "https://dcapswoz.ict.usc.edu/", on 2025–02–18. It contains the clinical interviews developed to support the psychological distress conditions, including post-traumatic stress disorder, depression, and anxiety. The data gathered included video and audio recordings, and more questionnaire responses. Moreover, this data source contains 189 sessions of communication ranging from 7 to 33 mins. In each session, there are interaction transcripts, member audio files, and details regarding facial features.

By adding Noise, data augmentation has been performed in this dataset to prevent the data imbalance problem. The augmented speech signals are further given as input to the designed depression and its severity identification mechanism.

Thus, from this resource, the augmented speech signals are acquired, and these speech signals are indicated as D_s, here $s = 1,2,3, \ldots, S$. The total count of the speech signals is considered as S. Fig. 2 shows some of the sample speech signals acquired from this dataset where x-axis is time (seconds) and y-axis shows amplitude (dB).

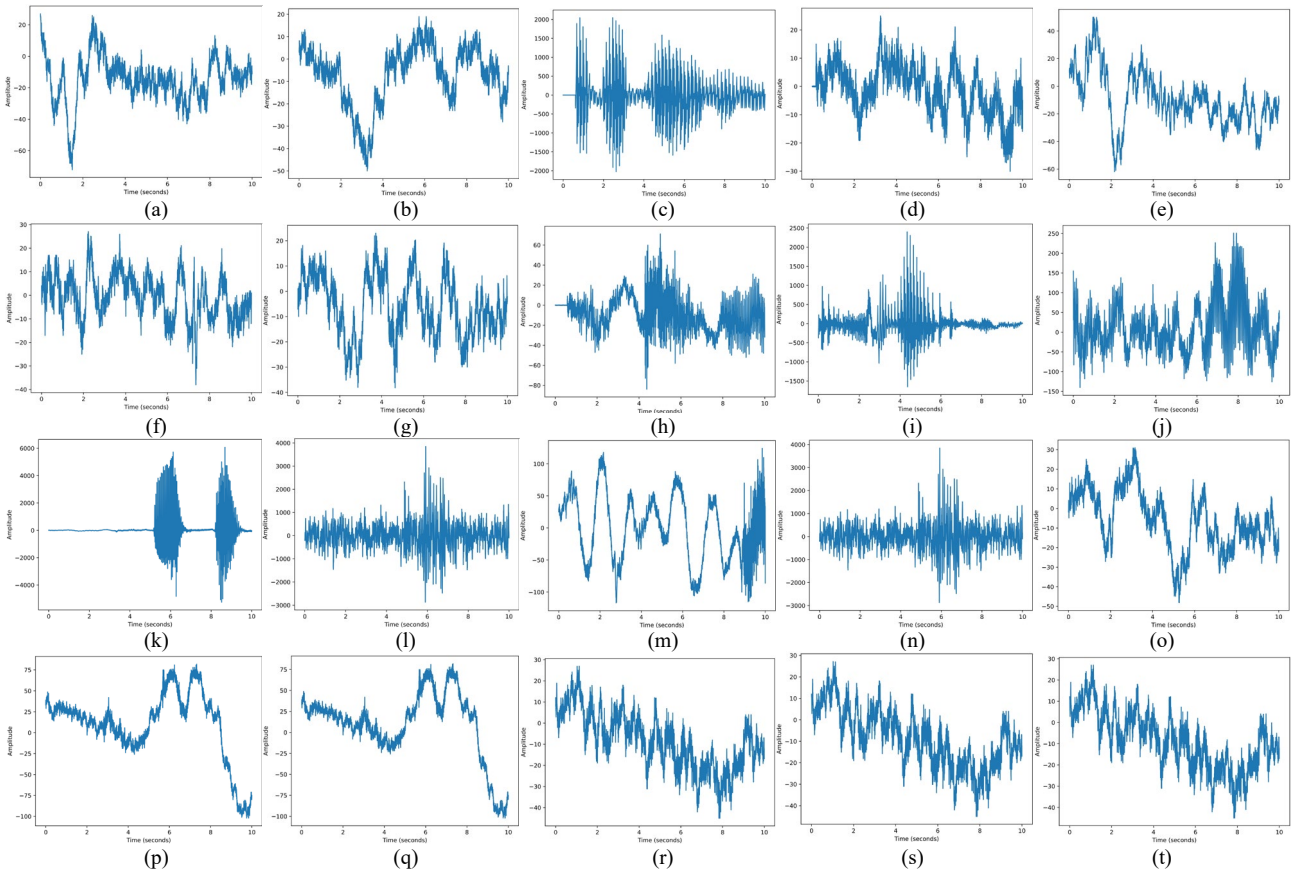


Fig. 2. Sample speech signals collected for the designed depression and its severity identification mechanism. (a)–(e): Normal-Signal 1–5; (f)–(j): Mild-Signal 1–5; (k)–(o): Moderate-Signal 1–5; (p)–(t): Severe-Signal 1–5.

IV. Removing Noise and Extracting Prominent Features from Speech Signals for Developed DeepDepressionNet

A. Signal Pre-processing

The garnered speech signals D_s involved with the signal pre-processing by employing two operations, which are explained below.

Noise reduction: For performing noise reduction, the model employs Empirical Mode Decomposition (EMD) [30], which is an innovative signal processing approach. The EMD process applied to the gathered speech signal $D_s(t)$ is given below.

(1) At first, the entire constrained high and low peaks of the speech signal are identified.

(2) The 3D spline interpolation is utilized for connecting entire local maxima, thereby forming the upper envelope of the signal.
(3) The same operation is conducted for the local minima to produce the lower envelope.
(4) The average a_1 of the upper and lower envelope is estimated, and the difference between a_1 and the signal $D_s(t)$ is computed and specified as $g_1(t)$.
(5) If the variable $g_1(t)$ meets the conditions of the Intrinsic Mode Function (IMF), then the variable g_1 is the initial amplitude and the frequency of the handled oscillatory mechanism $D_s(t)$.
(6) If the variable g_1 doesn't meet the IMF condition, then the operation of shifting, labelled again in phases 1 to 3, is performed on g_1.
(7) Consider that after j process cycles, the variable $g_1 j$ becomes an IMF; thus, it is specified as $k1 = g_1 j$, that is the initial IMF factor from the data.
(8) Eliminating $k1$ from $D_s(t)$, the variable $R1$ is obtained, which is referred to as the raw data from an upcoming cycle.
(9) Reprocess the overhead operation over multiple epochs, and the number of IMFs is obtained with the final residue R_n. This decomposition operation can be stopped when the residue R_n converts into a monotonic function, from which it is impossible to retrieve any more IMF. A common factor for terminating is to have the value of Normalised Standard Difference (NDS) inside an already estimated threshold, as given in Eq. (1).

$$NDS = \sum_{l=1}^{R} \frac{g_{j-1}(t) - g_j(t)^2}{g_j^2(t)} \quad (1)$$

Thus, by utilizing the EMD, the noise-reduced signal is achieved and is specified as D_s^{EMD}.

Normalization: The noise-reduced signal D_s^{EMD} is further given as input to the normalization process. Normalization is a data pre-processing method, where the signal is scaled to fit within a particular range, typically between 0 and 1 to guarantee that all features in the signal equally contribute and to eliminate the significant biases caused by differences in data scales. As a result, it enhances the functionality of the subsequent operations. Before performing feature extraction, it is necessary to guarantee that all features are equally treated. Moreover, by properly scaling the data, the normalization process minimizes the impacts of outliers.

Hence, the signals are normalized, and the pre-processed signals are achieved. The pre-processed signals are specified as D_s^{pre} which are shown in Fig. 3 considering amplitude (dB) on y-axis and time (seconds) on x-axis.

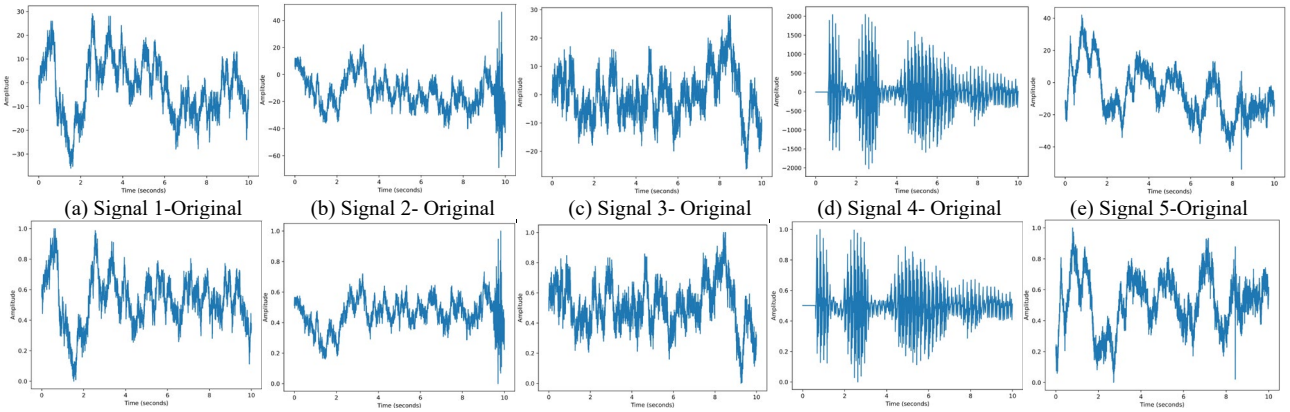


Fig. 3. Audio signals for the implemented depression and its severity identification mechanism (a)–(e): Original-Signal 1–5; (f)–(j): Pre-processed-Signal 1–5.

B. Feature Extraction

The pre-processed signals D_s^{pre} are passed as input to the feature extraction module, where the MFCC, spectral, and deep features are retrieved.

MFCC Feature extraction: From the pre-processed speech signals D_s^{pre}, the MFCC features are retrieved. The MFCC [31] is a short-term power spectrum specification of the speech signal. It utilizes the log power spectrum's sequential cosine transform on a non-linear frequency scale. Eq. (2) specifies the transmission from ordinary frequency r to Mel frequency p.

$$p = 2595\, log_{10}\big((r/700) + 1\big) \quad (2)$$

The following steps are considered for achieving the MFCC features from pre-processed speech signals.

- Initially, for an input pre-processed speech signal, Fourier transform is performed.
- The above spectrum's power is estimated within the triangular overlapping windows placed based on the Mel scale.
- The power logs are taken at each Mel-frequency.
- The Discrete Cosine Transform (DCT) is estimated for the Mel log power list.
- The resulting spectrum's amplitudes are subjected to MFCC.

The extracted MFCC features are specified as D_s^{MFCC}.

Spectral feature extraction: From the pre-processed speech signals D_s^{pre}, spectral features such as "spectral centroid, spectral entropy, spectral flux, and spectral roll-off" are extracted. The extraction of spectral features

involves validating the speech signal's frequency components by converting it into the frequency domain. Further, it extracts the particular features from the resulting spectrum across distinct frequency bands for specifying the significant speech acoustic properties. By efficiently capturing the speech's frequency content, this process allows for robust depression recognition and offers a detailed representation of the signal. The extracted spectral features are indicated as D_s^{Sp}.

Deep feature extraction: The developed work leverages the SAE for obtaining the deep features. The pre-processed speech signals are given to SAE for retrieving the deep features. By presenting the sparse penalty attribute, the SAE [32] employs the method of sparse coding. It helps to retrieve the highly precise deep features by understanding the sparsely limited conditions. Most of the time, the SAE makes the hidden neurons inactive. Considering the variable $g_b(u)$ that specifies the activated b^{th} module of the hidden layer. During the forward propagation operation, the hidden layer b^{th} module's average activation is estimated in Eq. (3) for the P training samples.

$$\gamma_b = \frac{1}{P}\sum_{i=1}^{P}[g_b(u_i)] \tag{3}$$

The amount of average activation γ_b should approximate the constant γ zero since multiple neurons should be inactive. The variable γ is pointed out as a sparse parameter. Thus, from SAE, the deep features are retrieved and the deep features are specified as D_s^{SAE}.

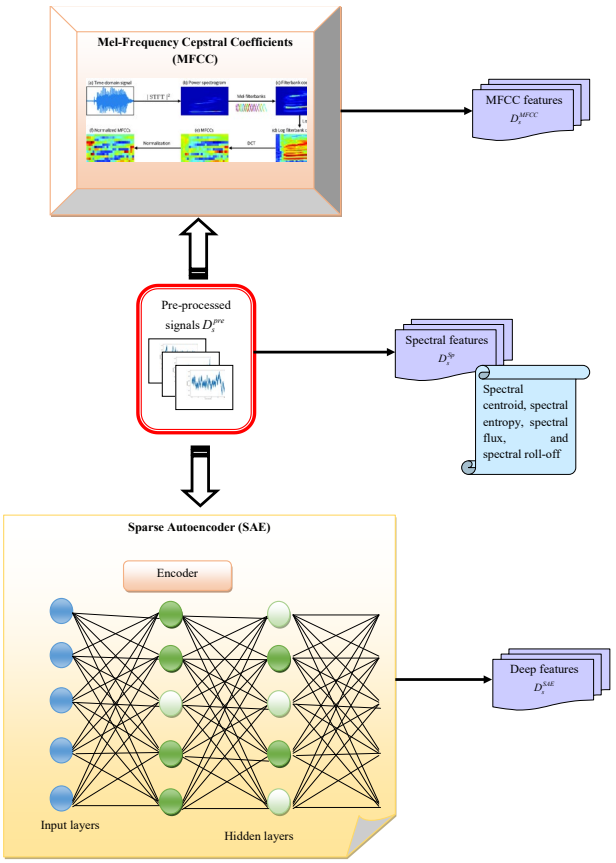


Fig. 4. Diagrammatic view of MFCC, spectral, and deep feature extraction.

Thus, from the pre-processed speech signals, the MFCC, spectral, and deep features are extracted for enhancing the efficacy and accuracy of the developed depression and its severity detection operation. Fig. 4 represents the feature extraction operation's diagrammatic view.

V. Implementation of an Effective Depression and Severity Detection Approach using Adaptive Multi-Scale Fused Feature-Based Deep Network

A. Bi-RNN

The Bi-RNN [33] technique is a relatively efficient model. The RNN offers a highly elegant way of designing sequential healthcare data. Nevertheless, the RNN cannot accurately forecast the data when the sequence length starts to become large. For rectifying this problem, the Bi-RNN is constructed. This technique can train the entire input details from two directions to boost the functionality of the prediction. The backward RNN and forward RNN are included in the Bi-RNN. The input data sequences are read by the forward RNN $\vec{f}$ from p_1 to p_B. Then, it determines the forward hidden state's sequence $(\overrightarrow{y_1}, \ldots\ldots, \overrightarrow{y_B})$.

In reverse order, the backward RNN $\overleftarrow{f}$ understands the input sequence from p_B to p_1. It determines the backward hidden state's sequence $(\overleftarrow{y_1}, \ldots\ldots, \overleftarrow{y_B})$. The backward $\overleftarrow{y_k}$ and forward $\overrightarrow{y_k}$ hidden states are integrated. Thus, the final latent vector representation is obtained as $y_k = [\overrightarrow{y_k}; \overleftarrow{y_k}]^T$.

B. Refined Attacking Strategy of GAO

This research work introduced a new bio-inspired heuristic algorithm named RASGAO.

The developed work designed an AMF2-DNet technique for identifying depression and its severity level. Here, the retrieved three MFCC, spectral, and deep features are given as input. With these three features, the network weights are multiplied. To minimize the computational overhead, the network weights are optimally selected. Moreover, the Bi-RNN technique's learning rate in the AMF2-DNet is also selected optimally. For this purpose, the RASGAO is developed. The RASGAO helps to optimally select the weights and parameters of AMF2-DNet, thus helping to increase the MCC and precision rates of depression and severity recognition process.

The RASGAO is developed from the traditional GAO. The GAO [34] copies the giant armadillo hunting mechanism. The conventional GAO has a high rate of convergence and resolves engineering design problems. Moreover, the existing GAO also mitigates the risk of local optima. Thus, the GAO is considered here.

The GAO helps to rectify the optimization issues with high convergence. However, the traditional GAO can't properly balance the exploration phase with the exploitation phase. The exploration stage is the attacking strategy of the giant armadillo. In the attacking mechanism, the GAO presents a random factor, which hinders the proper balancing of the algorithm when facing dynamic problems. For addressing these issues, the GAO's

attacking strategy is refined; thus, the RASGAO is designed. The developed RASGAO replaces the random integer with a fitness-based derivation, thus helping to balance the exploration phase with the exploitation stage. Thus, the RASGAO becomes effective for the optimization issues even if it is complex. The refined attacking strategy of GAO is formulated in Eq. (4).

$$H_{d,l}^{New} = h_{d,l} + \left(1 + 2 \times \frac{Cfit}{Cfit+Wfit}\right) \times \left(st_{d,l} - U_{d,l} \times h_{d,l}\right) \quad (4)$$

Here, the d^{th} new position of the giant armadillo's attacking mechanism is given as $H_{d,l}^{New}$ with l^{th} dimension. The d^{th} giant armadillo's attacking mechanism's old position is taken as $h_{d,l}$ for the l^{th} dimension. The current fitness is indicated as $Cfit$, and the giant armadillo's chosen termite mound is given as $st_{d,l}$. The worst fitness is specified as $Wfit$ and $U_{d,l}$ is the random attribute which is either 1 or 2.

Thus, the RASGAO is implemented in this study, which chooses the weights and parameters of the developed AMF2-DNet optimally for achieving highly satisfactory results. The RASGAO's pseudo-code is depicted in Algorithm 1.

Algorithm 1: Developed RASGAO

Input: Weights of the multi-scale network and learning rate of Bi-RNN
Output: Optimized weights of the multi-scale network and optimized learning rate of Bi-RNN
Set issue details: constraints, fitness function, and variables
Input population size P and maximum iterations M_i
Calculate fitness function
For $m = 1$ to M_i
  For $d = 1$ to P
    Refined Attacking Strategy
    Estimate the set of termite mounds for d^{th} candidate
    Choose the termite mounds for d^{th} candidate
    Estimate new place of d^{th} candidate by Eq. (4)
    Upgrade d^{th} candidate
    Digging phase
    Estimate new place of d^{th} member
    Upgrade d^{th} candidate
  End for
  Save the best solution
End for
Return optimal solution

C. Developed AMF2-DNet for Detection

Depression and its severity identification operation include a complex interplay of distinct factors such as speech patterns, emotional cues, and physiological signals. The conventional approaches depended on hand-crafted features and shallow machine learning approaches that have limited accuracy. However, with the emergence of deep learning models, it has become possible to implement highly effective and sophisticated depression identification models.

Hence, this research work develops an AMF2-DNet for identifying depression and its severity level. The developed AMF2-DNet utilizes the extracted MFCC D_s^{MFCC}, spectral D_s^{Sp}, and deep features D_s^{SAE} as input. This designed AMF2-DNet network includes some powerful networks and operations for identifying depression and its severity level. The AMF2-DNet contains multi-scale feature fusion, a deep network, optimization process, and novel loss function. Here, the deep network is considered as residual Bi-RNN.

At first, the extracted MFCC D_s^{MFCC}, spectral D_s^{Sp}, and deep features D_s^{SAE} are subjected to the multi-scale feature fusion phase [35]. The multi-scale feature fusion defines the operation of integrating the features extracted from the pre-processed speech signals for producing a highly robust and comprehensive representation of the data. In each scale, the obtained features were passed through the multi-scale operation. Further, these features are fused. The fusion of features from distinct scales enables the capture of the complex patterns in the data that may not be apparent to an individual scale. The multi-scale feature fusion operation provides some merits, including its ability to capture vast amounts of spectral and temporal features of speech signals. By integrating features from distinct scales, the network can learn to represent complex relationships and patterns in the data, resulting in improved functionality in the depression identification process. Moreover, the multi-scale feature fusion enables the incorporation of distinct kinds of features, such as "MFCC, spectral, and deep features" that can offer complementary detail and enhance the robustness of the approach. In addition, the multi-scale mechanism can help to minimize the impact of outliers and noise in the data, as features from distinct scales can offer complementary details.

The network weights are multiplied by each scale feature. The multiplication of multi-scale features with weights is a significant operation in the feature fusion operation, enabling the assignment of importance to each feature on the basis of its relevance to the task. The weighted feature fusion mechanism allows the network to selectively suppress or emphasize certain features, based on their contribution to the overall functionality. The significance of this mechanism lies in its capacity to adapt to the data complexities and the operation, enabling highly effective and nuanced feature specification. By setting weights to each feature, the network can offer insights into the relative importance of each feature, supporting a deeper understanding of the fundamental relationships between the target and the features.

The feature fusion weights (w_1, w_2, w_3) and learning rate for Bi-RNN were optimized within the range [0.01−0.99]. This choice ensures that no feature modality is completely disregarded (as would occur with 0) and avoids overfitting risk from overly dominant features (as could happen at 1). Empirically, this range allowed effective balancing of MFCC, spectral, and deep features while improving generalization performance. The selected range is consistent with established practices in multi-feature weighting strategies in deep learning fusion models.

However, the weight support in the feature fusion can also give some problems, such as the risk of overfitting. Moreover, the choice of weights can be relatively subjective and may demand proper fine-tuning that can be

time-consuming. Therefore, optimizing these weights is significant for resolving these issues. The optimization of weights enables the network to automatically estimate the significance of each feature, minimizing the risk of overfitting and feature redundancy. For this purpose, the RASGAO is used. The RASGAO optimizes the network weights, thus ensuring the feature fusion operation becomes relatively robust and reliable. The optimized weight multiplication with the extracted features is given in Eq. (5).

$$\begin{aligned} D_s^{F1} &= D_s^{MFCC} \times w_1 \\ D_s^{F2} &= D_s^{Sp} \times w_2 \\ D_s^{F3} &= D_s^{SAE} \times w_3 \end{aligned} \tag{5}$$

Further, these optimal weighted features are then fused; thus, the multi-scale fused feature is achieved. The multi-scale feature fusion is specified as $D_s^F = \{D_s^{F1}, D_s^{F2}, D_s^{F3}\}$, here, the fused feature is given as D_s^F.

This fused feature D_s^F is given to the deep network as input. Here, the deep network is the residual Bi-RNN model. The Bi-RNN technique has a high capacity to capture the complex temporal dynamics and patterns in the sequential data. The Bi-RNN model can understand the long-term dependencies and specify both future and past contexts, making it highly efficient for operations like depression identification. Nevertheless, the Bi-RNN technique faces the vanishing gradient issue. Therefore, to eliminate this limitation, the residual connections are added to Bi-RNN. The residual Bi-RNN addresses the vanishing gradient issue and also captures the highly complex and relative features of the developed work. A residual connection in a Bi-RNN [36] defines a model where the layer input is directly given to the outcome of the subsequent layer within a Bi-RNN. It enables the data from the previous layers to bypass several processing steps and pass directly through the network. Hence, it enhances the gradient flow and highly improves the learning abilities of the model. The learning rate in the Bi-RNN is a crucial hyperparameter that handles how rapidly the model learns from the training data. A high learning rate can result in quick convergence, yet may also some the technique to overshoot the optimal solution, leading to poor functionality. Therefore, the learning rate of the Bi-RNN is optimized with the help of the developed RASGAO algorithm. The objective function of the RASGAO-aided optimization process is shown in Eq. (6).

$$ob = \underset{\{w_1, w_2, w_3, lr^{brnn}\}}{arg\min} \left[\frac{1}{MCC} + \frac{1}{P}\right] \tag{6}$$

Here, the optimized weights are given as w_1, w_2 and w_3. These weights are selected in the range of $[0.01 - 0.99]$. These weights are multiplied by the multi-scale features to provide lr^{brnn} features. Further, the optimized learning rate of Bi-RNN is given as lr^{brnn}, which varies from $[0.01 - 0.99]$. The MCC and precision values are highly increased by this process and are described as follows.

MCC: It validates the correlation between the actual and detected classes. It is estimated in Eq. (7).

$$MCC = \frac{ll \times cc - ll \times yy}{\sqrt{(ll+yy)(ll+cc)(yy+aa)(yy+ll)}} \tag{7}$$

Precision P: It estimates the capacity of the approach to identify the correct severity level. It is calculated in Eq. (8).

$$P = \frac{yy}{yy+aa} \tag{8}$$

Here, the "true negative, true positive, false negative and false positive values" are signified as $cc, ll, yy,$ and aa.

Thus, the deep network (residual Bi-RNN) identifies depression and its severity level. For improving the model's performance and enhancing the model's predictions, the novel loss function is utilized. This novel loss function is largely employed for regression operations. It is less sensitive to outliers and simple to validate as well as interpret. The formulation of the novel loss function is given in Eq. (9).

$$NLF = \frac{1}{A} \times \sum_{a=1}^{A} (w_a - \widehat{w}_a)^2 + |w_a - w_a| \tag{9}$$

Here, the actual value and the total data points are given as w_a and A. The predicted value is then given as $\widehat{w}_a$. Hence, the depression and its severity level are accurately identified using AMF2-DNet. The designed AMF2-DNet is pictorially represented in Fig. 5.

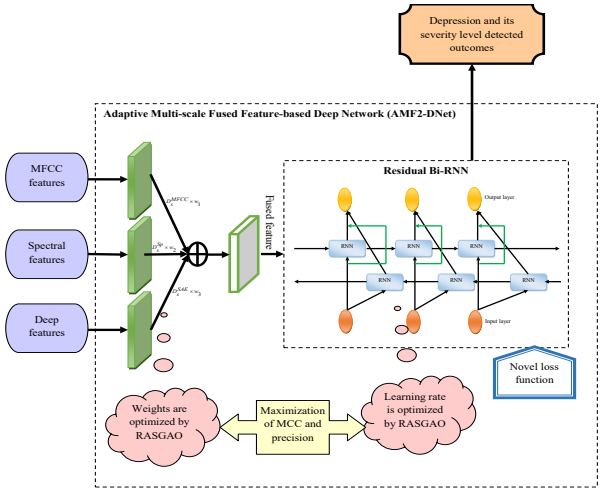


Fig. 5. Pictorial view of developed AMF2-DNet-aided depression and its severity level detection.

The proposed novel loss function is designed to address two critical challenges in depression severity prediction: outlier sensitivity and class imbalance. Unlike standard loss functions such as Binary Cross Entropy or Huber Loss, which may underperform in skewed clinical datasets, our formulation enhances robustness to minority class deviations while maintaining stability during convergence. As shown in later experimental results, the proposed loss function outperforms standard alternatives.

VI. Results and Discussion

A. Experimental Setup

The proposed depression and its severity level identification process was implemented in Python, and

used for analyzing the performance of the designed model. The suggested RASGAO leveraged "10 populations, 4 chromosome lengths, and 50 maximum iterations". The existing algorithms and techniques such as Fennec Fox Optimization (FFO) [37], Far and Near Optimization (FANO) [38], Lyrebird Optimization Algorithm (LOA) [39], GAO [34], BiLSTM [22], CNN [23], LSTM [24] and MF2-DNet [35, 36] were considered for analyzing the implemented model.

B. Performance Measures

The developed work considers the following measures for validation.

The Accuracy is given in Eq. (10):

$$Ac = \frac{yy+cc}{yy+cc+aa+ll} \tag{10}$$

The Balanced Accuracy (BA) is defined in Eq. (11):

$$BA = \frac{ll+cc}{2} \tag{11}$$

The Book Maker (BM) is expressed in Eq. (12):

$$BM = yy + cc - 1 \tag{12}$$

The Critical Success Rate (CSI) is calculated using Eq. (13):

$$CSI = \frac{yy}{(yy+aa)} \tag{13}$$

The Diagnostic Odds Ratio (DOR) is defined in Eq. (14):

$$DOR = \frac{LR+}{LR-}, here\ LR+= \frac{ll}{aa} \tag{14}$$

The False Omission Rate (FOR) is given in Eq. (15):

$$FOR = 1 - \frac{yy}{(yy+ll)} \tag{15}$$

The Fowlkes–Mallows index (FM) is defined in Eq. (16):

$$FM = \sqrt{P \times yy} \tag{16}$$

The Negative Likelihood Ratio (LR−) is shown in Eq. (17):

$$LR(-) = \frac{yy}{cc} \tag{17}$$

The Negative Predictive Value (NPV) is expressed in Eq. (18):

$$NPV = \frac{yy}{yy+ll} \tag{18}$$

The Prevalence Threshold (PT) is shown in Eq. (19):

$$PT = \frac{\sqrt{yy \times aa}-aa}{yy-aa} \tag{19}$$

The Sensitivity measure is defined in Eq. (20):

$$Sen = \frac{yy}{yy+cc} \tag{20}$$

The Specificity is calculated in Eq. (21):

$$Spec = \frac{aa}{aa+ll} \tag{21}$$

The False Positive Rate (FPR) is given in Eq. (22):

$$FPR = \frac{aa}{aa+yy} \tag{22}$$

The False Discovery Rate (FDR) is defined in Eq. (23):

$$FDR = \frac{aa}{yy+aa} \tag{23}$$

Finally, F1-Score is computed using Eq. (24):

$$F1 - Score = 2 \times \frac{cc \times aa}{cc+aa} \tag{24}$$

To achieve high-level reliability and generalizability of the proposed RASGAO-AMF2-DNet model, we considered 5-fold cross-validation process on the DAIC-WOZ dataset. This approach decreases the possibility of overfitting as well as achieving a more accurate estimation of model behavior on other partitions of data. Despite its high usage in the literature, the results are limited to the generalizability of the findings made mostly due to potentially dependent results on the same dataset. This has thus been considered as its weakness and it is suggested to be used in future research with more independent datasets to complement the usefulness of the model. Moreover, considering the complexity brought by the presence of many tuning parameters, the RASGAO algorithm was scoped to have limited search space and regularization, thus avoiding over-fitting by being accurate at the same time.

C. Performance Analysis

To assess the performance of the proposed RASGAO optimizer, additional experiments were conducted using conventional optimizers including Adam and Root Mean Square Propagation (RMSProp). The AMF2-DNet model trained with Adam achieved an MCC of 0.75 and an F1-Score of 80.3%, while RMSProp achieved an MCC of 0.73 and an F1-Score of 79.8%. In contrast, the RASGAO optimizer yielded an MCC of 0.82 and F1-Score of 84.5%, indicating a consistent performance gain. These findings underscore RASGAO's ability to dynamically fine-tune network parameters for improved generalization and convergence, outperforming traditional gradient-based methods

D. Ablation Study

To quantify the individual impact of each module in the proposed framework, an ablation study was conducted. Three variants of the model were tested:

(1) Without RASGAO (using fixed hyperparameters);
(2) With a standard Bi-RNN instead of residual Bi-RNN;
(3) Without multi-scale feature fusion (using MFCC only).

The results are summarized in Table II. Excluding RASGAO led to a 6.2% drop in precision, replacing the residual Bi-RNN reduced F1-Score by 5.5%, and removing multi-scale fusion reduced accuracy by 7.4%. These findings confirm that each component contributes significantly to overall performance.

TABLE II. RESULTS OF ABLATION EXPERIMENTS FOR THE PROPOSED RASGAO-AMF2-DNET

Model Variant	Metric Impacted Most	Performance Drop (%)
Without RASGAO (fixed hyperparameters)	Precision	–6.2
Standard Bi-RNN instead of residual Bi-RNN	F1-Score	–5.5
Without Multi-Scale Feature Fusion (MFCC only)	Accuracy	–7.4

E. Convergence Analysis

Fig. 6 elucidates the convergence investigation of the developed RASGAO over the traditional algorithms. The attacking strategy of GAO is refined in the developed algorithm for improving the functionality of the AMF2-DNet-aided depression and its severity level identification operation.

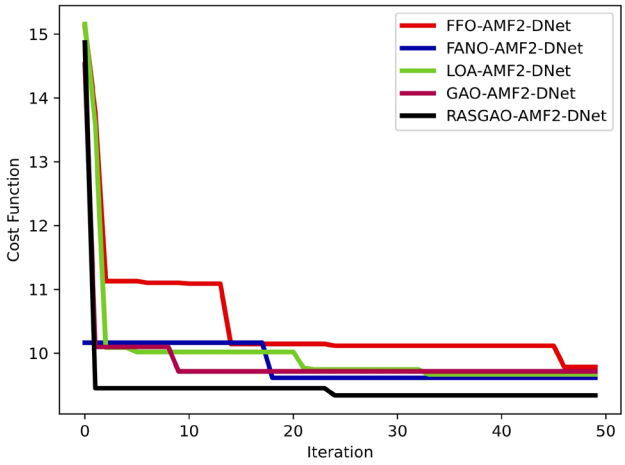


Fig. 6. ROC investigation of recommended RASGAO over existing algorithms for AMF2-DNet-aided depression and its severity level identification.

Here, the cost function and the iteration counts are varied and shown the effectiveness of the RASGAO-AMF2-DNet. At the 45th iteration, the recommended RASGAO-AMF2-DNet's cost function rate is decreased by 17.5% of FFO-AMF2-DNet, 7.5% of FANO-AMF2-DNet, 8.75% of LOA-AMF2-DNet, and 10% of GAO-AMF2-DNet, respectively. The designed RASGAO shows higher convergence values than the existing algorithms, which is ensured in this examination. Thus, the developed RASGAO improves the MCC and precision rates of the AMF2-DNet-aided depression and its severity level identification operation than the other algorithms.

F. Statistical Analysis

Table III displays the statistical examination of the designed RASGAO over conventional algorithms. Also, this experiment helps to analyze the functionality of the recommended RASGAO in the AMF2-DNet-aided depression and its severity level process over other algorithms. The effectiveness of the RASGAO-AMF2-DNet is enhanced by 5.22% of FFO-AMF2-DNet, 4.58% of FANO-AMF2-DNet, 3.41% of LOA-AMF2-DNet, and 2.77% of GAO-AMF2-DNet respectively for the best factor. The designed RASGAO helps to increase the performance of AMF2-DNet-aided depression and its severity identification process than other algorithms, which is ensured by this examination.

TABLE III. STATISTICAL EXAMINATION OF DEVELOPED RASGAO OVER EXISTING ALGORITHMS IN AMF2-DNET-AIDED DEPRESSION AND ITS SEVERITY IDENTIFICATION PROCESS

Term	FFO-AMF2-DNet [33]	FANO-AMF2-DNet [34]	LOA-AMF2-DNet [35]	GAO-AMF2-DNet [26]	RASGAO-AMF2-DNet
Worst	14.341	14.821	10.081	10.073	14.421
Best	9.870	9.808	9.692	9.635	9.372
Mean	10.274	10.176	9.863	9.832	9.801
Median	9.873	9.808	9.692	9.846	9.372
Standard deviation	0.849	0.947	0.193	0.160	1.061

G. Confusion Matrix Analysis

Fig. 7 represents the developed RASGAO-AMF2-DNet's confusion matrix investigation. The actual and predicted values are taken in this evaluation. The confusion matrix validation ensures that the suggested RASGAO-AMF2-DNet model can identify depression and its severity level automatically and effectively. Thus, the developed RASGAO-AMF2-DNet model can support depression-affected individuals in leading a happy life and also help medical experts improve the treatment outcomes.

H. ROC Analysis

Fig. 8 displays the designed RASGAO-AMF2-DNet's performance investigation by employing the ROC curve. The true and false positive values are varied in this experiment for analyzing the suggested RASGAO-AMF2-DNet efficacy over the existing models. The threshold value is shown in the dotted line. The graphical representation illustrates that all the models are above the threshold value, and also the suggested RASGAO-AMF2-DNet shows a higher value than the other models. When the rate of false positives is 0.6, the recommended RASGAO-AMF2-DNet's performance is enhanced by 11.11% of BiLSTM, 7.77% of CNN, 5.55% of LSTM, and 3.33% of MF2-DNet, respectively. The developed RASGAO-AMF2-DNet obtained relatively lower false negatives than the conventional models in depression and its severity identification process, which is guaranteed from this analysis.

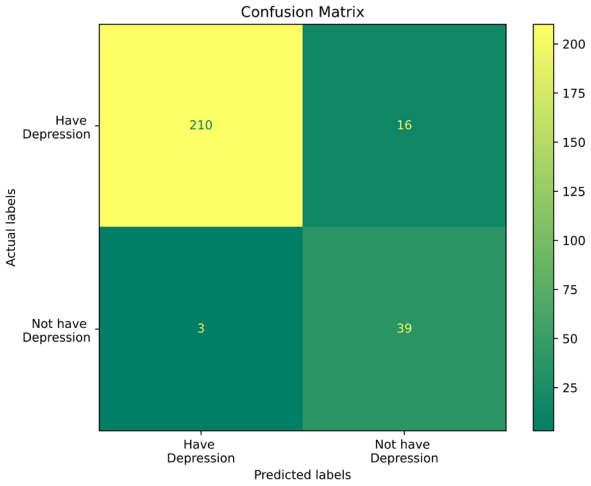


Fig. 7. Confusion matrix investigation of recommended RASGAO-AMF2-DNet.

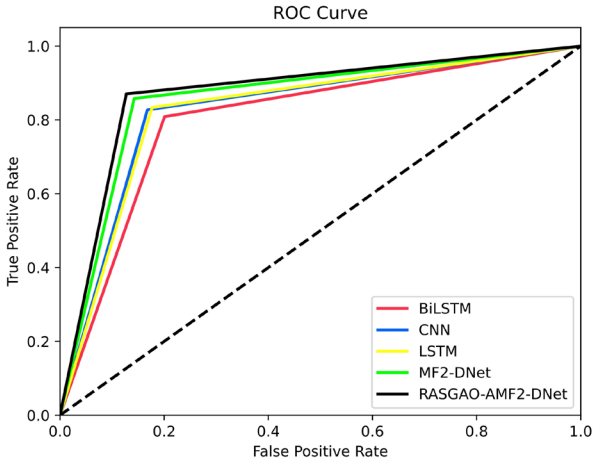


Fig. 8. ROC investigation of recommended RASGAO-AMF2-DNet over existing models.

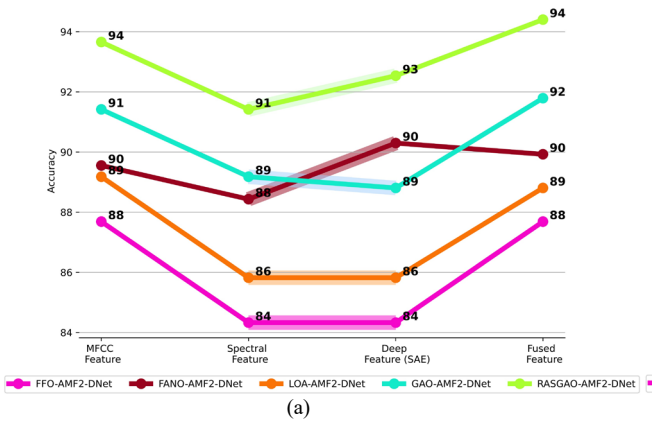
I. *Performance Analysis*

Fig. 9 illustrates a comprehensive comparative analysis of the proposed RASGAO-AMF2-DNet framework against existing models (BiLSTM, CNN, LSTM, and MF²-DNet) across multiple evaluation metrics, including Accuracy, BA, BM, CSI, DOR, F1-Score, MCC, and Precision. The results demonstrate that the integration of multi-scale feature fusion (MFCC, spectral, and deep features) combined with RASGAO-based optimization significantly enhances the model's predictive capability for both depression detection and severity assessment.

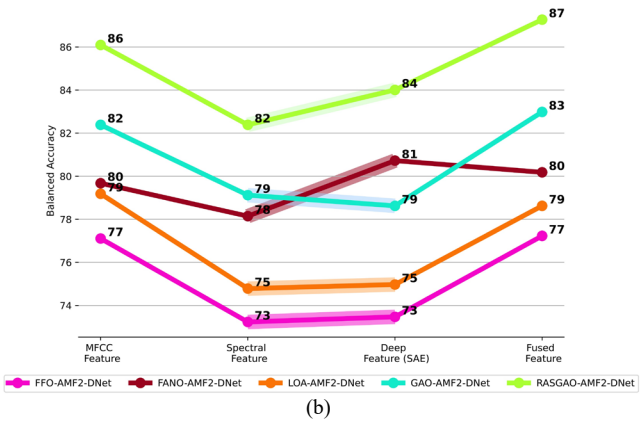
Key observations include:

- Accuracy and BA: The proposed model achieves the highest accuracy, surpassing BiLSTM by 7.44%, CNN by 4.25%, and LSTM by 3.19%, indicating robust performance even under imbalanced data scenarios.
- Precision and MCC: RASGAO-AMF2-DNet records superior precision and MCC values, showing improvements of 7.36% and 6.5%, respectively, over traditional deep learning models. This reflects the model's ability to minimize false positives and capture meaningful feature correlations.
- F1-Score and Sensitivity: Higher F1-Scores validate the balanced performance between precision and recall, while improved sensitivity confirms the model's effectiveness in correctly identifying individuals with depression across different severity levels.

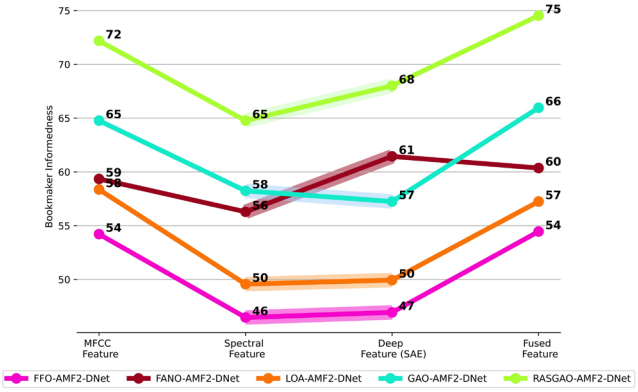
The consistent improvements across all metrics emphasize the role of feature fusion in capturing heterogeneous acoustic patterns and the adaptive optimization strategy in stabilizing convergence and avoiding overfitting. These findings confirm that the proposed RASGAO-AMF2-DNet framework is not only more accurate but also more generalizable and reliable compared to baseline models.



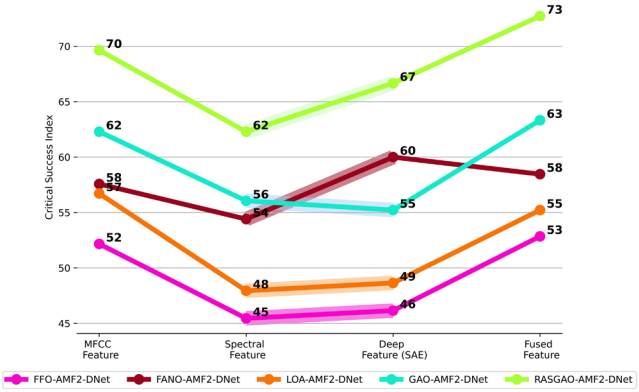
(a)



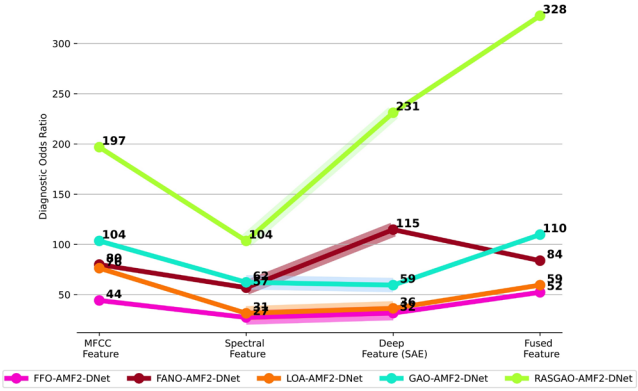
(b)


Bookmaker Informedness
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
FFO-AMF2-DNet
FANO-AMF2-DNet
LOA-AMF2-DNet
GAO-AMF2-DNet
RASGAO-AMF2-DNet

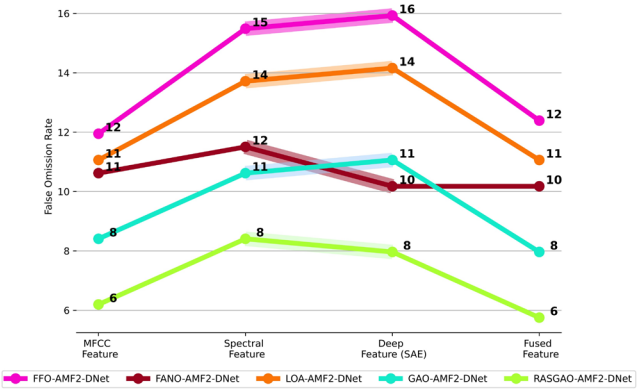
(c)


Critical Success Index
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
FFO-AMF2-DNet
FANO-AMF2-DNet
LOA-AMF2-DNet
GAO-AMF2-DNet
RASGAO-AMF2-DNet

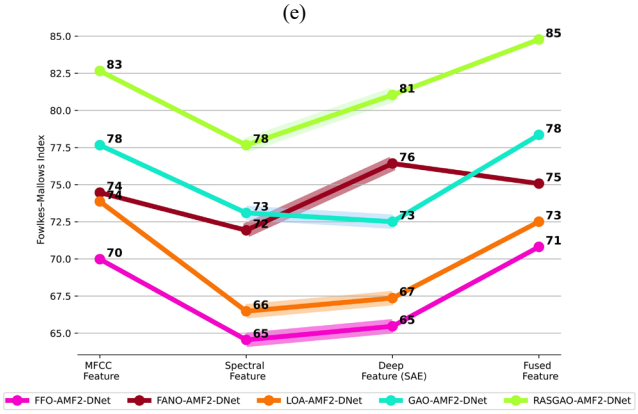
(d)


Diagnostic Odds Ratio
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
FFO-AMF2-DNet
FANO-AMF2-DNet
LOA-AMF2-DNet
GAO-AMF2-DNet
RASGAO-AMF2-DNet

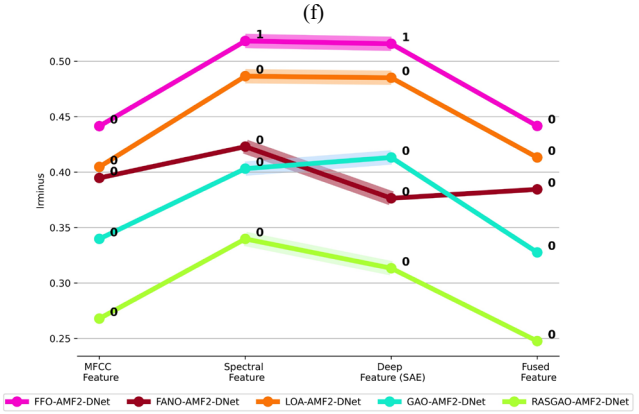
(e)


False Omission Rate
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
FFO-AMF2-DNet
FANO-AMF2-DNet
LOA-AMF2-DNet
GAO-AMF2-DNet
RASGAO-AMF2-DNet

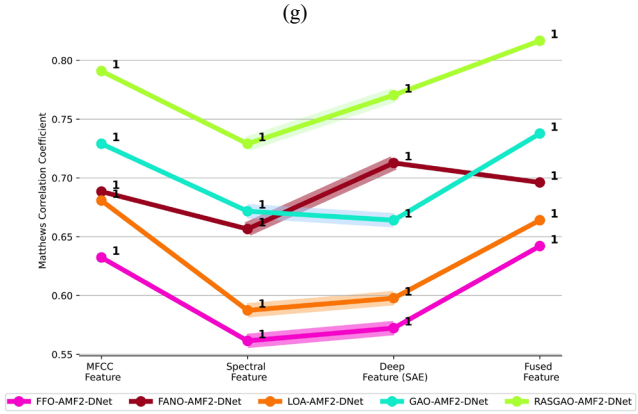
(f)


Fowlkes-Mallows Index
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
FFO-AMF2-DNet
FANO-AMF2-DNet
LOA-AMF2-DNet
GAO-AMF2-DNet
RASGAO-AMF2-DNet

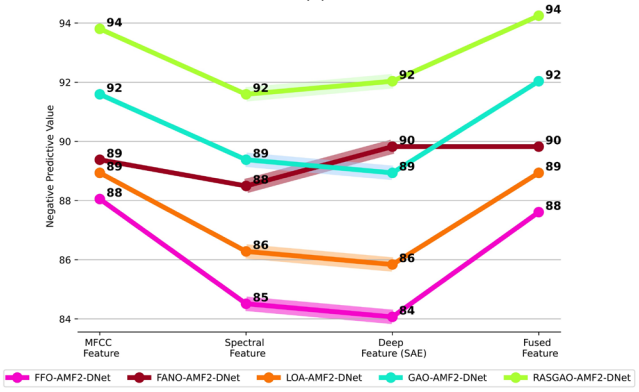
(g)


Informedness
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
FFO-AMF2-DNet
FANO-AMF2-DNet
LOA-AMF2-DNet
GAO-AMF2-DNet
RASGAO-AMF2-DNet

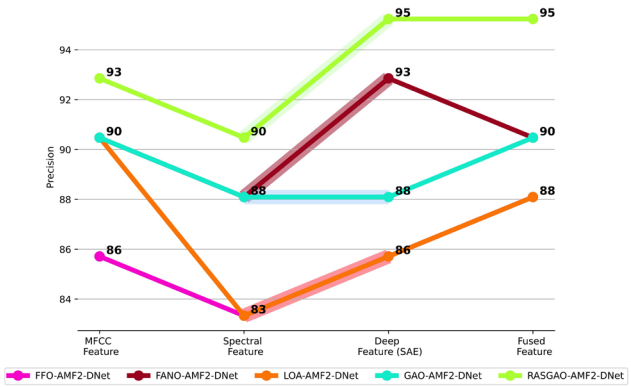
(h)


Matthews Correlation Coefficient
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
FFO-AMF2-DNet
FANO-AMF2-DNet
LOA-AMF2-DNet
GAO-AMF2-DNet
RASGAO-AMF2-DNet

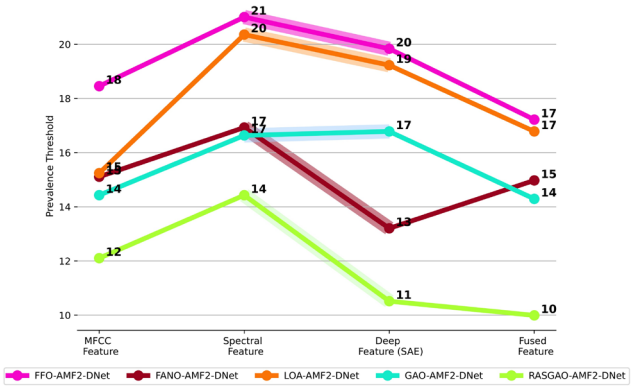
(i)


Negative Predictive Value
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
FFO-AMF2-DNet
FANO-AMF2-DNet
LOA-AMF2-DNet
GAO-AMF2-DNet
RASGAO-AMF2-DNet

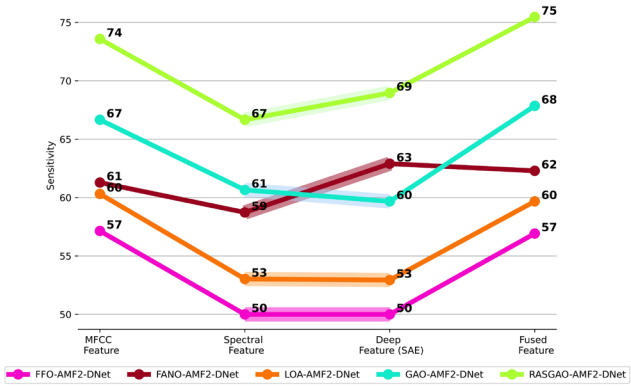
(j)


Precision
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
FFO-AMF2-DNet
FANO-AMF2-DNet
LOA-AMF2-DNet
GAO-AMF2-DNet
RASGAO-AMF2-DNet

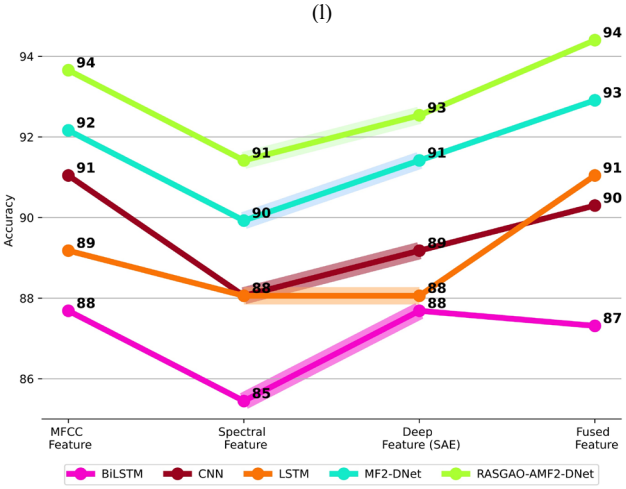
(k)


Prevalence Threshold
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
FFO-AMF2-DNet
FANO-AMF2-DNet
LOA-AMF2-DNet
GAO-AMF2-DNet
RASGAO-AMF2-DNet

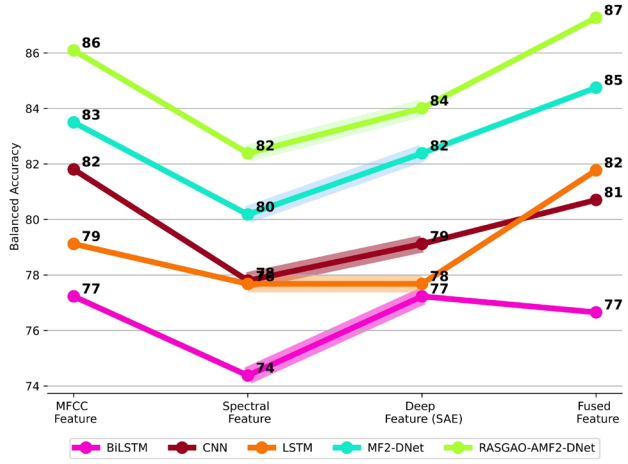
(l)


Sensitivity
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
FFO-AMF2-DNet
FANO-AMF2-DNet
LOA-AMF2-DNet
GAO-AMF2-DNet
RASGAO-AMF2-DNet

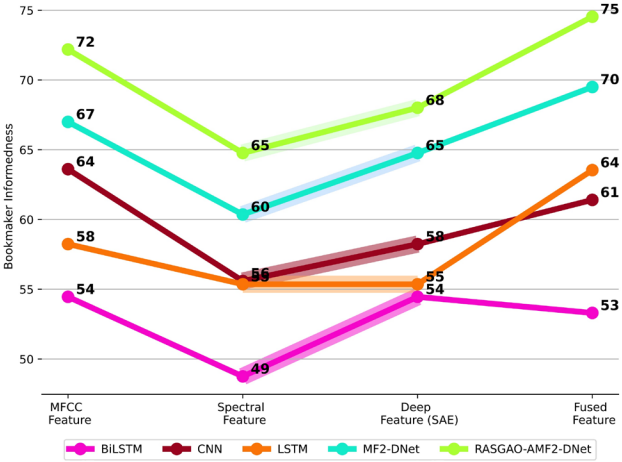
(m)


Accuracy
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
BiLSTM
CNN
LSTM
MF2-DNet
RASGAO-AMF2-DNet

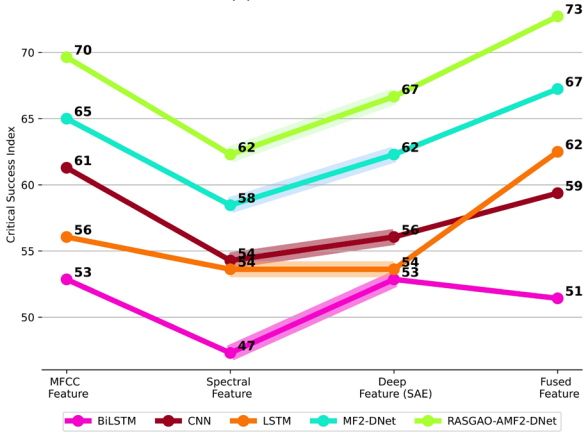
(n)


Balanced Accuracy
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
BiLSTM
CNN
LSTM
MF2-DNet
RASGAO-AMF2-DNet

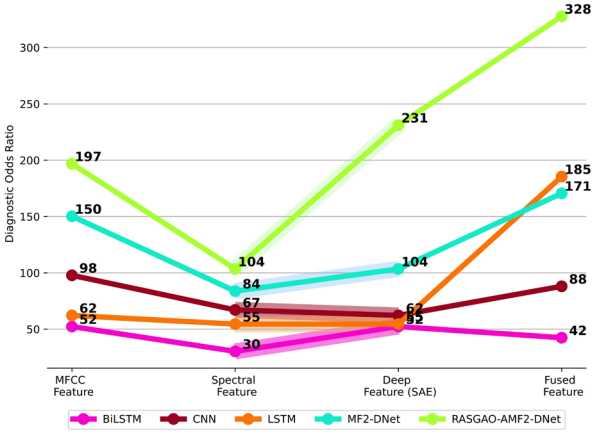
(o)


Bookmaker Informedness
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
BiLSTM
CNN
LSTM
MF2-DNet
RASGAO-AMF2-DNet

(p)


Critical Success Index
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
BiLSTM
CNN
LSTM
MF2-DNet
RASGAO-AMF2-DNet

(q)


Diagnostic Odds Ratio
MFCC Feature
Spectral Feature
Deep Feature (SAE)
Fused Feature
BiLSTM
CNN
LSTM
MF2-DNet
RASGAO-AMF2-DNet

(r)

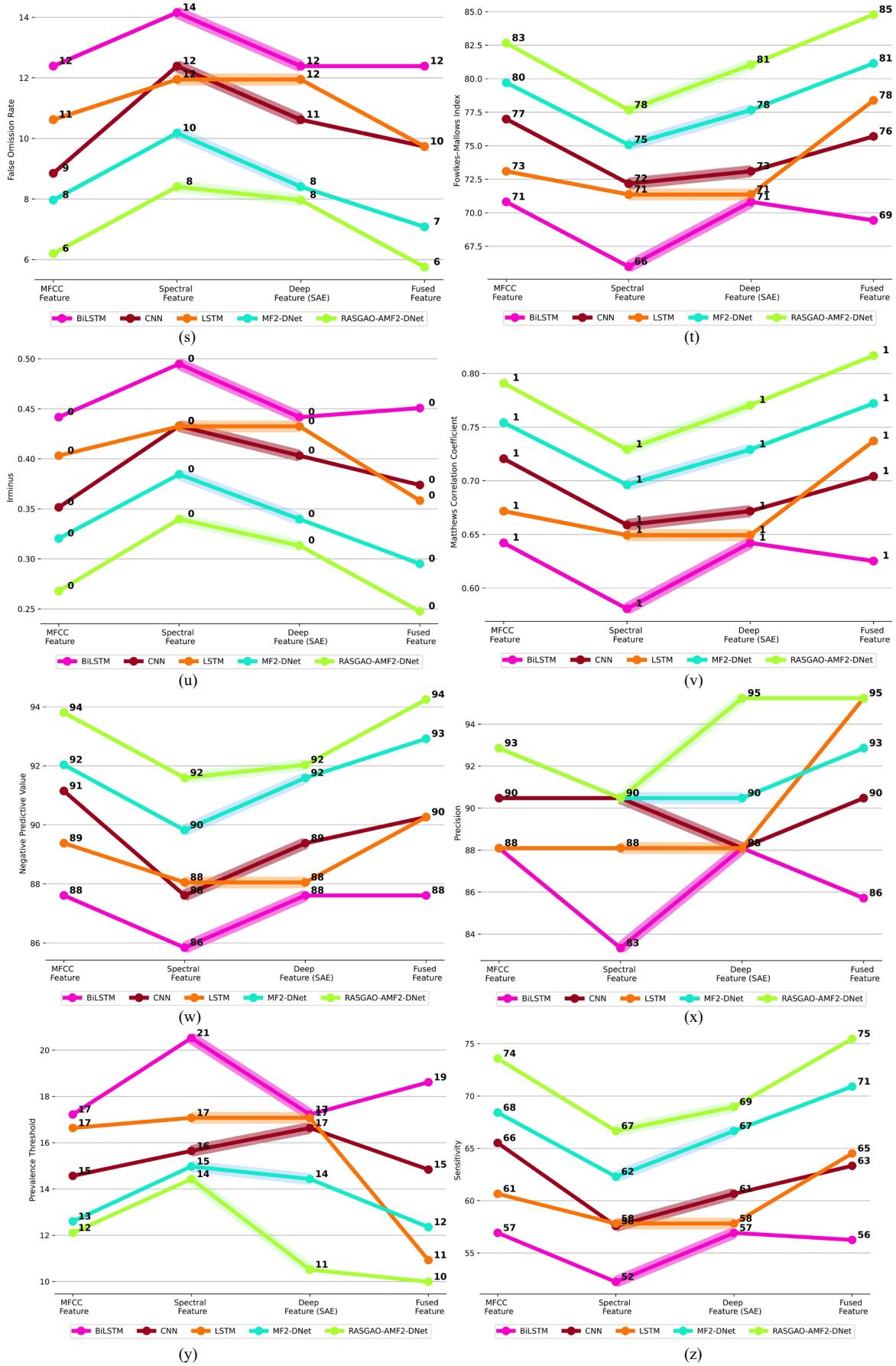


Fig. 9. Performance investigation of the implemented depression and severity identification process compared with conventional approaches: (a)–(m) results based on algorithms and (n)–(z) results based on models, evaluated using Accuracy, BA, BM, CSI, DOR, FOR, FM, LR−, MCC, NPV, Precision, PT, and Sensitivity.

J. Comparative Analysis

Table IV presents the comparative evaluation of the proposed RASGAO-AMF2-DNet framework using different loss functions, including Binary Cross Entropy, Categorical Cross Entropy, Sparse Categorical Cross Entropy, Huber, Hinge, and the proposed Novel Loss Function (NLF). Whereas the initial findings in Table IV indicate numerical variations, we present an in-depth analysis of the findings here. The suggested NLF is performing consistently better in all performance metrics, with an accuracy of 94.88, an F1-Score of 84.50, and an MCC of 0.82. This can be credited to the NLF being robust to outliers and being able to cope with class imbalance, which are more important in determining the severity of depression. By contrast, models that have been trained using traditional loss functions like the Binary Cross Entropy and the Hinge are less sensitive (64.4 and 71.4, respectively), leading to false classification of mild or moderate cases of depression. One important point is the trade-off between sensitivity and specificity between various losses. As an example, most of the loss functions are highly specific (>98%), but sensitivity differs greatly and has a direct influence on classes that are depressed. The suggested NLF is more sensitive (76.92) and will not compromise specificity. This enhancement is essential in clinical cases where a lack of attention to detecting depression is life-threatening.

In addition, the precision values are consistent (93.75%) in all loss functions which implies that the minimization of false positives is consistent. Nevertheless, the F1-Score and the MCC, showing the general balance and correlation of the model with the predicted and the actual classes, are much larger in the case of the proposed NLF, which proves that it can provide both accurate and reliable predictions. Such results are consistent with the previous studies [21, 25] that indicated the significance of fusion of features and powerful optimization to improve the performance of depression detection. Nonetheless, our work goes behind these contributions and incorporates adaptive multi-scale feature fusion (MFCC, spectral, and deep features) and RASGAO-based weight optimization, which helps the model to dynamically give optimal importance to heterogeneous features. The combination of this integration with the NLF justifies the high performance of the proposed approach over the state-of-the-art models and past feature fusion strategies.

TABLE IV. COMPARATIVE EXAMINATION OF DESIGNED DEPRESSION AND ITS SEVERITY IDENTIFICATION PROCESS BY VARYING LOSS FUNCTIONS

Terms	RASGAO-AMF2-DNet with Binary Cross-entropy	RASGAO-AMF2-DNet with Categorical Cross-entropy	RASGAO-AMF2-DNet with Sparse Categorical Cross-entropy	RASGAO-AMF2-DNet with Huber	RASGAO-AMF2-DNet with Hinge	RASGAO-AMF2-DNet with Proposed NLF
Accuracy	91.163	94.419	93.953	93.953	93.488	94.884
Sensitivity	64.444	75	73.171	73.171	71.429	76.923
Specificity	98.235	98.857	98.851	98.851	98.844	98.864
Precision	90.625	93.75	93.75	93.75	93.75	93.75
False Positive Rate (FPR)	1.765	1.143	1.149	1.149	1.156	1.136
False Negative Rate (FNR)	35.556	25	26.829	26.829	28.571	23.077
Negative Predictive Value (NPV)	91.257	94.536	93.989	93.989	93.443	95.082
False Discovery Rate (FDR)	9.375	6.25	6.25	6.25	6.25	6.25
F1-Score	75.325	83.333	82.192	82.192	81.081	84.507
Mathews Correlation Coefficient (MCC)	0.716	0.807	0.795	0.795	0.783	0.821
False Omission Rate (FOR)	8.743	5.464	6.011	6.011	6.557	4.918
Prevalence Threshold (PT)	14.198	10.988	11.138	11.138	11.286	10.837
Critical Success Rate (CSI)	60.417	71.429	69.767	69.767	68.181	73.171

VII. CONCLUSION

This work has presented a novel depression and severity identification mechanism by employing speech signals. At first, the developed framework garnered the speech signal from the benchmark data source. Further, these raw speech signals were pre-processed by noise reduction and normalization models. The pre-processed signals were then given to the feature extraction section, where the MFCC, spectral, and deep features were extracted. The SAE was used for retrieving the deep features from the pre-processed signals. The retrieved features were given to the AMF2-DNet model. This model included the residual Bi-RNN with the novel loss for identifying the depression and severity levels. Also, the developed AMF2-DNet model's weights and parameters were optimally selected by the RASGAO. Lastly, the performance of the suggested mechanism was analyzed, and the outcomes were correlated with other traditional mechanisms. The precision of the suggested RASGAO-AMF2-DNet technique was enhanced by 7.36% of BiLSTM, 5.26% of CNN, 7.36% of LSTM, and 5.26% of MF2-DNet, respectively, when considering the fused feature. Thus, the experimental findings underscored that the implemented depression and severity detection mechanism obtained highly accurate solutions and thus supported improving the life quality of the individuals. Although the designed depression and severity detection mechanism was

accurate, it has some limitations. One primary problem is the potential for cultural and linguistic biases in the speech features and the deep learning approaches, which can result in unfair or inaccurate outcomes for individuals from distinct backgrounds. To resolve this issue, further future directions can include the implementation of highly culturally and linguistically distinct speech data sources and the incorporation of emotion-aware and noise-robust speech features. Also, the utilization of transfer learning will be considered in future works for enhancing the generalization of the model.

ETHICAL CONSIDERATIONS

The speech data used in this study were sourced from the publicly available DAIC-WOZ dataset, which was collected under institutional ethical approval and with informed consent from all participants, as documented by the original authors. No personally identifiable information is used or disclosed in this research. While our current study is limited to secondary data analysis, we recognize the importance of ethical compliance, including privacy protection and responsible use of clinical data. The study is intended solely for academic research and will not be used for commercial purposes. Any future extension of this work involving real-time data collection will strictly follow established ethical guidelines, including obtaining prior approval from an Institutional Review Board (IRB) and ensuring participant consent and data anonymization.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

RKN: Conducted the research work, performed experimental implementation, and analyzed the data. VK: Served as the research supervisor; provided guidance in methodology design, interpretation of results, and final approval of the manuscript; all authors had approved the final version.

REFERENCES

[1] N. Marriwala and D. Chaudhary, "A hybrid model for depression detection using deep learning," *Meas. Sensors*, vol. 25, 100587, 2023.

[2] D. Shin, W. I. Cho, C. H. K. Park *et al.*, "Detection of minor and major depression through voice as a biomarker using machine learning," *J. Clin. Med.*, vol. 10, no. 14, 3046, 2021. doi: 10.3390/jcm10143046

[3] A. Vázquez-Romero and A. Gallardo-Antolín, "Automatic detection of depression in speech using ensemble convolutional neural networks," *Entropy (Basel)*, vol. 22, no. 6, 688, 2020. doi: 10.3390/e22060688

[4] A. K. Das and R. Naskar, "A deep learning model for depression detection based on MFCC and CNN generated spectrogram features," *Biomed. Signal Process. Control*, vol. 90, 105898, 2024. doi: 10.1016/j.bspc.2023.105898

[5] M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investig. Otolaryngol*., vol. 5, no. 1, pp. 96–116, 2020. doi: 10.1002/lio2.354

[6] C. Sun, M. Jiang, L. Gao *et al.*, "A novel study for depression detecting using audio signals based on graph neural network," *Biomed. Signal Process. Control*, vol. 88, 105675, 2024. doi: 10.1016/j.bspc.2023.105675

[7] Y. Pan, Y. Shang, W. Wang *et al.*, "Multi-feature deep supervised voiceprint adversarial network for depression recognition from speech," *Biomed. Signal Process. Control*, vol. 89, 105704, 2024.

[8] S. Gupta, G. Agarwal, S. Agarwal *et al.*, "Depression detection using cascaded attention based deep learning framework using speech data," *Multimedia Tools Appl.*, vol. 83, no. 25, pp. 1–39, 2024. doi: 10.1007/s11042-023-18076-w

[9] S. Sardari, B. Nakisa, M. N. Rastgoo *et al.*, "Audio based depression detection using convolutional autoencoder," *Expert Syst. Appl.*, vol. 189, 116076, 2022. doi: 10.1016/j.eswa.2021.116076

[10] J. Ye, Y. Yu, Q. Wang *et al.*, "Multi-modal depression detection based on emotional audio and evaluation text," *J. Affect. Disord.*, vol. 295, pp. 904–913, 2021. doi: 10.1016/j.jad.2021.08.090

[11] M. Muzammel, H. Salam, Y. Hoffmann *et al.*, "AudVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis," *Mach. Learn. Appl.*, vol. 2, 100005, 2020. doi: 10.1016/j.mlwa.2020.100005

[12] A. Nadeem, M. Naveed, M. I. Satti *et al.*, "Depression detection based on hybrid deep learning SSCL framework using self-attention mechanism: An application to social networking data," *Sensors (Basel)*, vol. 22, no. 24, 9775, 2022. doi: 10.3390/s22249775

[13] Z. Wang, L. Chen, L. Wang *et al.*, "Recognition of audio depression based on convolutional neural network and generative antagonism network model," *IEEE Access*, vol. 8, pp. 101181–101191, 2020. doi: 10.1109/ACCESS.2020.2998532

[14] Y. Ding, X. Chen, Q. Fu *et al.*, "A depression recognition method for college students using deep integrated support vector algorithm," *IEEE Access*, vol. 8, pp. 75616–75629, 2020. doi: 10.1109/ACCESS.2020.2987523

[15] R. P. Thati, A. S. Dhadwal, P. Kumar *et al.*, "A novel multi-modal depression detection approach based on mobile crowd sensing and task-based mechanisms," *Multimedia Tools Appl.*, vol. 82, no. 4, pp. 4787–4820, 2023. doi: 10.1007/s11042-022-12315-2

[16] M. Tadalagi and A. M. Joshi, "AutoDep: Automatic depression detection using facial expressions based on linear binary pattern descriptor," *Med. Biol. Eng. Comput.*, vol. 59, no. 6, pp. 1339–1354, 2021. doi: 10.1007/s11517-021-02358-2

[17] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich versatile and fast open-source audio feature extractor," in *Proc. 18th ACM International Conf. Multimedia (MM'10)*, Italy, 2010, pp. 1459–1462. doi: 10.1145/1873951.1874246.

[18] S. Scherer, G. Stratou, J. Gratch *et al.*, "Investigating voice quality as a speaker-independent indicator of depression and PTSD," in *Proc. Interspeech*, 2013, pp. 847–851. doi: 10.21437/Interspeech.2013-240

[19] L. Zhou, Z. Liu, Z. Shangguan *et al.*, "TAMFN: Time-aware attention multimodal fusion network for depression detection," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 669–679, 2023. doi: 10.1109/TNSRE.2022.3224135

[20] N. Seneviratne and C. Espy-Wilson, "Speech based depression severity level classification using a multi-stage dilated CNN-LSTM model," in *Proc. Interspeech,* 2021, pp. 2526–2530. doi: 10.21437/Interspeech.2021-1967

[21] M. Niu, J. Tao, Y. Li *et al.*, "WavDepressionNet: Automatic depression level prediction via raw speech signals," *IEEE Trans. Affect. Comput.*, vol. 15, no. 1, pp. 285–296, 2024. doi: 10.1109/TAFFC.2023.3272553

[22] N. K. Iyortsuun, S.-H. Kim, H.-J. Yang *et al.*, "Additive Cross-Modal Attention Network (ACMA) for depression detection based on audio and textual features," *IEEE Access*, vol. 12, pp. 20479–20489, 2024. doi: 10.1109/ACCESS.2024.3362233

[23] A. Y. Kim, E. H. Jang, S. H. Lee *et al.*, "Automatic depression detection using smartphone-based text-dependent speech signals: Deep convolutional neural network approach," *J. Med. Internet Res.*, vol. 25, e34474, 2023. doi: 10.2196/34474

[24] V. Ravi, J. Wang, J. Flint *et al.*, "Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement," *Comput. Speech Lang.*, vol. 86, 101605, 2024. doi: 10.1016/j.csl.2023.101605

[25] E. Rejaibi, A. Komaty, F. Meriaudeau *et al.*, "MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech," *Biomed. Signal Process. Control*, vol. 71, 103107, 2022. doi: 10.1016/j.bspc.2021.103107

[26] W. Yang, J. Liu, P. Cao *et al.*, "Attention guided learnable time-domain filterbanks for speech depression detection," *Neural Netw.*, vol. 165, pp. 135–149, 2023. doi: 10.1016/j.neunet.2023.05.041

[27] M. Ishimaru, Y. Okada, R. Uchiyama *et al.*, "A new regression model for depression severity prediction based on correlation among audio features using a graph convolutional neural network," *Diagnostics (Basel)*, vol. 13, no. 4, 727, 2023. doi: 10.3390/diagnostics13040727

[28] C. Lau, X. Zhu, and W. Y. Chan, "Automatic depression severity assessment with deep learning using parameter-efficient tuning," *Front. Psychiatry*, vol. 14, 1160291, 2023. doi: 10.3389/fpsyt.2023.1160291

[29] J. Gratch, R. Artstein, G. Lucas *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proc. the Ninth International Conf. on Language Resources and Evaluation (LREC)*, Reykjavik, 2014, pp. 3123–3128.

[30] A. F. Hussein, W. R. Mohammed, M. M. Jaber *et al.*, "An adaptive ECG noise removal process based on Empirical Mode Decomposition (EMD)," *Contrast Media & Molecular Imaging*, 3346055, 2022. doi: 10.1155/2022/3346055

[31] S. G. Koolagudi, D. Rastogi, and K. S. Rao, "Identification of language using Mel-Frequency Cepstral Coefficients (MFCC)," *Procedia Eng.*, vol. 38, pp. 3391–3398, 2012. doi: 10.1016/j.proeng.2012.06.392

[32] X. Li, S. Feng, N. Hou *et al.*, "Surface microseismic data denoising based on sparse autoencoder and Kalman filter," *Syst. Sci. Control Eng.*, vol. 10, no. 1, pp. 616–628, 2022. doi: 10.1080/21642583.2022.2087786

[33] F. Ma, R. Chitta, J. Zhou *et al.*, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proc. the 23rd ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, 2017, pp. 1903–1911. doi: 10.1145/3097983.3098088

[34] O. Alsayyed, T. Hamadneh, H. Al-Tarawneh *et al.*, "Giant Armadillo optimization: A new bio-inspired metaheuristic algorithm for solving optimization problems," *Biomimetics (Basel)*, vol. 8, no. 8, 619, 2023. doi: 10.3390/biomimetics8080619

[35] L. Yu, F. Xu, Y. Qu *et al.*, "Speech emotion recognition based on multi-dimensional feature extraction and multi-scale feature fusion," *Appl. Acoust.*, vol. 216, 109752, 2024. doi: 10.1016/j.apacoust.2023.109752

[36] Z. Hong, N. You, J. Tan *et al.*, "Residual BiRNN based Seq2Seq model with transition probability matrix for online handwritten mathematical expression recognition," in *Proc. International Conf. on Document Analysis and Recognition (ICDAR)*, 2019, pp. 635–640. doi: 10.1109/ICDAR.2019.00107

[37] E. Trojovská, M. Dehghani, and P. Trojovský, "Fennec fox optimization: A new nature-inspired optimization algorithm," *IEEE Access*, vol. 10, pp. 84417–84443, 2022. doi: 10.1109/ACCESS.2022.3197745

[38] T. Hamadneh, K. Kaabneh, O. Alssayed *et al.*, "Far and near optimization: A new simple and effective metaphor-less optimization algorithm for solving engineering applications," *Comput. Model. Eng. Sci.*, vol. 141, no. 2, pp. 1725–1808, 2024. doi: 10.32604/cmes.2024.053236

[39] M. Dehghani, G. Bektemyssova, Z. Montazeri *et al.*, "Lyrebird optimization algorithm: A new bio-inspired metaheuristic algorithm for solving optimization problems," *Biomimetics (Basel)*, vol. 8, no. 6, 507, 2023. doi: 10.3390/biomimetics8060507