International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 10-No.3 (2024) pp. 491-506
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

KWHO-CNN: A Hybrid Metaheuristic Algorithm Based Optimized Attention-Driven CNN For Automatic Clinical Depression Recognition

Priti Parag GAIKWAD[1*,2], Mithra VENKATESAN[3]

[1*]Electronics and Telecommunication Engineering Research Scholar Dr. D. Y. Patil Institute of Technology Pimpri, Pune, Maharashtra Pimpri, Pune
[2] Sinhgad College of Engineering, Vadgaon Pune, Maharashtra
* **Corresponding Author Email:** ppgaikwad.scoe@sinhgad.edu- **ORCID:** 0009-0008-9919-9828

[3]AI&DS HoD and D. Y. Patil Institute of Technology Pimpri Pune Maharashtra Pimpri, Pune, Maharashtra
Email: mithra.v@dypvp.edu.in - **ORCID:** 0000-0002-6541-447X

Article Info:

DOI: 10.22399/ijcesen.359
Received : 27 June 2024
Accepted : 6 August 2024

Keywords :

Deep learning
Biomedical Signal processing
Clinical depression diagnosis
Speech depression recognition

Abstract:

Depression is a widespread mental disorder with inconsistent symptoms that make diagnosis challenging in clinical practice and research. Nevertheless, the poor identification may be partially explained by present approaches ignoring patients' vocal tract modifications in favor of merely considering speech perception aspects. This study proposes a novel framework, KWHO-CNN, integrating a hybrid metaheuristic algorithm with Attention-Driven Convolutional Neural Networks (CNNs), to enhance depression detection using speech data. It addresses challenges like variability in speech patterns and small sample sizes by optimizing feature selection and classification. Initial preprocessing involves noise reduction using Adaptive filtering techniques, data normalization employs Vocal Tract Length Normalization (VTLN), and segmentation with Independent Component Analysis (ICA), followed by feature extraction, primarily utilizing Mel-frequency cepstral coefficients (MFCCs). The Krill Wolf Hybrid Optimization (KWHO) Algorithm optimizes these features, overcoming issues of over-fitting and enhancing model performance. The Attention-Driven CNN architecture further refines classification, leveraging dense computations and architectural homogeneity. The result shows that the suggested model demonstrates superior performance and outperforms depression diagnosis, with over 98% accuracy, 96% precision, and 96% recall. The proposed framework represents a promising advancement in depression detection, offering a robust solution that could significantly impact clinical practice and research in the field of mental health.

1. Introduction

Depression is a common but dangerous psychological illness that is defined by social dysfunction, cognitive deterioration, and a persistent sense of sadness [1]. According to the World Health Organisation (WHO), over 300 million people worldwide suffer from depression, and the number is rising, particularly among the elderly [2]. Major Depressive Disorder (MDD) is the scientific term for depression, which is a mental illness marked by low mood, low self-esteem, loss of interest, low energy, and pain that persists for a long time without a clear reason. It has detrimental effects on a person's eating, sleeping, working, and family life. According to [3] half of all completed suicides are related to depression and other mood disorders. As a result, a lot of studies has gone into creating methods for identifying and preventing this mental illness so that psychologists and psychiatrists can start helping patients right away [4]. Adolescents are frequently the first to experience depression, which can persist or recur in adulthood and eventually develop a chronic mental illness that lasts a lifetime [5]. Conventional mental health disorder reports, like the Patient Health

Questionnaire (PHQ-8), Geriatric Depression Scale (GDS), Hamilton Rating Scale for Depression (HRSD), and Beck's Depression Inventory (BDI-II), are not entirely helpful or require too much thought on the part of the psychiatrist to diagnose the condition [6]. Additionally, the patients find them to be rather uninteresting, which may skew their opinions. However, in recent years, there has been an increasing amount of interest in the idea of using automated techniques to diagnose depression by utilising the interaction between people and computer-aided devices through an appropriate description of measurable behavioural depression traits [7]. While some writers examine speaking rates and silences in read and unprompted speech [8, 9], others take into account behavioural traits like handwriting and sketching [10].

Research on depression diagnosis with the application of AI technology has generally been carried out with several different types of data, such as face image, electroencephalogram (EEG), human voice, behaviour, and text. Linguistic data, such as speech and text, especially reveal the characteristics of patients with depression. This is because persons with symptoms of depressive disorder exhibit speech characteristics, such as reduced vocal intensity, reduced pitch range, and slower speech [11]. Numerous studies have demonstrated a strong correlation between speech signals and state of mind. Recently, several speech-based machine-learning techniques for depression detection have been developed to aid doctors in diagnosing specific cases of depression [12]. Over the last decade, social media has emerged to be an extensively used platform for the exchange of ideas and information. A small chunk of text can echo the mental state of a person. Hence, practitioners can gain a good amount of intuition about the mental well-being and health of an individual from a Facebook post, a tweet, or an Instagram post [13]. While these speech and text data can be used to analyse the characteristics of depression, since depression is a complex mental disorder, analyzing a specific aspect using a single mode of data may not be sufficient for effective evaluation. Several studies have demonstrated that the fusion of various modalities of data significantly increases the accuracy of depression diagnosis [14, 15].
The authors have employed diverse methodologies to scrutinize speech patterns, such as natural language processing or prosodic and acoustic analysis. It has been discovered that characteristics such as the fundamental frequency and intensity, which can be extracted using acoustic analysis, are associated with depression. Prosodic analysis has been used to examine speech patterns and intonation, which can reveal information about a person's emotional state [16]. The substance of the speech, including lexical features and the subjects covered, has also been analysed using natural language processing; depressed people typically favour particular words and talk about particular subjects more frequently [17]. Support Vector Machines are among the top machine learning models for speech-based depression identification that have been documented in recent research. Among the best machine learning models that have been described in recent research for depression identification from speech are Support Vector Machines (SVMs), Random Forests (RFs) [18, 19], Deep Neural Networks (DNNs), and Convolutional Neural Networks (CNNs) [20]. Therefore, research is needed to enable accurate diagnosis of depression considering multiple types of data and their characteristics.

2. Related works

Rejaibi, E., et al., [21] A deep Recurrent Neural Network-based framework has been created for speech-based clinical depression identification and prediction. The framework extracts high-level and low-level audio features from audio recordings to predict the Patient Health Questionnaire and the binary classification of depression. The method outperforms cutting-edge algorithms on the DAIC-WOZ database, with high accuracy and low root mean square error. The framework is appropriate for real-time applications because of its speed, non-invasiveness, and non-intrusion. However, real-time applications are challenging and inconvenient.

Du, M., et al., [22] An MSCDR has been suggested, which captures text-independent depressive voice representation from speaker to listener. The model describes voice perception and generation processes using Mel-frequency cepstral coefficients and linear predictive coding characteristics. The study verifies the MSCDR's generalization ability and superiority, implying that vocal tract abnormalities in individuals with depression warrant consideration for audio-based depression diagnosis. However, its use in clinical translation may be limited due to the small sample size.

Marriwala, N., et al., [23] A study presented a hybrid model that employs deep learning techniques to identify depression by combining textual and auditory characteristics. The model was comprised of three parts: an audio CNN model, an LSTM-based hybrid model, and a textual CNN model. Bi-LSTM, a more advanced variant of LSTM, is also employed. The study discovered that

deep learning is more successful at diagnosing depression, with textual CNN and audio Bi-LSTM achieving high accuracy. However, obstacles include dealing with various linguistic expressions, ethical concerns, and potential prejudices.

Huang, Y., et al., [24] A study suggests using a (HATCN) acoustic depression recognition model for interview conversation speech in consultation situations. The model gathers acoustic characteristics from sentences and segments using regional attention approaches, addressing imbalances with a periodic focused loss function. Tests reveal that the model exceeds other strategies in recognition performance. The study also investigates the effect of voice noise on melancholy identification in consultation settings. However, enhancements are required for large-scale applications.

Yin, F., et al., [25] A deep learning model that combines a transformer module for temporal sequential information and a parallel convolutional neural network for local knowledge has been suggested to detect depression using audio. On two datasets, the model outperforms cutting-edge techniques; nonetheless, scalability concerns and challenges in achieving universal effectiveness of acoustic low-level features across varied demographic groups may emerge.

The gaps identified from the Literature Survey are as follows:

- The current public depression databases, AVEC2013, AVEC2014, DAIC-WOZ, and BD, have a limited scope and are unsuitable for scientific research because to ethical problems and the sensitivity of depressive speech.
- The data is not universal since interactive clinical interviews cannot adequately depict depressed individuals' daily lives. Model generalization is difficult due to the few datasets employed in investigations.
- Furthermore, the pathophysiology and behavior of bipolar illness differ from major depressive disorders, and few studies have looked into how speech cues can distinguish between the two.
- Future research trends in depression analysis should include merging multiple modalities to increase study success.

Towards addressing some of the gaps identified the following are the objectives of this work:

- To create a Local Speech Language dataset i.e Marathi Dataset to address the unavailability of the dataset
- To remove the background noise from the raw dataset, resulting in the enhanced usable dataset.
- To create a Novel optimized framework for optimal feature selection from the speech dataset.
- To design a novel unique Attention CNN algorithm for the classification of depressed levels. Hence the major contribution of the work is given as follows:
- Need for creating a local language Marathi dataset of size 54 non-professional speakers.
- A novel framework called KWHO-CNN is proposed, integrating a hybrid Krill Herd and GWO algorithm which results in reduction of 16134 features to 1663 features.
- A unique Attention-Driven Convolution Neural Network implemented for optimal feature selection and classification of speech data, resulting in improved performance metrics of 98% accuracy and 96% precision and recall.

From the above literature review, the existing study assesses an approach to identify depression based on auditory signals, and it recommends fine-tuning datasets with significant variability to lower error rates. The method's accuracy is limited by a small sample size and difficulty in addressing diverse language expressions. It raises ethical concerns and may introduce biases in algorithmic decision-making. To overcome the above issues, there is a need to propose a novel framework using deep learning models with optimized feature selection for processing speech data. The study's overall goal is to improve depression identification using a unique framework that combines optimized deep learning models with hybrid metaheuristic algorithms to increase feature selection, classification accuracy, and data privacy.

3. Proposed Methodology

The proposed approach consists of three steps: creating a Marathi dataset, pre-processing to remove background noise from a voice database, and feature extraction and classification using attention-driven CNN blocks. Figure 1 depicts the proposed framework for the work carried out.

The first block represents the creation of a dataset, for initially the consent form was submitted by all 54 non-professional speakers after that Marathi speech dataset was recorded by using software under the guidance of a well-known psychologist in their clinic. The BDI-IV is used to authenticate speech recordings for depression. Speeches are scored on a scale of mild to severe, with normal scores indicating mild depression, borderline

clinical depression indicating moderate depression, and severe depression indicating severe depression.

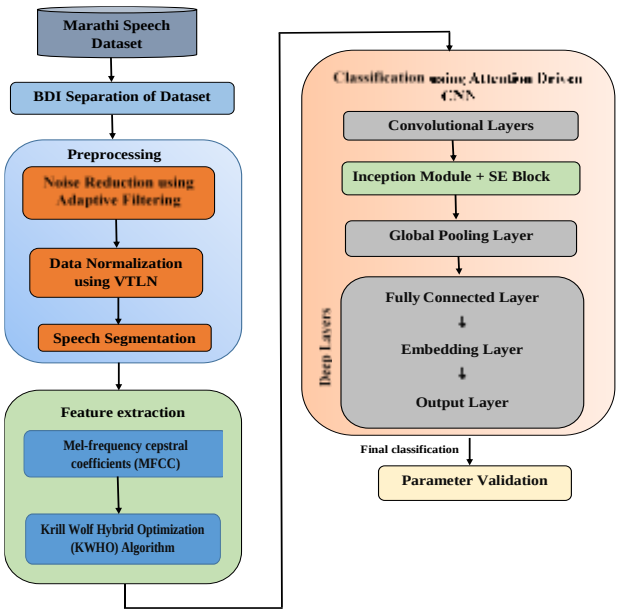

Figure 1. *The Proposed Framework.*

The next block consists of pre-processing removes background noise. The existing study evaluates a depression detection method using auditory signals. It suggests fine-tuning datasets to reduce error rates. Three steps are usually involved in pre-processing methods for speech emotion recognition: noise reduction, data normalization, and speech segmentation. These procedures are used to analyze the various speakers and emotions that are expressed in speech.

The next block is Feature extraction which includes feature extraction as well as feature optimization by using the Krill Herd Wolf Optimization Algorithm. The different method faces challenges like small sample sizes, ethical issues, and potential biases. To overcome the above issues, there is a need to propose a novel framework using deep learning models with optimized feature selection for processing speech data. In feature extraction, the Cepstral descriptors or Mel-frequency cepstral coefficients are extracted which helps to detect different depression levels. After feature extraction, the data augmentation was done on the dataset due small set of real-world Marathi datasets.

The extracted features are of 16134 features, and the Krill Wolf Hybrid Optimization (KWHO) Algorithm elects the relevant features to extract the different depression levels. In particular, there have been a lot of recent studies on the diagnosis of depression using MDL algorithms that allow for the merging of different data sources to yield a variety of information. However, it faces challenges of inherent variability in speech patterns among individuals, which requires robust feature extraction and modeling techniques to capture subtle emotional cues. To overcome these issues, a novel framework of KWHO-CNN: A Hybrid Metaheuristic Algorithm based Optimized Attention-Driven CNN is proposed.

The next block is Attention CNN is used to categorize depression with high accuracy. The Attention-Driven Convolutional Neural Networks are a valid and effective approach for optimal feature selection and classification for speech data. The current research challenge is to increase emotion identification accuracy through the use of low-level descriptors and sentence-level data.

3.1 Dataset Description

The data was obtained from 54 non-professional Marathi speakers, captured with a microphone, and analyzed in the presence of a certified psychologist in Pune, Maharashtra. The data was validated using the BDI, a self-report rating assessment consisting of 21 items that assess depression-related attitudes and symptoms. The BDI-IV, intended for those aged 13 and up, includes items linked to depressive symptoms such as hopelessness, irritation, guilt, and somatic symptoms. To evaluate the degree of depression, test scores are assigned values ranging from 0 to 4 and compared to a key. The normal cut-off scores are as follows.

3.2 Pre-Processing

Pre-processing methods for speech emotion recognition typically involve three key steps: noise reduction, data normalization, and speech segmentation. Noise reduction aims to eliminate background sounds and disturbances, ensuring the clarity and quality of the audio signal. Data normalization adjusts the audio features to a common scale, improving the consistency and accuracy of emotion recognition across different recordings. Speech segmentation divides the continuous audio stream into smaller, meaningful units, such as phonemes or words, facilitating detailed analysis of the emotional content. Together, these steps enable the effective identification and classification of emotions expressed by different speakers in their speech.

3.2.1. Noise Reduction using Adaptive Filtering

3.2.1.1. Signal Model

In this study, the noise reduction task is to recover an SOI $x(n)$ from an observation signal $y(n)$that has been distorted by noise $v(n)$.

$$y(n) = x(n) + v(n) \quad (1)$$

where $v(n)$ is the Gaussian random process that represents the additive noise. It is presumed that the noise $v(n)$does not correlate with the SOI signal $x(n)$.

3.2.1.2 The Adaptive LMS Algorithm

Because of its simplicity and resilience, adaptive signal processing makes extensive use of the least-mean-square (LMS) method. It is renowned for being straightforward and for operating well in fixed contexts [26]. In general, an N-tap filter's weight vector $w(n)$ at time instant n is represented by

$$w(n) = [w_1(n)w_2(n) \dots w_N(n)]^T \quad (2)$$

The input sequence is $\{x(n)\}$and $x(n) = [x(n)x(n-1) \dots x(n-N+1)]^T$to which is a vector representation, comprising the N most recent samples of$\{x(n)\}$. The desired signal $d(n)$ is followed by the filter output $y(n) = w^T(n) \times (n)$, and the estimated error $e(n)$is determined by

$$e(n) = d(n) - y(n) \quad (3)$$

Based on the observed value of $e(n)$, an adaptive filtering algorithm modifies the filter tap weight $w(n)$ at each time instant. The conventional LMS algorithm is modified as [27]

$$w(n+1) = w(n) + \mu e(n)x(n) \quad (4)$$

whereas $\mu\ is$ the step-size parameter that influences the filter weights' convergence behavior.

3.2.1.3 Noise Reduction Algorithm Based on LMS Filter

A noise cancellation approach based on the LMS filtering algorithm for optimal performance is used. Figure. 2 displays the block diagram for the noise reduction technique. Since most audio signals fluctuate over time, segmenting the input signal is necessary for the LMS filtering approach to effectively reduce noise. Every 40ms, the raw noisy signal is divided into segments. Consider the noisy test signal $y = \{y_1 : t = 1,2, \dots, T\}$, where T is the number of frames and is the frame at the time t.

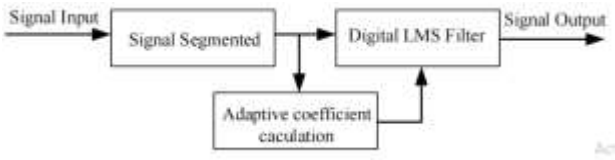

Figure 2. *Block diagram of the noise reduction method*

The noise cancellation technique entails identifying a corresponding weight vector $w_t$for each noisy frame y_t, which may be discriminated more precisely than individual frames. The LMS filter employs the NLMS approach to normalize weight vector updates using the regressor's squared norm. This approach is less sensitive to variations in the adaptive filter's input strength, making it an appropriate choice for applications where the filter input power is highly unpredictable. Normalized weight updates complicate the statistical analysis of the NLMS algorithm. It can be thought of as a subset of the LMS approach, with a step size that varies with input signal strength.

$$w(n+1) = w(n) + \frac{\hat{\mu}}{X^T(n)x(n)}\text{e(n)x(n)} \quad (5)$$

where $\hat{\mu}/X^T(n) \times (n)$is the real step size and $\hat{\mu}$is a factor to be selected. Fig. 2 illustrates that this approach's output signal is

$$z_t(k) = \sum_{k=0}^{N-1} y_t\,(k-n)w(k) \quad (6)$$

In fact, a regularisation parameter ε is needed for a more reliable implementation of the NLMS algorithm. This results in the ε -NLMS algorithm [27], which has an enhanced step size of$\hat{\mu}/(X^T(n) \times (n) + \varepsilon)$. Despite sharing comparable tap-weight adaption equations, LMS-type algorithms are typically not compared to NLMS-type algorithms. It has been demonstrated that algorithms of the NLMS type offer a quicker rate of convergence.

3.2.2 Data Normalization using VTLN

VTLN is a speech signal power spectrum warping function that accounts for differences in vocal tract length between speakers by scaling the frequency axis using a transformation parameter α [28].

$$g_a : [0,\pi] \to [0,\pi]$$
$$\omega \to \hat{\omega}_m(\alpha) = g_a(\omega) \quad (7)$$

Consider ω_m as the centre frequency of filter m in a filter bank of M filters. The warped central frequency of filter m is some $\hat{\omega}_m$ [29],

$$\hat{\omega}_m(\alpha) = \begin{cases} \alpha \cdot \omega_m & \omega_m \leq \omega_0 \\ \alpha \cdot \omega_0 + \frac{\omega_{max} - \alpha \cdot \omega_0}{\omega_{max} - \omega_0}(\omega_m - \omega_0) & \omega_m \geq \omega_0 \end{cases} \quad (8)$$

Where α denotes the warping factor or parameter, $\omega_{max}$is the maximum filter-bank frequency and ω_0 is defined as follows,

$$\omega_0 = \begin{cases} \frac{7}{8}\omega_{max} & \alpha \leq 1 \\ \frac{7}{8\cdot\alpha}\omega_{max} & \alpha > 1 \end{cases} \quad (9)$$

Typically, to implement conventional VTLN, a filter-bank is generated for each warping component αthat needs to be analysed. Function, $g_a\left(\left(z_t(k)\right)_r\right)$, where $g_a\left(z_t(k)\right)$is the piecewise function given in equation (10) and $\left(z_t(k)\right)_r$ is the series of acoustic data.

$$\alpha_{optimal} = arg\max_{\alpha}\left\{\log\left[Pr\left(g_a\left(\left(z_t(k)\right)_r\right)|W_r\right)\right]\right\} \quad (10)$$

where W_r is the word sequence that was identified during the first recognition attempt.

3.2.3 Speech Segmentation using ICA

Independent Component Analysis (ICA) [30] is a technique used in speech segmentation that separates sources in an audio mixture. Specifically, it extracts distinct voices from a conversation among two speakers. When attempting to identify and extract individual sources from observed mixes without prior knowledge of the sources or their mixing process, individual Component Analysis is a potent statistical approach that is employed. When it comes to speech segmentation, ICA works under the premise that the audio signal being observed is a linear mixture of distinct sources, each of which adds unique features to the mixture, such as temporal patterns, pitch, and timbre. To use ICA for speech segmentation, the mixed audio signal that contains the two speakers' dialogue must first be broken down into its sources. Using ICA techniques, a set of statistically independent components is estimated from the observed mixture to achieve this breakdown. These algorithms recognize and distinguish between the sources by making use of their statistical characteristics, such as independence and non-Gaussianity.

The next stage is to determine which independent components belong to each speaker's voice once the independent components have been calculated. Analyzing the individual components' spectral characteristics, temporal patterns, and energy distributions will help accomplish this. The distinct voices of the speakers are represented by components that display characteristics associated with speech signals. Reconstructing the divided speech signals by merging the pertinent independent components comes last, followed by the identification of the independent components that belong to each speaker. Only the elements that accurately represent the speaker's voice are kept after the undesired ones have been filtered away. Depression analysis can then be performed on the separated speech signals.

Thus, identifying distinct voices from conversations involving several speakers and separating sources in an audio mixture may be accomplished effectively by speech segmentation using Independent Component Analysis. Through the use of the statistical characteristics of the sources, ICA makes it possible to separate speech signals without visual aids and makes it easier to extract valuable information from complex audio recordings.

3.3 Feature Extraction

Following pre-processing, the next critical phase in speech emotion recognition is feature extraction, which primarily utilizes MFCCs. MFCCs are frequently used because of their ability to effectively reflect sound's short-term power spectrum, allowing speech nuances and emotions to be recognized. To further enhance these extracted features, the KWHO Algorithm is employed. This algorithm optimizes the MFCC features by fine-tuning them to better capture the emotional content of the speech, thereby overcoming common issues such as overfitting. By refining the feature set, KWHO improves the generalizability and robustness of the emotion recognition model, leading to enhanced performance and more accurate emotion detection.

3.3.1Mel-frequency Cepstral Coefficients

The energies of the cepstrum are described by the MFCC coefficients in the mel-scale, a non-linear scale. They are said to be the most discriminating sound characteristics that mimic the way the spoken signal is perceived by the "human peripheral auditory system." These coefficients' first and second derivatives make it possible to monitor how they change over time and, therefore, how the voice tone changes. For these reasons, the MFCC coefficients are extracted in this suggested work to investigate their resilience in a speech-based automated depression diagnostic application. The following can be used to approximate the connection between the Mel scale and frequency:

$$Mel(f) = 2595 \times \lg(1 + \frac{f}{700}) \quad (11)$$

MFCC is extracted through the following processes: 1) Determine the frequency using the Fast Fourier Transform spectrum

3.3.2 Data Augmentation and Random Sampling

Data augmentation is the process of creating new data from existing data to train machine learning

models. This is especially crucial for Marathi datasets, which require big and diverse samples to begin training. Data augmentation expands the dataset by making minor changes to the original data. To address imbalance class concerns for various depression classes, random oversampling is used, which repeats cases from the minority class in the training dataset.

3.3.3 Krill Wolf Hybrid Optimization (KWHO) Algorithm

The krill herd algorithm [31] is an intelligent swarm optimization method that mimics the characteristics of krill, allowing the group to migrate based on each member's fitness. This method is based on the nutritional features of the krill herd and the mutual local search process of neighboring residents. The world's best reaction is thought to stem from the krill patient's fitness level, as designed with fewer factors.

$$\widehat{K}p.q = \frac{K_p - K_q}{K^{worst} - K^{best}} \quad (12)$$

The greatest and worst Krill individuals' fitness values are represented by $K^{best} and K^{worst}$, the fitness of the p^{th} Krill individual is represented by K_p, the q^{th} neighbor's fitness is represented by K_q.

3.3.2.2 Grey Wolf Optimization Algorithm:

The encirclement strategy is based on the herding and stalking behaviors of grey wolves in the wild [32]. Grey wolves hunt in groups of five to twelve and are classified into four types: α, β, γ, and ω. The α holds more authority over the pack than the β, while the γ leads and makes decisions on behalf of the group. The third-level dominator, γ, aids, guards, and monitors all members, especially weak and old wolves. The final set, ω, includes the remainder of the pack. The hierarchy orders the wolves to circle and hunt, with stalking acts including finding the prey, encircling and intimidating it, and finally attacking the victim. This method can be constructed analytically with equations.

$$\vec{V} = \left|\vec{D} \times \vec{X}_S(l) - \vec{Y}(l)\right| \quad (13)$$

$$\vec{X}(l+1) = \vec{X}_S(l) - \vec{G} \times \vec{V} \quad (14)$$

$$\vec{G} = 2 \times \vec{c} \times \vec{u}_1 - \vec{c} \quad (15)$$

$$\vec{D} = 2 \times \vec{u}_2 \quad (16)$$

$$b = 2 - g \times \frac{2}{L} \quad (17)$$

Where G and D are two regression coefficients, L appears to be the largest number of iterations, and $\overrightarrow{Y_s}(l)$ is the location of the prey at iteration g and $\vec{Y}(l)$ is the position of the wolf at iterations l and (l + 1), respectively. The primary goals of D and G, respectively, are avoiding local minimum stagnation and achieving a balance between supply and demand. Grey wolf optimization can prevent local optima stagnation by randomly varying the amount of D. It can also exploit and explore a specific search space if $|G| < 1$ and $|G| > 1$. Depending on the E and F-level solutions, the H-level solutions should be updated at each iteration.

$$\vec{V}_E = \left|\vec{D}_1 \times \vec{Y}_E - \vec{Y}\right| \quad (18)$$

$$\vec{V}_F = \left|\vec{D}_2 \times \vec{Y}_F - \vec{Y}\right| \quad (19)$$

$$\vec{V}_H = \left|\vec{D}_3 \times \vec{Y}_H - \vec{Y}\right| \quad (20)$$

$$\vec{Y}_1 = \vec{X}_E \times \vec{G}_1 - \vec{V}_E \quad (21)$$

$$\vec{Y}_2 = \vec{X}_F \times \vec{G}_2 - \vec{V}_F \quad (22)$$

$$\vec{Y}_3 = \vec{X}_H \times \vec{G}_3 - \vec{V}_H \quad (23)$$

$$\vec{Y}(l+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (24)$$

The merged equation for the grey wolf and krill herd is as follows:

$$K\,p.q = \frac{K_p - K_q}{K^{worst} - K^{best}} + \vec{Y}(l+1) \quad (25)$$

Where K_p is the fitness of the pth Krill individual, K_q is the fitness of the qth neighbour, and K^{worst} and K^{best} are the fitness ratings of the greatest and worst Krill individuals. Where $\overrightarrow{Y_s}(g)$ prey position at iteration g, wolf position at iteration g and (l + 1) is represented by $\vec{Y}(g)$.

The Krill Wolf Hybrid Optimization (KWHO) Algorithms elect and refine relevant factors that are indicative of depressive states iteratively to extract features from audio data and detect depression. This is how it usually works:

1. Initialization: To start, the algorithm initializes a population of candidate solutions, each of which is a collection of features taken from the audio file.

2. Feature Selection: To choose the most pertinent features from the audio input, KWH combines the GWO and KHO algorithms. To ensure diversity among the characteristics chosen, KHO first explores the search space by dynamically

modifying the feature selection. Subsequently, GWO refines the chosen features according to their fitness, taking advantage of favorable areas within the search space.

3. Fitness Evaluation: The algorithm measures how effectively each feature contributes to the detection of depression in the audio data after choosing a group of features. A classification model trained on labeled audio samples is usually used for this, with the model's performance acting as the fitness metric.

4. Iterative Refinement: By repeating the feature selection, fitness evaluation, and optimization processes, KWH iteratively improves the features that have been chosen. This process keeps on until a predetermined endpoint is reached, like finishing with a sufficient amount of iterations or meeting performance requirements.

5. Final Feature Set: Following optimization, KWH determines a final feature set that best distinguishes between audio samples with and without depression. After that, a classification model for depression detection uses these features as inputs.

6. Model Training and Evaluation: Lastly, the chosen features and labeled audio data are used to train a classification model, such as an Attention Driven CNN.

By leveraging the complementary strengths of KHO and GWO, KWH optimizes the feature extraction process to identify the most relevant features for detecting depression in audio data. This approach aims to improve the accuracy and reliability of depression detection systems, potentially leading to better diagnostic tools and treatment outcomes for individuals suffering from depression.

Algorithm 1: Krill Wolf Hybrid Optimization (KWHO)

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 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 influence, foraging activity, and physical diffusion
            Update krill position based on these movements
        Evaluate the fitness of the updated solutions
        Select top-performing solutions to form the next generation
    if Even iterations;
        update α, β, δ based on new fitness values
    Increment iteration counter
Return the best feature subset found

3.4 Classification using Attention Driven CNN

The study employs Attention-driven CNN to improve depression level categorization by incorporating a global pooling layer, an SE module, and an inception module into the standard convolutional neural network. This improves the network's discriminatory ability and performance in recognizing depression levels. The deep learning network is made up of numerous convolutional layers that are batch normalized, as well as a combination of the SE module and Inception framework as shown in Figure 3.

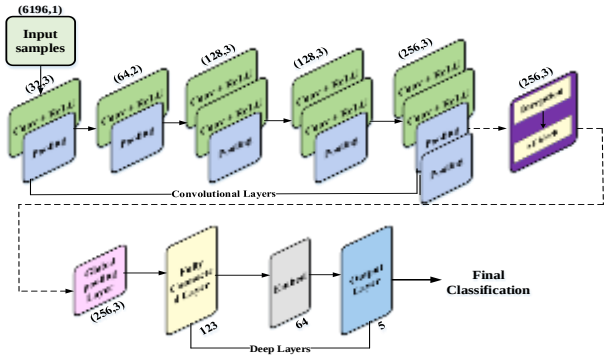

Figure 3. *The Proposed Attention-Driven CNN*

The Attention Driven CNN accelerates training and prevents overfitting by employing convolution filters and batch normalization. It normalizes feature values using a minimum and maximum value. The convolution layers' outputs are combined with an activation function. The max pooling layer filters and sub-samples background noise to reduce input data dimensionality and improve feature fusion. The fixed SE module employs the greatest global pooling layer, recalibrating original features in the channel dimension, lowering training parameters, accelerating model convergence, and improving classification accuracy.

Attention-driven CNNs are well-known for their compositionality and resilience to location fluctuation. They can recognize patterns without being aware of their location, which helps with input data rotation and scaling. These filters convert low-level attributes into higher-level feature representations in deeper layers while preserving compositionality. Hyperparameters such as filter size, stride, and pooling type determine the sliding window distance in a convolution process. The data's feature sets are combined and fed into classifiers to determine the effect of each feature type on the classification process. The original nine-layer architecture has a single input layer, six convolution layers, one output layer, one completely linked layer, and a single output layer.

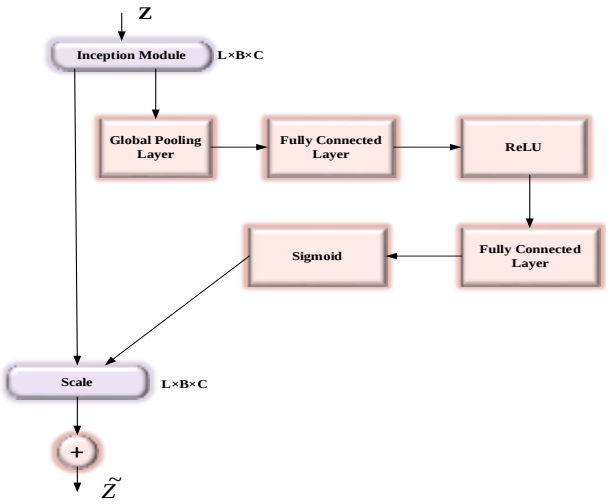

Figure 4. *Combined Framework for SE and Inception Module*

3.4.1. Inception Module

The initiation module is the network's main component, which increases identification performance by collecting data from several fields and incorporating multi-scale information. The network's depth and width are enhanced while remaining sparse, which reduces overfitting and free parameters. The Inception module employs three convolution kernels: a 3x3 max pooling layer, 1x1 convolution, 3x3 convolution, and 5x5 convolution. It captures macroscopic and microscopic features, while the pooling layer safeguards unprocessed input data. The module combines various layers to provide a multi-scale feature map.

3.4.2. Global Pooling Layer

The CNN network has always been configured as a fully integrated system. Unfortunately, the fully linked layer's numerous additional parameters slow the network's training rate and make overfitting easy. The GPL aims to produce an output for each feature map by averaging all of its pixels [33]. The output characteristics' component vector is delivered directly to SoftMax for categorization.

3.4.3. Squeeze-and-Excitation Module

The SE module [34] purges the initial SE module's previous convolutional sequence. The quantity of feature channels, given an input of Z, is C.Three steps are necessary to calibrate previously acquired features, as opposed to the usual CNN. The Squeeze operation is the initial step. Assume the inputs are $Z = (Z_1, Z_2, \dots \dots Z_C)$, $Z_C \in P^{L\times B}$. Formally, reducing Z across its spatial dimensions $L \times B$ results in a statistic $x \in P^C$. The c-th element of X is determined by

$$X_c = G_{sq}(Z_c) = \frac{1}{B\times L}\sum_{i=1}^{B}\sum_{j=1}^{L} Z_c(i,j) \quad (26)$$

The Squeeze operation converts the $L \times B \times C$ input to a $1 \times 1 \times C$ output, which corresponds to the G_{sq}operation. This results in the layer's global data or C feature maps, while the $X_c$output describes local descriptors for the entire channel map.

To regulate the model's complexity and generalization, the embedding process is governed by two nonlinear linked layers.

$$S_{act} = G_{ex}(X, B) = \sigma\big(g(X, B)\big) = \sigma(B_2\delta(B_1 * X)) \quad (27)$$

When δ denotes it is possible to think of the ReLu function and the output X as a collection of local attributes for the whole channel map, $B_1 \in P^{\frac{C}{p}\times C}$&$B_2 \in P^{C\times\frac{C}{p}}$. Two nonlinear layers govern the embedding process's complexity and generalization. A reweight procedure weights prior features in the channel dimension, establishing the relevance of each feature channel using the weight of the Excitation output. When rescaling Z with the activations s, the block's output is obtained:

$$\bar{z}_c = G_{scale}(z_C, s_C) = z_C . s_C \quad (28)$$

where can where$G_{scale}(z_C, s_C)$ denotes channel-wise multiplication of the scalar s_C and the feature map $z_C \in P^{L\times B}$&$\bar{z} = |\bar{z}_1, \bar{z}_2, \dots \dots, \bar{z}_C|$.

The SE module may be integrated into both the inception and conventional network architectures, as seen in Figure 4. In Figure. 4, the SE module and the Inception module are combined.

Algorithm 2: A Novel Attention Driven CNNAlgorithm

Input: Accept input data of size (6196,1).

Convolutional Layers with SE and Inception:

Step 1: Convolutional layer with 32 to 256 filters of size (3x3), using ReLU activation.

SE module:

Step 2: Apply squeeze-and-excitation operation to the output of the previous convolutional layer.

Inception module:

Step 3: Apply the Inception structure to the output of the previous SE module.

Step 4: Global Max Pooling layer with a pool size of (256, 3).

Fully Connected Layers:

Step 5: Dense layer (123) with neurons, using ReLU activation

Step 6: Denselayer (64) with neurons, using ReLU activation.

Step 7: Dense layer (5) with the desired number of output neurons, using softmax activation.

Output: The final output is a probability distribution over the classes in the target dataset.

4. System Implementation

4.1 Database

The Marathi speech audio file of 54 individuals is used while implementing the algorithm. These 54 databases are of engineering different branch students. Before recording their audio file the consent form is submitted by each student. All the datasets are labeled according to the BDI tool score and stored in the different folders according to their labeled class for Normal, Mood disturbance, slightly depressed, moderately depressed, and Depressed in folders 0, 1,2,3,4 respectively.

4.2. Pre-Processing

4.2.1 Noise Reduction

Noise reduction is a critical step in speech data preparation that improves audio signal quality by reducing background noise. Adaptive filtering approaches, such as the LMS and RLS algorithms, are used to update filter coefficients in real time based on input signal characteristics, successfully dealing with non-stationary noise.

4.2.2 Data Normalization and Segmentation

VTLN is used to normalize the data and account for speaker-dependent changes in vocal tract length. Independent Component Analysis (ICA) separates components in an audio mixture based on statistical independence. The speech is then separated into 7-second segments with no overlap, resulting in a unified speech with variable lengths and an increased number of training examples.

4.3 Feature Extraction

4.3.1 Feature extraction

Pre-processing involves extracting feature sets and statistical metrics from audio recordings. MFCCs are commonly employed in speech-processing applications. The MFCC extraction method extracts spectrum properties such as MFCC, SDCC, spectral centroid, roll-off, flatness, contrast, bandwidth, chroma-soft, zero crossing rate, root mean square energy, and LPCC from audio signals. These coefficients capture the spectral properties of the voice stream and are calculated by dividing it into small frames and using a filter bank.

4.3.2. Krill-wolf herd Optimization Algorithm

The KWHO Algorithm optimizes the MFCC extraction of voice signals. This algorithm iteratively explores the solution space for the subset of features that maximizes the specified criteria. Multi-agent systems are used to select characteristics according to various patterns. The search algorithms are modified to broaden the search space without being limited by local optimums. When selecting traits, krill's swarming behavior, as well as grey wolves' social hierarchy and hunting behavior, are considered. This iterative updating increases prediction model performance while minimizing overfitting issues. This leads to reduced complexity, faster training, and a robust training model.

4.4 Classification using Attention Driven CNN

The Attention Driven CNN is proposed as a method for accurately categorizing patients based on extracted data. The CNN employs a convolutional layer to extract features, followed by a subsampling layer to retrieve important information while reducing spatial resolution. A structure consisting of an inception module, a squeeze and excitation module, and a global pooling layer is used to determine if a patient has depression or not. Filtering allows the CNN to approximate the ideal sparse structure using dense calculations and a ReLU. A fully connected layer classifies features by utilizing posterior probabilities to train kernels via back-propagation. The output layer is preceded by an embedding layer, which converts each signal into a fixed-length vector of a specific size. The resulting vector is dense and has real values ranging from 0 to 4, allowing for better signal labeling and classification into five classifications based on the Beck Depression Inventory. The model

outperforms traditional convolutional neural networks.

4. Result and Discussion

The result part discusses the usefulness of the suggested strategy and the outcomes of its implementation. The results of the comparison of the suggested work with existing approaches are also reported. The suggested deep learning-based method to a benchmark filter is evaluated using simulations, data pre-processing, and Python 3 libraries (such as NumPy, pandas, seaborn, and Sk learn packages). The model is created with TensorFlow 2.10 and the Keras Python library. A suitable emotional speech database is an important requirement for any emotional recognition model. The quality of the database determines the efficiency of the system.

5.1 Dataset Description

The data was gathered from 54 non-professional Marathi speakers and verified using the BDI, a 21-item self-report rating assessment that assesses depression-related attitudes and symptoms, making it a popular psychometric measure of depression severity.

This dataset contains 40 datasets of males and 14 datasets of females. This complete dataset is divided into 0-4 classes based on the BDI tool in different class folders given below.

Class BDI Score Result

Class 0- 1-10_____These ups and downs are considered normal

Class 1- 11-16_____Mild mood disturbance

Class 2-17-20_____Borderline clinical depression

Class 3-21-30_____ Moderate depression

Class 4-31-40_____Severe depression

The pre-processing and feature extraction were performed on this database according to class 0-4, so all the databases were labeled according to class 0-4 as per their BDI score. This dataset is separated into training, validation, and test datasets: By dividing it in half, 90/10, the full dataset was separated into a training and a test dataset.

5.2. Pre- Processing

Pre-processing includes different processes like noise reduction, data normalization, speech segmentation, feature extraction, and optimization, as discussed below.

5.2.1 1Noise Reduction using Adaptive Filtering

Raw speech comprises internal noise caught during collection, such as the interviewer's voice and mute clips, which is unrelated to depression and hence affects recognition performance. As seen in Fig. 5, we used the adaptive filtering approach to eliminate the noise portion.

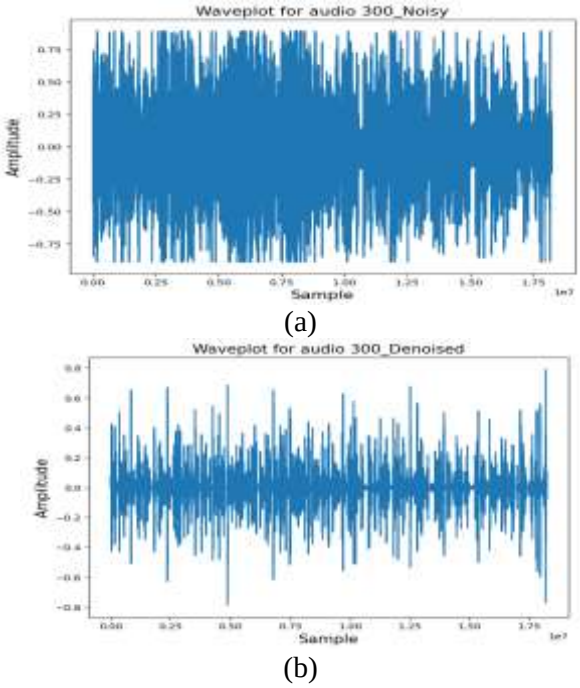

Figure 5. *Preprocessing results of one subject with (a) raw speech with noise, (b) speech without internal noise*

5.3 Feature Extraction

Machine learning algorithms identifying conditions such as depression depend on the properties found in the data. To train their models, we can take meaningful features out of the raw data. By using "feature extraction," complex data sets can be reduced to more manageable inputs. To identify depression, audio data must be analyzed for significant elements that reveal information about a person's feelings and speech patterns. A person may have depression if they have certain vocal characteristics. A total of 16556 features are extracted from all 54 speech datasets.

5.4 Data Augmentation and Random Oversampling

The study employs 54 non-professional speakers' voice samples to investigate data augmentation, a technique that artificially enhances the training set by producing modified copies of the dataset. The study employs random oversampling to balance the dataset, making minor tweaks or creating new data

points using deep learning. The results reveal that all classes 0-4 have 219 samples, making this method suitable for tiny datasets. The Random Over Sampler software was used to carry out this operation, yielding a total of 219 samples for all classes.

5.5 Krill Wolf Hybrid Optimization (KWHO) Result

Fitness achievement through the KWHO method is strengthened by its efficient population-based optimization approach. By harnessing the collective intelligence of krill and wolf populations, this methodology exhibits superior performance compared to various optimization methods. Through simulations, it has been demonstrated that the KWHO method consistently outperforms alternatives, yielding optimal solutions of higher quality across numerous iterations. Notably, even after a hundred iterations, the KWHO method maintains its ability to generate significantly improved optimal solutions, achieving a fitness function point as low as 0.0050. This highlights the method's robustness and its capacity to continuously refine solutions towards greater fitness. One of the key advantages of KWHO lies in its ability to deliver such impressive results with minimal parameterization. This means that users can effectively apply the methodology without the need for extensive fine-tuning, thereby streamlining the optimization process and enhancing overall efficiency. Consequently, the adoption of the KWHO method not only facilitates the attainment of superior fitness levels but also optimizes resource utilization, making it a compelling choice for a wide range of optimization tasks, as illustrated in Fig 6.

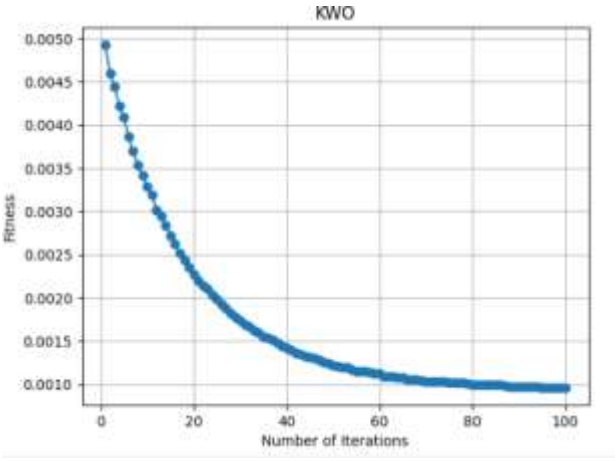

Figure 6. *Krill Wolf Hybrid Optimization*

Finally, the finding reveals that by using the KWHO algorithm the 16565 features are optimized to 6196 features.

5.6 Attention-driven CNN

The Attention Driven CNN is used in the study to improve disease detection by dealing with long training convergence durations and an abundance of model parameters. To improve disease recognition, an inception module, global pooling layer, and SE module were added to the standard convolutional neural network. The input data was optimized using a convolutional layer with 32 to 256 filters, a squeeze-and-excitation operation, the Inception structure, a Global Max Pooling layer, Fully Connected Layers (dense layers with neurons), and a Dense layer (5) with the required number of output neurons. The final result is a probability distribution over the classes in the target dataset.

5.7 Performance Evaluation:

This section delves into the performance evaluation of the proposed KWHO-CNN, integrating a hybrid metaheuristic algorithm with Attention-Driven CNNs in distinguishing between normal and depressed mental health states. It charts the accuracy and loss of the model across 50 training epochs, providing insights into its convergence speed. It also discusses the potential trade-offs between model accuracy, computational efficiency, and generalizability, hinting at a point where increased accuracy might come at the cost of other factors. Overall, the section emphasizes the efficacy of the framework in accurately discerning between normal and depressed mental health conditions.

5.7.1 Performance Parameters:

The report gives simulation results for the suggested deep learning approach, emphasizing accuracy, precision, and recall performance criteria to demonstrate the study's efficacy. Fig. 7 below showcases the performance parameters of the proposed method.

a) Accuracy

Accuracy is used to assess how well a categorization model works.

$$ACC = \frac{TP+TN}{T\,P\,+T\,N\,+F\,P\,+F\,N} \qquad (29)$$

"true positives," "false positives," "true negatives," as well as "false negatives," accordingly.

b) Precision

Precision is defined as the proportion of confirmed positive cases to all expected positive patterns. The following formula can be used to calculate it.

$$Precision = \frac{TP}{T\,P+F\,P} \quad (30)$$

c) Recall

A classifier's recall is determined by the percentage of true positives it receives, demonstrating its ability to recognize positive class patterns that can be sensitive or exact but not susceptible.

$$Recall = \frac{TP}{TP+FN} \quad (31)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (32)$$

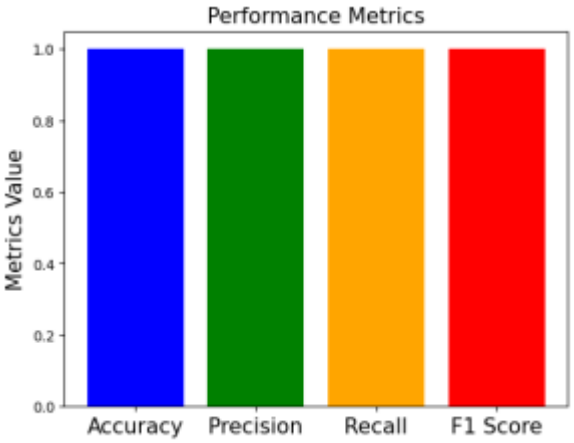

Figure 7. *Performance Parameters of the KWHO-CNN Model*

The confusion metrics about TP, FP, TN, and FN have a significant impact on automatic clinical depression recognition by the proposed system. The number of speech data that correctly classified the depression state as normal, mild, BL, moderate, and severe is also determined by the confusion measures. The confusion metrics of the proposed model are shown in Figure 8. Fig 8 shows how successfully the proposed models classify the depression level types. There are five classes in total, ranging from class 0 to class 4, and each class is represented by one of the five types of depression levels that are currently affected in the given speech data, which include normal, mild, BL, moderate, and severe. As a result, the TP value is predicted for 93% of data from the total data volume in class 0, 99% of data from the total data volume in class 1, 96% of data from the total data volume in class 2, 97% of data from total data volume in class 3 and 96% data from total data volume in class 4 respectively. The greatest predicted value (TP) for the type of depression level for the true labels normal, mild, BL, moderate, and severe is 99% for class 1, respectively.

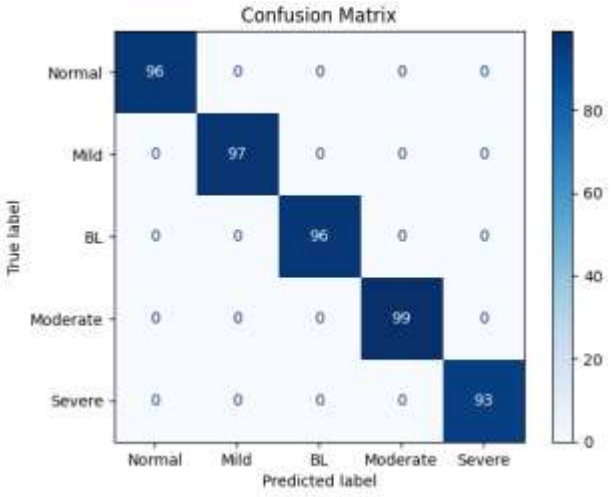

Figure 8. *Confusion Matrix for the KWHO-CNN Model*

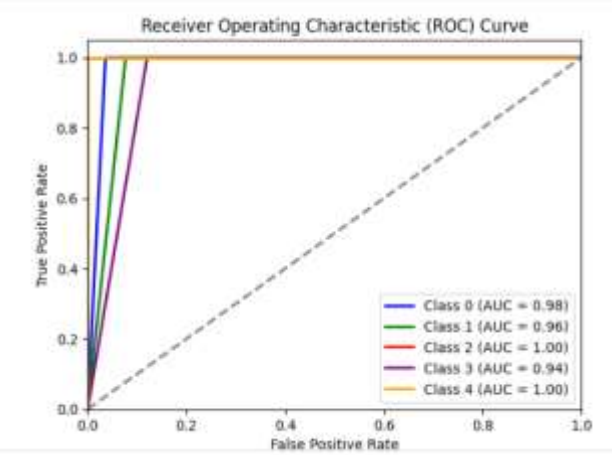

Figure 9. *Receiving Operating Characteristic (ROC) for class 0-4*

ROC analysis is a tool for evaluating the performance of diagnostic tests and statistical models. It defines a patient's depression as positive or negative based on test findings, to determine the ideal cut-off value for the greatest diagnostic performance. The ROC curve has five classifications that represent different levels of accuracy as shown in Figure 9.

5.4 Model Training and Validation Result

Training accuracy is a model's ability to recognize irrelevant samples. Overfitting is easily detected by comparing losses or accuracy over epochs. Figure. 9 displays thatover 50 epochs, the model achieved a 0.22 accuracy rate by using a hybrid Krill Herd and Grey Wolf optimization method with Attention-Driven Convolutional Neural Networks. This model saves time and effort by offering good

performance without requiring a full hyperparameter search, hence improving feature selection and categorization of speech data.

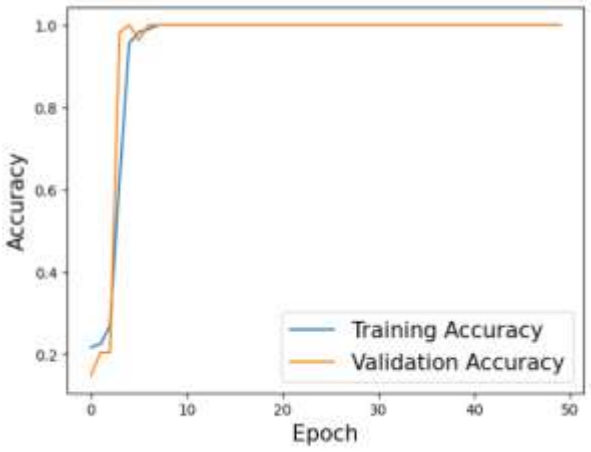

Figure 10. *Training and Validation Accuracy of the KWHO-CNN Model*

The suggested deep learning model, which employs a hybrid Krill Herd and Grey Wolf optimization method, and Attention-Driven Convolutional Neural Networks, yielded less than 1.60 train loss over 50 epochs as shown in Figure 10. This is because min-max normalization is used to minimize the dimensionality of the input signal, and the optimization process is capable of adjusting the learning rate for each parameter, successfully updating model parameters, and maintaining convergence and stability.

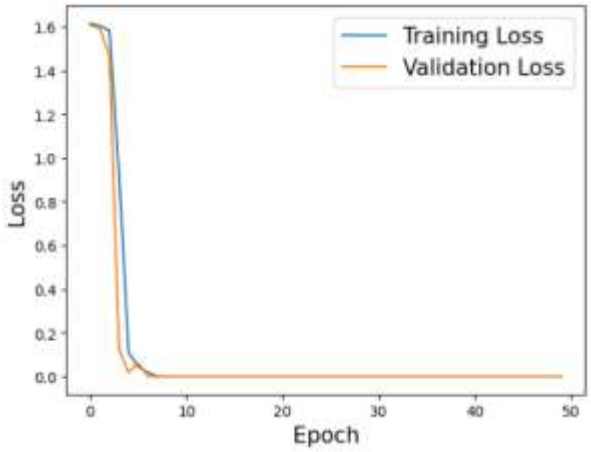

Figure 11. *Training and Validation Loss of the KWHO-CNN Model*

5.8 Comparison Result

The suggested methodology is tested and compared to five existing classifiers, including CNN, Parallel CNN, HATCN, and DNN, utilizing metrics such as Accuracy, Precision, and Recall. The approach is compared to current classifiers using CNN, Parallel CNN, HATCN, and DNN, with Table 1 providing an overview of the proposed procedures vs current classifiers.

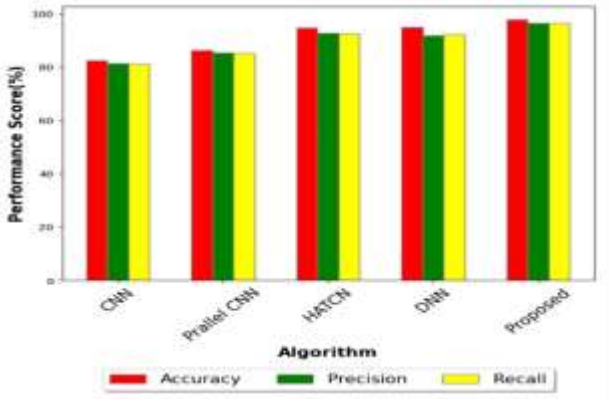

Figure 12. *Comparison of Proposed Methodology with Existing Methods*

The suggested method outperforms other classifiers, including CNN, Parallel CNN, HATCN, and DNN, in terms of accuracy, precision, and recall. It combines MFCC characteristics with the hybrid Krill Herd and Grey Wolf optimization algorithms, yielding high values of 98%, 96%, and 96%, respectively. Figure 12 presents a comparative comparison.

Table 1. *Comparison of Proposed Methodology*

Methods	Accuracy (%)	Precision (%)	Recall (%)
CNN	83	81	81
Parallel CNN	86	85	85
HATCH	95	94	94
DNN	95	92.5	93
Proposed	98	96	96

Figure11 and Table 1 shows the comparison of extisting Methodology with the proposed methdology providing an overview of the proposed procedures vs current classifiers. Figure 11 and The table1 compares the performance of various methods—CNN, Parallel CNN, HATCH, DNN, and a Proposed Method—on an unspecified task using accuracy, precision, and recall as metrics. The CNN [21] method achieves an accuracy of 83%, with both precision and recall at 81%. The Parallel CNN [22] improves upon this, reaching 86% accuracy and 85% for both precision and recall. The HATCH [23] method shows significant improvement, boasting 95% accuracy, 94% precision, and 94% recall. Similarly, the DNN [24] method maintains high performance with 95% accuracy, 92.5% precision, and 93% recall. The Proposed Method, however, surpasses all others, achieving the highest metrics across the board with

98% accuracy, 96% precision, and 96% recall. This indicates a substantial enhancement in performance, suggesting that the Proposed Method is more effective and reliable than the other techniques evaluated.

6. Conclusion

In conclusion, depression presents a significant challenge in clinical diagnosis due to its varied symptoms. Audio-based diagnosis offers promise for early mass screening, yet current methods often overlook vital vocal tract modifications, limiting their effectiveness. This study introduces a novel framework, KWHO-CNN, which integrates a hybrid metaheuristic algorithm with Attention-Driven CNN, to enhance depression detection using speech data. By addressing issues such as speech pattern variability and small sample sizes through optimized feature selection and classification, the proposed framework demonstrates superior performance in depression diagnosis. Initial preprocessing steps, including noise reduction and data normalization, are followed by feature extraction using MFCCs. The KWHO Algorithm optimizes these features from 16565 features to 6196 optimized features, mitigating overfitting and enhancing model performance. The Attention-Driven CNN architecture further refines classification, leveraging dense computations and architectural homogeneity. The proposed system outperforms accuracy, precision, and recall, achieved by integrating MFCC features with the KWHO algorithm, achieving 98% accuracy and 96% precision and recall. The findings reveal that the proposed model achieves impressive overall accuracy, precision, recall, and F1-score, exceeding 90% with the proposed method. Future work could explore the integration of additional data modalities to enhance the model's diagnostic capabilities. Additionally, the FL could also be incorporated to address privacy concerns and enhancing model robustness by training on distributed data without compromising individual privacy.

Author Statements:

- **Ethical approval:** Institutional Review Board approval was not required.
- **Conflict of interest:** They author declare that have no conflict of interest
- **Acknowledgement:** The authors would like to thank the Deanship of Dr. D. Y. Patil Institute of Technology for supporting this work.
- **Author contributions:** The authors confirm contribution to the paper as follows and all authors reviewed the results and approved the final version of the manuscript.
- **Funding information:** The authors state that this work has not received any funding.
- **Data availability statement:** No data, models, or code were generated or used during the study

References

[1]. Hammar, Å., Ronold, E.H., &Rekkedal, G.Å. (2022). Cognitive impairment and neurocognitive profiles in major depression—a clinical perspective. *Frontiers in Psychiatry*. 13: 764374.

[2]. Yang, W., Liu, J., Cao, P., Zhu, R., Wang, Y., Liu, J. K., & Zhang, X. (2023). Attention guided learnable time-domain filterbanks for speech depression detection. *Neural Networks*.

[3]. Bachmann, S. (2018). Epidemiology of suicide and the psychiatric perspective. *International journal of environmental research and public health*. 15(7): 1425.

[4]. Vázquez-Romero, A., & Gallardo-Antolín, A. (2020). Automatic detection of depression in speech using ensemble convolutional neural networks. *Entropy*. 22(6): 688.

[5]. Altwaijri, Y. A., Al- Subaie, A. S., Al- Habeeb, A., Bilal, L., Al- Desouki, M., Aradati, M., & Kessler, R. C. (2020). Lifetime prevalence and age- of- onset distributions of mental disorders in the Saudi National Mental Health Survey. *International journal of methods in psychiatric research*. 29(3): e1836.

[6]. Vitale, F., Carbonaro, B., Cordasco, G., Esposito, A., Marrone, S., Raimo, G., & Verde, L. (2021). A Privacy-Oriented Approach for Depression Signs Detection Based on Speech Analysis. *Electronics*. 10(23): 2986.

[7]. Esposito, A., Callejas, Z., Hemmje, M. L., Fuchs, M., Maldonato, M. N., &Cordasco, G. (2021). Intelligent Advanced User Interfaces for Monitoring Mental Health Wellbeing. In *Advanced Visual Interfaces. Supporting Artificial Intelligence and Big Data Applications: AVI 2020 Workshops, AVI-BDA and ITAVIS, Ischia, Italy, June 9, 2020 and September 29, 2020, Revised Selected Papers*. 83-95.

[8]. Aloshban, N., Esposito, A., &Vinciarelli, A. (2020). Detecting depression in less than 10 seconds: Impact of speaking time on depression detection sensitivity. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. 79-87.

[9]. Tao, F., Esposito, A., &Vinciarelli, A. (2020). Spotting the Traces of Depression in Read Speech: An Approach Based on Computational Paralinguistics and Social Signal Processing. In *INTERSPEECH*. 1828-1832.

[10]. Esposito, A., Raimo, G., Maldonato, M., Vogel, C., Conson, M., & Cordasco, G. (2020). Behavioral sentiment analysis of depressive states. In *2020 11th IEEE International Conference on Cognitive*

Infocommunications (CogInfoCom). 000209-000214.

[11]. Jo, A.H., &Kwak, K.C. (2022). Diagnosis of Depression Based on Four-Stream Model of Bi-LSTM and CNN From Audio and Text Information. *IEEE Access*. 10: 134113-134135.

[12]. Cai, C., Niu, M., Liu, B., Tao, J., & Liu, X. (2021). TDCA-Net: Time-Domain Channel Attention Network for Depression Detection. In *Interspeech*. 2511-2515.

[13]. Nadeem, A., Naveed, M., Islam Satti, M., Afzal, H., Ahmad, T., & Kim, K.I. (2022). Depression detection based on hybrid deep learning SSCL framework using self-attention mechanism: An application to social networking data. *Sensors*. 22(24): 9775.

[14]. Guo, T., Zhao, W., Alrashoud, M., Tolba, A., Firmin, S., & Xia, F. (2022). Multimodal educational data fusion for students' mental health detection. *IEEE Access*. 10: 70370-70382.

[15]. Park, J., & Moon, N. (2022). Design and implementation of attention depression detection model based on multimodal analysis. *Sustainability*. 14(6): 3569.

[16]. Prabhudesai, S., Mhaske, A., Parmar, M., & Bhagwat, S. (2021). Depression Detection and Analysis Using Deep Learning: Study and Comparative Analysis. In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*. 570-574.

[17]. Huang, Z., Epps, J., Joachim, D., &Sethu, V. (2019). Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection. *IEEE Journal of Selected Topics in Signal Processing*. 14(2): 435-448.

[18]. Yalamanchili, B., Kota, N.S., Abbaraju, M.S., Nadella, V.S.S., & Alluri, S.V. (2020). Real-time acoustic based depression detection using machine learning techniques. *In 2020 International conference on emerging trends in information technology and engineering (ic-ETITE)*. 1-6.

[19]. Wu, P., Wang, R., Lin, H., Zhang, F., Tu, J., & Sun, M. (2023). Automatic depression recognition by intelligent speech signal processing: A systematic survey. *CAAI Transactions on Intelligence Technology*. 8(3): 701-711.

[20]. Miao, X., et al. (2022). Fusing features of speech for depression classification based on higher-order spectral analysis. *in Speech Communication.* 143(1): 46–56.

[21]. Rejaibi, E., Komaty, A., Meriaudeau, F., Agrebi, S., &Othmani, A. (2022). MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*. 71: 103107.

[22]. Du, M., Liu, S., Wang, T., Zhang, W., Ke, Y., Chen, L., & Ming, D. (2023). Depression recognition using a proposed speech chain model fusing speech production and perception features. *Journal of Affective Disorders*. 323: 299-308.

[23]. Marriwala, N., & Chaudhary, D. (2023). A hybrid model for depression detection using deep learning. *Measurement: Sensors*. 25: 100587.

[24]. Huang, Y., Ma, Y., Xiao, J., Liu, W., & Zhang, G. (2023). Identification of depression state based on multi- scale acoustic features in interrogation environment. *IET Signal Processing*. 17(4): e12207.

[25]. Yin, F., Du, J., Xu, X., & Zhao, L. (2023). Depression Detection in Speech Using Transformer and Parallel Convolutional Neural Networks. *Electronics*. 12(2): 328.

[26]. Alouane, M.T.H., & Jaı, M. (2006). A new nonstationary LMS algorithm for tracking Markovian time varying systems. *Signal processing.* 86(1): 50-70.

[27]. Haykin, S. (2001). Minimum mean square error adaptive filter. Adaptive Filter Theory, 4th ed. *Prentice Hall, Upper Saddle River.* 183-228.

[28]. Pitz, M., & Ney, H. (2005). Vocal tract normalization equals linear transformation in cepstral space. *IEEE Transactions on Speech and Audio Processing*. 13(5): 930-944.

[29]. Lee, L., & Rose, R. (1998). A frequency warping approach to speaker normalization. *IEEE Transactions on speech and audio processing.* 6(1): 49-60.

[30]. Tharwat, A. (2021). Independent component analysis: An introduction. *Applied Computing and Informatics.* 17(2): 222-249.

[31]. Gandomi, A.H., &Alavi, A.H. (2012). Krill herd: a new bio-inspired optimization algorithm. *Communications in nonlinear science and numerical simulation*. 17(12): 4831-4845.

[32]. Mirjalili, S., Mirjalili, S.M., & Lewis, A. (2014). Grey wolf optimizer. *Advances in engineering software.* 69: 46-61.

[33]. Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

[34]. Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition.* 7132-7141.