Machine Learning with Applications 22 (2025) 100748

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Evolutionary AdaBoost ensemble: A machine learning framework for depression detection

Ruhollah Sayeri [a], Behnam Barzegar [a,*], Yaser Bozorgi rad [a], Nasser Mikaeilvand [a], Mohammad Hassan Tayarani Najaran [b,*]

[a] Department of Computer Engineering, Bab.C., Islamic Azad University, Babol, Iran
[b] University of Hertfordshire, Hatfield, UK

ARTICLE INFO

Keywords:
Depression detection
Convolutional neural networks
Signal processing
Affective computing

ABSTRACT

Depression is a prevalent and debilitating mental health disorder that often goes undiagnosed due to the lack of accessible, objective screening tools. This paper introduces EVAdaBoost, an Evolutionary AdaBoost ensemble framework designed for automated depression detection from voice signals. The method leverages a diverse set of signal processing techniques—including Fourier, Wavelet, Walsh, Hilbert–Huang, and OpenSmile, as well as time–frequency transformations for convolutional neural networks (CNNs). Each feature set is used to train a specialised AdaBoost ensemble, with Broad Learning Systems (BLS) serving as efficient weak learners. A key innovation of EVAdaBoost is its use of a quantum-inspired evolutionary algorithm to optimise the feature subsets assigned to each AdaBoost model. Instead of using all extracted features, which may include noise, redundancy, and irrelevant data, EVAdaBoost evolves to select diverse and high-performing subsets of features for each AdaBoost base learner, automatically discarding non-informative features. This evolutionary selection enhances both classification accuracy and computational efficiency. Additionally, an evolutionary pruning algorithm is employed to find the optimal subset of AdaBoost algorithms that offer the best performance at reduced computational cost. Experiments across nine feature types and multiple benchmark classifiers show that EVAdaBoost consistently outperforms state-of-the-art methods in accuracy, sensitivity (TPR), specificity (TNR), and precision (PPV). The results underscore the potential of hybrid evolutionary ensemble learning for non-invasive, speech-based mental health screening.

1. Introduction

Depression and other common mental health disorders, such as anxiety, affect millions of people across all age groups globally (Reece et al., 2017), posing severe risks not only to mental well-being but also to physical health, including increased susceptibility to cardiovascular diseases, diabetes, and even mortality due to suicide (Almas et al., 2015). According to the World Health Organisation, over 280 million people suffer from depression worldwide, with nearly 800,000 suicide cases attributed to it annually (World Health Organization, 2023). Despite the significant burden, both personal and societal, mental health disorders often remain undiagnosed or untreated due to the stigma surrounding them, limited access to early diagnostic services, and the high cost and inefficiency of traditional diagnostic methods such as clinical interviews and psychological assessments (Cacheda et al., 2019). Individuals with depression exhibit diverse symptoms, including sleep disturbances, mood fluctuations, reduced cognitive function, altered speech and facial patterns, and hormonal imbalances, which may vary by the type and severity of the disorder (Yates et al., 2017). These variations complicate the diagnosis process and hinder timely intervention. With the rapid advancement of artificial intelligence and machine learning, new possibilities are emerging for automatic, scalable, and non-invasive detection of depression through analysis of multimodal data such as facial expressions, voice signals, text, and physiological indicators like EEG (Yazdavar et al., 2017). Emotion recognition, especially when extended to digital behaviour such as social media text, can serve as a valuable tool in identifying emotional

* Corresponding authors.
E-mail addresses: ruhollah.sayeri@iau.ac.ir (R. Sayeri), behnam.barzegar@iau.ac.ir (B. Barzegar), yaser.bozorgi@iau.ac.ir (Y.B. rad), nasser.mikaeilvand@iau.ac.ir (N. Mikaeilvand), m.tayaraninajaran@herts.ac.uk (M.H.T. Najaran).

https://doi.org/10.1016/j.mlwa.2025.100748
Received 11 July 2025; Received in revised form 29 September 2025; Accepted 3 October 2025
Available online 6 October 2025

imbalance, detecting suicidal ideation, and ultimately facilitating early intervention and destigmatisation (Hasin et al., 2018).

Given the increasing global concern about depression, researchers and developers are turning to machine learning (ML) as a powerful tool to enhance early detection and intervention strategies (Jiang et al., 2017). ML have been used for analysing complex patterns in multimodal data, to detect depression (Guntuku et al., 2017). For instance, emotion recognition systems utilise facial micro-expressions captured via video to detect sadness, lethargy, or disengagement, hallmarks of depressive states (Deshpande & Rao, 2017). Similarly, sentiment analysis and natural language processing (NLP) techniques can mine textual responses from chatbots, interviews, or social media posts to identify linguistic markers of depression, such as negative sentiment, reduced affect, or cognitive distortions (Alghamdi et al., 2020). Deep learning models like CNNs and Bi-LSTMs have demonstrated strong performance in distinguishing depressed from non-depressed individuals by analysing the syntax, semantics, and temporal patterns of user-generated content (Kour & Gupta, 2022). Some systems even integrate these features with audio inputs, using prosodic cues like speech pauses, tone, and rhythm to detect vocal indicators of psychological distress (Low et al., 2020). Moreover, large-scale social media platforms like Twitter and Reddit offer rich, real-world behavioural data that, when processed through ML classifiers such as SVMs, Random Forests, and 1D-CNNs, allow for scalable depression screening (Tadesse et al., 2019). However, despite promising results, with some models achieving over 90% accuracy, challenges remain in ensuring robustness, reducing false positives, and addressing ethical concerns related to privacy and misclassification (Islam et al., 2024). Nevertheless, the integration of ML into depression detection systems offers a transformative approach to bridging diagnostic gaps and personalising mental health care.

Depression and other common mental health disorders, such as anxiety, affect millions of people across all age groups globally (Reece et al., 2017), posing severe risks not only to mental well-being but also to physical health, including increased susceptibility to cardiovascular diseases, diabetes, and even mortality due to suicide (Almas et al., 2015). According to the World Health Organisation, over 280 million people suffer from depression worldwide, with nearly 800,000 suicide cases attributed to it annually (World Health Organization, 2023). Despite the significant personal and societal burden, mental health disorders often remain undiagnosed or untreated due to stigma, limited access to early diagnostic services, and the inefficiency of traditional methods such as clinical interviews and psychological assessments (Cacheda et al., 2019). Individuals with depression exhibit diverse symptoms, including sleep disturbances, mood fluctuations, reduced cognitive function, altered speech and facial patterns, and hormonal imbalances, which vary across types and severities of the disorder (Yates et al., 2017). These variations complicate diagnosis and delay timely intervention. With the rapid advancement of artificial intelligence (AI) and machine learning (ML), new opportunities have emerged for automatic, scalable, and non-invasive approaches to depression detection (Yazdavar et al., 2017).

Machine Learning algorithms have become powerful tools for analysing complex behavioural and physiological patterns linked to depression (Jiang et al., 2017). For instance, emotion recognition systems leverage facial micro-expressions captured via video to detect sadness, lethargy, or disengagement, which are representative of depressive states (Deshpande & Rao, 2017). Similarly, sentiment analysis and natural language processing (NLP) techniques can extract linguistic markers of depression from textual responses in chatbots, interviews, or social media posts, including negative sentiment, reduced affect, and cognitive distortions (Alghamdi et al., 2020). Deep learning models, such as CNNs and Bi-LSTMs, have demonstrated strong performance in detecting individuals with depression by analysing the syntax, semantics, and temporal dynamics of user-generated content (Kour & Gupta, 2022). Other systems combine these with audio features, using prosodic cues like pauses, tone, and rhythm to detect vocal indicators of psychological distress (Low et al., 2020). Large-scale platforms such as Twitter and Reddit also provide rich, real-world behavioural data, which can be processed through classifiers like SVMs, Random Forests, and 1D-CNNs to enable scalable depression screening (Tadesse et al., 2019).

Despite encouraging results, challenges remain in improving robustness, reducing false positives, and addressing ethical concerns such as privacy and misclassification (Islam et al., 2024). Most existing studies focus on single modalities or rely on standard feature extraction and classification methods, often neglecting the role of systematic feature selection and ensemble diversity. Research specifically targeting feature subset optimisation for depression detection remains very limited, with only a handful of studies exploring it in depth. This creates a critical gap for frameworks that can integrate heterogeneous feature types while dynamically selecting and optimising feature subsets to improve robustness, generalisability, and efficiency. The work in this paper addresses this gap by introducing an evolutionary ensemble framework that systematically combines diverse feature representations with optimised feature selection for speech-based depression detection.

1.1. Literature review

Recent advances in depression detection increasingly leverage multimodal data and diverse machine learning (ML) techniques. Several studies have highlighted the effectiveness of audio, textual, and physiological data in enhancing detection accuracy. For instance, Vandana et al. proposed a hybrid deep learning model that combines audio and textual features from the DAIC-WoZ dataset, finding that CNN models trained on audio achieved higher accuracy (98%) than those trained on text (92%), with Bi-LSTM models also yielding promising results (Vandana et al., 2023). Similarly, Philipthekkekara et al. introduced a CNN-BiLSTM model with attention mechanisms and reported a remarkable 96.71% accuracy using the CLEF2017 dataset (Philip Thekkekara et al., 2024). These findings suggest that deep learning architectures, particularly those integrating multiple modalities and attention mechanisms, can provide high precision in depression classification.

Parallel to multimodal approaches, a large body of research has emerged on detecting depression via social media platforms. Helmy et al. demonstrated the utility of Twitter in early depression identification by deploying ML models across both Arabic and English datasets, achieving F1-scores up to 96.6% (Helmy et al., 2024). Ghosal et al. developed a framework to differentiate depression and suicidal ideation using Reddit data, combining fastText, TF-IDF, and XGBoost to yield strong classification metrics (Ghosal & Jain, 2023). This direction aligns with other studies that apply machine learning on Reddit posts and specialised subreddits like SuicideWatch to identify users at risk (Desu et al., 2022). In the Chinese context, researchers have created a depression lexicon to extract semantic features from Sina Weibo posts, showing improved classification performance with feature fusion and boosting techniques (Guo et al., 2023). Collectively, these works underscore the growing potential of social media text as a rich, non-invasive resource for large-scale mental health screening.

Physiological signals, particularly EEG, offer another frontier in depression detection. EEG-based studies aim to overcome the limitations of self-reported symptoms and online behavioural cues by analysing brainwave patterns. For example, Khadidos et al. compared traditional ML and deep learning methods using band power features and found that CNN achieved the best results with 98.13% accuracy (Khadidos et al., 2023). Similarly, Song et al. introduced LSDD-EEGNet, an end-to-end framework integrating CNN and LSTM with domain adaptation, which showed superior performance in subject-independent settings (Song et al., 2022). Mohammed et al. also explored EEG with advanced feature extraction techniques (e.g., Fourier-Bessel series) and domain adaptation, achieving improved accuracy through LS-SVM and ensemble models (Mohammed & Diykh, 2023). These findings highlight that EEG signals, when analysed with robust deep learning and signal processing techniques, hold substantial promise for early and objective detection of depression.

2. Background

Detecting depression via audio signals is a challenging task, primarily due to the subtle and often subjective nature of depressive symptoms in speech (Koops et al., 2023). Depression can manifest through changes in voice pitch, tone, speed, and rhythm, but these variations may be subtle and overlap with other conditions or emotions, making it difficult to distinguish (Darby et al., 1984). Additionally, audio features like speech patterns can be influenced by factors such as cultural background, individual speech traits, or environmental noise, which adds complexity to accurate detection (Long et al., 2017). Moreover, depression's diverse symptoms and its potential to fluctuate over time further complicate the task of consistently identifying it through audio signals alone.

Ensemble learning algorithms outperform traditional machine learning models by combining multiple weak or base learners to create a stronger, more robust model, thereby improving generalisation and reducing overfitting (Dong et al., 2020). While individual models may perform poorly on specific data points, ensemble methods leverage the diversity of different models (or iterations) to minimise errors and enhance prediction accuracy. By aggregating the outputs of multiple classifiers, ensemble methods effectively capture complex patterns in the data, leading to better performance on unseen examples. This characteristic is particularly beneficial in detecting depression, where subtle variations in speech, text, or behavioural data may be hard to capture with a single model (Ansari et al., 2022). For instance, combining models that focus on different feature types (e.g., acoustic features, linguistic features, and non-verbal cues) can help detect depression more accurately, as each model contributes its specialised knowledge to the final prediction (Zhang et al., 2019). Additionally, ensemble methods like AdaBoost can handle imbalanced datasets, common in depression detection, by focusing on misclassified instances and iteratively improving the model's ability to detect more nuanced depressive symptoms.

The AdaBoost (Adaptive Boosting) algorithm is an ensemble learning technique designed to improve the performance of weak classifiers by iteratively combining them to form a robust, strong classifier (Ying et al., 2013). Initially, AdaBoost assigns equal weights to all training samples. In each iteration, it trains a weak learner, and evaluates their performance by calculating the weighted classification error. The error rate is used to compute a weight, referred to as α_t, which reflects the model's contribution to the final prediction. In this algorithm, misclassified samples are given higher weights, ensuring that the subsequent weak learner focuses on these harder-to-classify instances. This process continues iteratively, and the final strong classifier is a weighted sum of the individual weak classifiers, where the weight of each classifier is determined by its error rate. The algorithm's advantage lies in its ability to reduce bias by combining weak learners while simultaneously controlling variance using relatively simple models ultimately leading to improved generalisation, particularly effective in scenarios where the data is noisy or contains imbalanced classes.

The Broad Learning System (BLS) is a neural network framework based on the Random Vector Functional-Link Neural Network (RVFLNN) (Pao & Takefuji, 1992), designed to enhance learning efficiency through a shallow but wide architecture. It consists of a single-layer neural network with two main components in its hidden layer: feature nodes and enhancement nodes. First, feature nodes are generated by applying a linear transformation Φ to the input data, expanding the feature space. These feature nodes act as an intermediate representation that preserves essential input characteristics. Next, enhancement nodes are created by further transforming feature nodes through a nonlinear activation function ζ, which increases model expressiveness and nonlinearity without requiring deep layers. The final step involves computing the output weights W_m using the Moore–Penrose pseudoinverse, efficiently finding a least-squares solution without iterative backpropagation. Unlike traditional deep learning models that require layer-wise training, BLS can incrementally learn new data by adding additional nodes dynamically, making it highly scalable and computationally efficient. This architecture enables fast training (Gong et al., 2021), avoids vanishing gradient issues (Chen & Liu, 2017), and is well-suited for applications in classification, regression, and real-time learning tasks (Zhang et al., 2020).

The Universal Approximation Capability of the Broad Learning System (BLS) is a fundamental property allowing the system to approximate any continuous function on compact sets with sufficient feature and enhancement nodes (Chen et al., 2018). This capability is rooted in Random Vector Functional-Link Neural Networks (RVFLNN) principles, which BLS is built upon. BLS leverages randomly generated feature nodes and enhancement nodes, with the latter transforming the input data through a nonlinear activation function, thereby expanding the feature space (Chen & Liu, 2017). The theoretical foundation for BLS's universal approximation ability has been rigorously established, demonstrating that, even with randomly initialised weights and without the need for gradient descent or iterative training, BLS can probabilistically converge to the target function (Gong et al., 2021).

The Broad Learning System (BLS) is well-suited to serve as a weak learner in AdaBoost due to its inherent efficiency, flexibility, and scalability advantages (Chen et al., 2018). Unlike decision stumps or weak learners that may struggle with nonlinearity or limited feature representation, BLS utilises both linear and nonlinear transformations to expand the feature space, enhancing its ability to capture complex patterns while maintaining simplicity (Wu et al., 2022). Additionally, the use of the Moore–Penrose pseudoinverse for training ensures that BLS can be trained fast and efficiently, without requiring backpropagation or iterative optimisation, making it highly suitable for the iterative process of AdaBoost (Yang et al., 2021). Compared to weak learners like decision trees, BLS offers a more flexible and robust approach, as it can dynamically adapt by adding new nodes without retraining the entire model. This incremental learning ability makes BLS particularly effective in AdaBoost, as it allows for efficient updates and quick incorporation of misclassified samples, improving the overall performance with each boosting iteration. Thus, BLS not only complements the goals of AdaBoost by focusing on difficult-to-classify instances but also provides a more efficient and scalable solution than traditional weak learners.

Although feature selection is widely used in machine learning, research that has explicitly applied systematic feature selection to depression detection is limited. To the best of our knowledge, only two prior studies have directly targeted this challenge. In Chikersal et al. (2021), behavioural features are extracted and selected from smartphone and wearable data, and in Hassan and Kaabouch (2024), several classical feature selection methods on EEG-based depression detection are evaluated. These two works are restricted to single data modalities and rely on fixed selection techniques. The evolutionary method in this paper adaptively optimises diverse feature subsets across multiple heterogeneous speech feature families.

Numerous evolutionary and swarm-based algorithms have been proposed for feature selection, which mainly focus on optimisation frameworks for dimensionality reduction rather than on domain-specific applications like depression detection. For example, niching-based multiobjective FS methods (Wang et al., 2023), ant colony optimisation approaches for high-dimensional FS (Ma et al., 2021), weighted differential evolution for large-scale FS (Wang et al., 2022), multifactorial PSO for FS (Chen et al., 2022), coyote optimisation for binary FS (Thom de Souza et al., 2020), adaptive multi-objective GAs (Xue et al., 2021), dynamic sticky binary PSO (Nguyen et al., 2021), and BBPSO with mutual information (fang Song et al., 2021) have all been designed to evolve efficient and compact feature subsets. Other works, such as feature selection for scheduling heuristics (Zhang et al., 2021) or evolutionary ensembles for cross-subject emotion recognition (Zhang et al., 2024), address different problem domains. The work presented

in this paper is the first to adopt an evolutionary mechanism directly into an AdaBoost ensemble tailored for speech-based depression detection, where the aim is not only to reduce dimensionality but also to maximise ensemble diversity across heterogeneous feature families (Fourier, Wavelet, Hilbert–Huang, OpenSmile, and CNN-based representations). Moreover, while prior algorithms generally evolve a single nondominated set of feature subsets, EVAdaBoost dynamically evolves multiple specialised feature subsets, each assigned to different AdaBoost models, thereby improving robustness and capturing complementary depression-related cues that simpler feature selection or ensemble methods cannot. A comprehensive review on evolutionary approaches for feature selection can be found in Song et al. (2024).

3. The dataset

The Distress Analysis Interview Corpus (DAIC) (Gratch et al., 2014) is a collection of clinical interviews aimed at supporting the diagnosis of psychological distress conditions like anxiety, depression, and PTSD. The dataset is a multimodal corpus collected under multiple interaction settings, including face-to-face, teleconference, Wizard-of-Oz, and fully automated virtual agent interviews. Participants were recruited both online (via Craigslist) and in-person at a U.S. Vets facility in Southern California, ensuring diversity across civilian and veteran populations. All interviews were conducted in English with fluent speakers, and session durations ranged from 5 to 60 min. Recordings include high-quality audio, video, and depth sensor (Microsoft Kinect) streams, with some sessions additionally capturing physiological measures such as galvanic skin response, ECG, and respiration. Interview protocols followed a consistent semi-structured format covering neutral, symptom-focused, and cool-down phases, designed to elicit naturalistic speech while safeguarding participant well-being. The dataset is enriched with metadata such as demographic information, standardised psychological assessments (e.g., PHQ-9, PCL-C, STAI, PANAS), and subjective ratings of the interviewer or agent. The corpus provides extensive annotations of participant speech, including explicit mentions of mental health conditions, making it particularly well-suited for the study of depression detection. The combination of varied interaction modalities, multimodal recordings, and detailed psychological profiling offers a robust foundation for evaluating model performance and generalisability across populations, recording conditions, and affective states.

4. The proposed algorithm

This paper proposes an evolutionary AdaBoost ensemble learning to detect depression via audio signals. This section explains the algorithm.

4.1. Feature extraction

The proposed algorithm in this paper uses various feature extraction methods, as we believe that each type of feature captures certain characteristics of the voice signal that can be representative of depression. To the best of our knowledge, there is no paper to study this wide range of features in the detection of depression (Bhadra & Kumar, 2022; Cellini et al., 2022; Joshi & Kanoongo, 2022; Squires et al., 2023).

- *Time-Domain Features*: Time-domain features are critical for capturing the basic characteristics of speech signals. The mean of the signal represents the average amplitude, providing insights into the overall level of the voice, which can fluctuate due to emotional states such as depression. Entropy error and entropy estimation measure the predictability of the signal, with higher values indicating more complexity and unpredictability, which may be different among certain emotional states associated with depression. The histogram lower and histogram upper features represent the distribution of signal values, and extreme values in these features might suggest speech patterns with limited variability or expressiveness. RMS (Root Mean Square) measures the energy or loudness of the signal, which may be different in depressed individuals. Kurtosis quantifies the "tailedness" of the signal's distribution, with higher values indicating outliers, potentially reflecting more erratic speech patterns. Skewness measures the asymmetry of the distribution, which may be representative of depressed speech. The peak-to-peak amplitude reflects the difference between the highest and lowest points of the signal and can indicate reduced expressiveness. The crest factor, the ratio of the peak value to the RMS, may be different in depressed speech due to less emotional intensity. Finally, features such as shape factor, impulse factor, margin factor, and add factors capture the overall shape and impulsivity of the signal. These factors can reflect the reduced energy and expressiveness seen in depressed speech. For a more detailed description and mathematical formula of these features see Ben Ali et al. (2015).
- *Fourier Transform Features*: The Fourier transform provides insights into the frequency components of the signal. The frequency centre represents the central frequency of the signal, which may be different in depression as emotional states can alter vocal pitch and tone. RMS variance frequency measures the variance in the frequency domain, indicating the stability or fluctuation in the vocal frequencies. The variance in this feature may be different between depressed and non-depressed individuals. Similarly, the root variance frequency captures variations in the frequency that can be representative of depression. For a more detailed description of these features and the extraction process see Tran et al. (2013).
- *Wavelet Transform Features*: Wavelet transform allows for the extraction of both time and frequency information from the speech signal, making it particularly useful for detecting transient or dynamic speech characteristics. Statistical features such as mean, entropy error, entropy estimation, RMS, kurtosis, and skewness are then calculated for the first six decomposition levels. These statistical measures capture both global and local properties of the signal, including the energy distribution and variability at different scales. The RMS reflects the overall energy of the speech signal, which is usually different between people. Kurtosis and skewness offer insights into the distribution of the signal, where depressed speech may exhibit different variance and symmetry. The other features, such as peak to peak, crest factor, and add factors, further provide detailed characterisations of the signal's shape and behaviour, helping to distinguish speech characteristics associated with depression. For more information on how to extract these wavelet features from a signal see Hu et al. (2007).
- *Walsh Transform*: The Walsh transform is a non-sinusoidal, orthogonal transform that captures periodic components of the speech signal using square-wave basis functions rather than sinusoidal ones. This property makes it particularly effective at representing abrupt changes, block-like patterns, and energy distributions across time segments, which may not be as clearly captured by sinusoidal-based transforms such as the Fourier. The statistical features derived from Walsh coefficients (e.g., mean, RMS, kurtosis, entropy) can reveal subtle regularities and structural properties of speech. Depressed speech has been consistently associated with reduced prosodic variation, flatter intonation, and decreased spectral and temporal complexity (Garcia-Toro et al., 2000). The Walsh transform is well-suited to capture such reductions in variability, since its square-wave basis emphasises regions of stability or repetition in the signal. Furthermore, unlike wavelet or Hilbert–Huang transforms, which focus on multi-scale oscillatory patterns and intrinsic mode decomposition, the Walsh transform provides a complementary view by highlighting regularities and low-dynamic regions in the signal. This complementarity increases the diversity of representations within the ensemble framework and ensures that depression-related cues,

whether subtle prosodic flattening or more pronounced reductions in dynamic range, are captured. Prior studies have also noted that transforms sensitive to periodicity and regularity can be effective in detecting pathological changes in speech (Xiang et al., 2009).

- *Hilbert–Huang Transform (HHT)*: The Hilbert–Huang Transform (HHT), which combines Empirical Mode Decomposition (EMD) with Hilbert spectral analysis, is specifically tailored for non-linear, non-stationary signals, properties frequently exhibited by natural speech under emotional and cognitive modulation. Each speech signal is adaptively decomposed into Intrinsic Mode Functions (IMFs), which are then converted into instantaneous amplitude and frequency trajectories over time, yielding a high-resolution time-frequency-energy representation that captures both transient fluctuations and long-term trends (Huang et al., 2003). In the context of depression, speech often shows nonstationary characteristics such as slowed tempo, monotonic intonation, or reduced expressive modulation (König et al., 2022). The HHT is particularly well suited to detect these dynamics because its adaptive decomposition isolates evolving frequency components and amplitude modulations that static transforms may blur or miss. By capturing subtle shifts in instantaneous frequency, amplitude envelopes, and mode-specific behaviours over the utterance, HHT-based features can enrich the representation of depression-relevant vocal markers beyond what Fourier or wavelet-based features alone provide. For more information on this transform see Konar and Chattopadhyay (2015).
- *OpenSmile*: OpenSMILE (Eyben et al., 2010) extracts a diverse set of acoustic features from voice signals. Prosodic features, such as energy, pitch (F0), jitter, and shimmer, capture variations in loudness and intonation. Spectral features, including Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, and spectral flux, provide insights into speech timbre and articulation. Voice quality features, such as harmonics-to-noise ratio (HNR) and formants, help assess breathiness and resonance. Temporal features, like speech rate, pause duration, and voice activity detection (VAD), highlight psychomotor slowing. These extracted features serve as inputs for machine learning models, enabling the detection of subtle vocal markers of depression that may not be perceptible to human listeners, making them valuable for automated, objective, and non-invasive mental health assessments.

These features, when combined, offer a comprehensive approach to analysing speech for depression detection, allowing the model to capture both subtle and overt changes in vocal patterns that are indicative of depressive states. Creating an ensemble learning algorithm that uses all these diverse features to train a separate base learner offers several benefits. By leveraging multiple types of features, each base learner can specialise in detecting specific patterns in the speech signal, such as energy fluctuations, frequency shifts, or temporal variations, which may be linked to depression. Combining these specialised learners allows the ensemble to capture a wider range of acoustic and vocal characteristics that any single feature set might miss. This approach increases the model's robustness, as it can effectively handle the complex and multidimensional nature of speech data while addressing issues like feature redundancy and noise. Furthermore, ensemble methods enhance generalisation, improve classification accuracy, and help mitigate overfitting, making the system more reliable and effective in detecting depression through voice signals.

4.2. Convolutional Neural Networks

The Hilbert–Huang Transform (HHT), Short-Term Fourier Transform (STFT), and Wavelet Transform (WT) are used in this paper to extract features and convert raw audio signals into time–frequency representations, effectively creating spectrogram-like images. Once the audio is transformed into an image-like representation, CNNs can be employed to extract hierarchical features from these spectrograms. Unlike manually engineered features, CNNs automatically learn complex and subtle patterns by capturing spatial correlations within the image. These networks excel at identifying hidden structures, such as subtle variations in frequency and time that may be imperceptible to the human ear but are indicative of depression-related speech alterations. The strength of CNNs in this application lies in their ability to detect deep, non-linear dependencies in speech patterns, offering superior generalisation compared to traditional handcrafted features.

Convolutional Neural Networks (CNNs) are a class of feed-forward artificial neural network algorithms widely used for pattern recognition tasks. As Fig. 1 illustrates, a CNN comprises multiple convolutional and pooling layers. The example in the figure features two convolutional layers, two pooling layers, and a flattening layer, followed by four groups of feature maps and a fully connected layer at the end. The input to a CNN is typically a 2D image signal, which undergoes processing through various layers. Convolutional and pooling layers extract essential features from the input image, which are then passed to fully connected layers. A feature map is generated by applying a filter to the convolutional layers. This filter, represented as a matrix, moves across the image using a step size known as a stride, performing convolution operations. Each pixel in the feature map is computed as the dot product of corresponding pixels in the filter and the image. The key parameters of convolutional operations include connection weights, filter dimensions (width and height), the number of feature maps, and stride dimensions (width and height).

Like convolutional operators, pooling operators process an image by moving across it. These operators use a kernel matrix to compute and extract the maximum or average values from the previous layer. Pooling is designed to optimise computations by reducing the size of feature representations, thereby lowering computational costs, the number of parameters, and memory requirements. It achieves this by merging the outputs of multiple neurons from one layer into a single neuron in the next. The key parameters of the pooling operator include stride height, stride width, kernel height, kernel width, and pooling type. At the end of a CNN, one or more fully connected layers perform high-level reasoning and classification. In a fully connected layer, each neuron is linked to every neuron in the preceding layer. The CNN shown in Fig. 1 follows a specific sequence of convolutional, pooling, and fully connected layers, referred to as its architecture. The structure of this architecture significantly impacts the performance of the network. Additionally, numerical parameters such as kernel size, filter size, and stride size play a crucial role in determining effectiveness. Therefore, choosing an architecture tailored to the specific problem is essential for optimal performance.

In this paper, we use the method presented in Najaran (2023) to optimise the architecture of the CNN. This algorithm employs a genetic programming approach to optimise the structure of CNNs for diagnosing COVID-19 cases using X-ray images. It utilises a graph-based representation of CNN architecture, incorporating evolutionary operators such as crossover and mutation. The CNN architecture in this algorithm is defined by two sets of parameters: the skeleton, which specifies the arrangement and connections of convolutional and pooling layers, and the numerical parameters, which determine properties such as filter size and kernel size. The optimisation process follows a co-evolutionary scheme, refining both the skeleton and numerical parameters to enhance the performance of the CNN.

Extracting features using a CNN from time-domain features requires the signal to be transformed into an image. Fig. 2 illustrates the process of converting one-dimensional signals into an image suitable for CNN processing. To ensure compatibility with CNN architectures, all input images must have the same dimensions. In this approach, the signal is divided into M non-overlapping segments, each of size M. These segments are then stacked vertically to construct an $M \times M$ image. Since the total length of the signal typically exceeds M^2, the segments

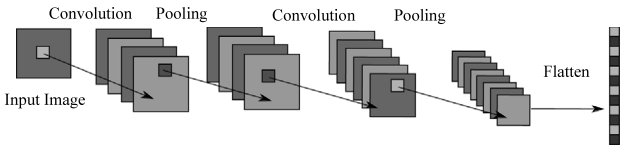

Fig. 1. The architecture of CNNs consists of a combination of convolution and pooling operators.

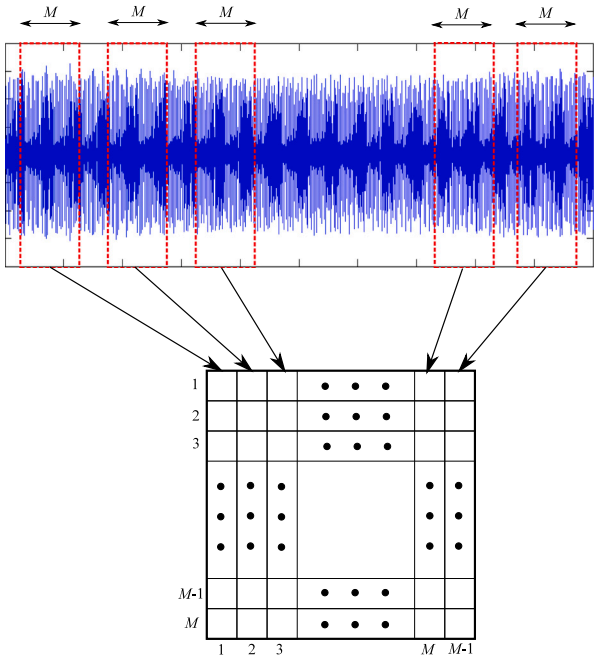

Fig. 2. The architecture of CNNs consists of a combination of convolution and pooling operators.

are randomly selected from different parts of the signal. Additionally, because the signal is converted into an image format, its values must be normalised to the [0, 255] range to match standard image processing requirements.

Training a CNN involves optimising its weights, filters, and kernels to achieve the best performance. In the literature, gradient-based optimisation algorithms are commonly used to determine these parameters efficiently. Compared to exhaustive or evolutionary search methods, gradient descent (GD) algorithms offer a significant advantage in speed, particularly for CNNs, which contain many parameters that must be fine-tuned during the learning process.

4.3. Evolutionary AdaBoost algorithm (EVAdaBoost)

Fig. 3, shows the proposed ensemble learning algorithm. In this algorithm, the voice signals are first passed to several transform functions. After applying the Fourier Transform, Walsh Transform, Hilbert–Huang Transform (HHT), Short-Term Fourier Transform (STFT), and Wavelet Transform (WT) to voice signals, statistical and deep learning-based features can be extracted for depression detection. Statistical feature extraction involves computing descriptive measures such as mean, variance, skewness, kurtosis, entropy, peak-to-peak amplitude, crest factor, and energy distribution from the transformed signals. These features capture irregularities, frequency shifts, and speech energy variations, which are crucial for identifying depression-related speech alterations, such as reduced articulation, monotonicity, and psychomotor slowing. Simultaneously, the outputs of these transforms can be visualised as spectrogram-like images, which are then fed into a

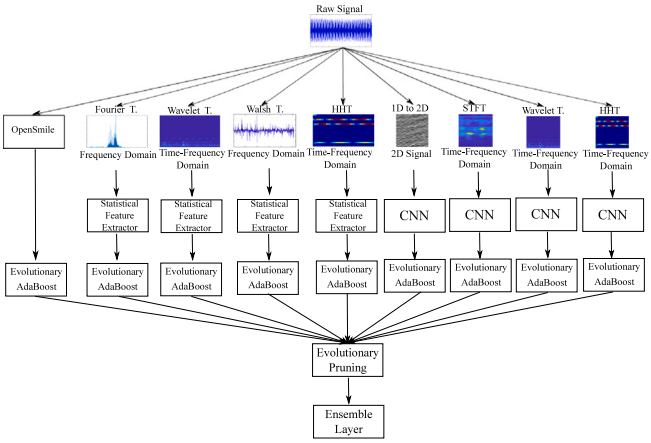

Fig. 3. The structure of the proposed ensemble learning algorithm.

Convolutional Neural Network (CNN) for automated feature extraction. CNNs excel at capturing hidden frequency-time relationships, texture patterns, and fine-grained spectral variations that may be imperceptible to traditional statistical methods. By leveraging hierarchical feature learning, CNNs can automatically identify depression-related spectral changes, such as flattened pitch, disrupted harmonic structures, and altered frequency energy distributions. Combining statistical and CNN-extracted features enhances the robustness of machine learning models, enabling a more comprehensive, data-driven approach to automated depression detection.

The extracted features are fed to the proposed evolutionary AdaBoost algorithm (EVAdaBoost) presented in algorithm 1. The EVAdaBoost algorithm receives as input the set of features extracted via each of the feature extraction algorithms, generating an ensemble of n AdaBoost algorithms, each trained on a specific subset of the features. These subsets of features are optimised through the evolutionary processes to create the set of AdaBoost algorithms with maximum performance and diversity. Since feature extraction methods often generate a large number of raw features, directly using all of them may introduce redundancy, noise, and unnecessary complexity. This approach performs an automated feature selection mechanism, as features that do not contribute meaningfully to classification performance are automatically discarded. This approach selects only the most informative and relevant features, eliminating redundant or highly correlated ones, resulting in enhanced accuracy and computational complexity.

Another advantage of this method is that it recognises that there is not a single optimal subset of features, as different feature combinations may capture different aspects of the data. The evolutionary process explores multiple feature subsets, ensuring that each AdaBoost model is trained on a unique, specialised set of features. This specialisation allows each AdaBoost classifier to focus on different discriminative patterns, improving the overall robustness of the ensemble. Moreover, by optimising the composition of these feature subsets, the algorithm ensures that selected features within each subset are complementary, allowing individual models to learn distinct yet relevant aspects of the data.

The evolutionary optimisation process also maximises diversity among the AdaBoost algorithms, a fundamental principle in designing effective ensemble learning models. Diversity in an ensemble refers to the degree of disagreement or variation in decision boundaries among the base learners. When classifiers in an ensemble make independent or weakly correlated errors, the ensemble can correct individual mistakes. If all base learners rely on the same highly correlated features, they may exhibit similar weaknesses, reducing the benefits of ensemble learning. This diversity is particularly beneficial in handling complex data distributions, class imbalances, and overlapping class regions

(as observed in highly complex tasks like depression detection), as different models capture different decision boundaries. Diversity-driven optimisation in feature selection ensures that the selected subsets are individually strong and complementary. Additionally, since AdaBoost focuses on misclassified samples, the diverse feature subsets allow each model to specialise in handling different types of misclassifications.

Another benefit of this method is its ability to adapt dynamically to different feature extraction methods, so each ensemble of AdaBoost is tailored to the type of features for which it is optimised.

One novelty of the proposed algorithm is the evolutionary selection of feature subsets. Instead of training all AdaBoost models on the full set of extracted features, which may introduce redundancy, noise, and overfitting, the evolutionary process searches through all possible subsets of selected features to find the optimised subsets tailored to each base learner. This ensures that non-informative or highly correlated features are discarded, while complementary and discriminative features are retained. The proposed algorithm does not converge to a single "optimal" subset, but rather evolves multiple diverse subsets, allowing each AdaBoost model to specialise in capturing different acoustic and temporal patterns in the speech signals. This diversity among feature subsets reduces correlated errors, increases robustness, and leads to consistently higher ensemble performance.

The AdaBoost algorithm presented in this paper adopts Broad Learning Systems (BLS) as its weak classifier. BLS is a strong choice as a weak classifier in the AdaBoost algorithm due to its efficient feature mapping, scalability, and adaptability to high-dimensional data. Unlike deep learning models, which require extensive training and parameter tuning, BLS incrementally expands its network structure, making it computationally efficient and well-suited for ensemble learning. Its ability to extract and transform features dynamically ensures that each weak classifier in AdaBoost can capture diverse aspects of the data, enhancing the ensemble's overall performance. Additionally, BLS is highly adaptable to feature diversity and redundancy reduction, which aligns well with AdaBoost's iterative weighting mechanism, allowing it to focus on harder-to-classify samples while maintaining generalisation. The combination of fast training, feature adaptability, and robustness makes BLS an excellent choice for weak learners in an AdaBoost ensemble, particularly for tasks involving complex, high-dimensional feature sets like depression detection. A detailed explanation of how BLS can be used as weak learners in AdaBoost can be found in Yun et al. (2024).

Quantum Evolutionary Algorithm (QEA) (Tayarani-N & Akbarzadeh-T, 2014) is specifically designed for the class of binary combinatorial optimisation problems (like satisfiability or knapsack problems), and because the feature selection problem in this paper is a binary combinatorial problem, this algorithm is well-suited to the problem. Section 5 performs experiments and shows that QEA offers the best performance among the existing algorithms.

A detailed description of the algorithm is presented as follows.

QEA employs a probabilistic representation for individuals, where each individual is encoded using a quantum bit (q-bit). A q-bit defines the probability of each bit in the individual being either zero or one. Consequently, in this representation, solutions are expressed as strings of probability density functions capable of representing binary strings and are denoted as follows:

$$q = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_j & \dots & \alpha_n \\ \beta_1 & \beta_2 & \dots & \beta_j & \dots & \beta_n \end{bmatrix}, \tag{1}$$

where $|\alpha_j|^2 + |\beta_j|^2 = 1$, $|\alpha_j|^2$ is the probability of the jth q-bit being zero, $|\beta_j|^2$ is the probability being one, and n is the problem dimension. In QEA, the evolutionary process is carried out using the "update" operator, which iteratively adjusts the values of α and β at each step of the algorithm to increase the probability of representing better solutions. The update operator in step 11 is defined as,

$$\begin{bmatrix} \alpha_j^{\tau+1} \\ \beta_j^{\tau+1} \end{bmatrix} = \begin{bmatrix} \cos(\delta\theta) & -\sin(\delta\theta) \\ \sin(\delta\theta) & \cos(\delta\theta) \end{bmatrix} \begin{bmatrix} \alpha_j^{\tau} \\ \beta_j^{\tau} \end{bmatrix}, \tag{2}$$

Algorithm 1: The proposed Evolutionary AdaBoost Algorithm (EVAdaBoost).

```
1  Initialise the algorithm parameters;
2  τ = 0;
3  Initialise the population Q^0 using Eq. (10);
4  observe Q^0 to generate X^0;
5  evaluate X^0;
6  store X^0 into B^0;
7  while not termination condition do
8      observe Q^τ to generate X^τ;
9      evaluate X^τ;
10     find the best neighbour of each q-individual and store in b^i
         if it is better than b^i ;
11     update Q^τ using Q-gate;
12     τ = τ + 1;
13 end
14 return the best solution;
```

where $\delta\theta$ is the rotation angle that controls the convergence speed of the algorithm and τ is the iteration of the algorithm. For a more detailed explanation of the QEA mechanism, see Tayarani-N and Akbarzadeh-T (2014). This paper uses $\delta\theta = 0.01$ as studies show that this value offers the optimal results for most of the problems (Tayarani-N & Akbarzadeh-T, 2014).

In step 1, the algorithm's parameters are initialised, including the population size and the number of AdaBoost learners in the ensemble, n.

In step 3 the population is initialised randomly. This algorithm aims to generate n different AdaBoost algorithms, each trained on a specific subset of features. Thus, the solutions in the population should represent the features used to train the AdaBoost algorithms. Each individual in this algorithm is represented by an $n \times m$ matrix x (m is the number of features), where the entries define the feature selection for each AdaBoost classifier in the ensemble: $x_{ij} = 1$ if the jth feature is selected to train the ith AdaBoost in the ensemble, and $x_{ij} = 0$ if it is discarded. Note that m differs for each feature type; for example, there are 384 OpenSmile features, and CNNs extract a specific number of features depending on their architecture.

In QEA, the initialisation is performed by setting the probability of the solutions being at zero or one state with the same probability as follows,

$$\begin{bmatrix} \alpha_{lf}^{i0} \\ \beta_{lf}^{i0} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}. \tag{3}$$

In step 5, the individuals are evaluated. To evaluate an individual, x, an ensemble of AdaBoost algorithms is generated and trained on the features suggested by x. The AdaBoost algorithms are trained based on a 5-fold cross-validation scheme on the training data, that is, 80% of the training data is used to train the models and 20% (validation data) to evaluate the performance. This is performed five times so all the data records are used at least once in the evaluation process. Note that this is performed only on the training data, not the test data. Once the n AdaBoost algorithms are trained, their diversity and performance on the validation data are calculated. The performance of the ensemble is measured as the average performance of the individual AdaBoosts, and the diversity among the AdaBoost algorithms in the ensemble (denoted by L) is measured as the sum of pairwise diversity among each pair of AdaBoost algorithms, as,

$$D(L, x) = \frac{1}{n(n-1)} \sum_{l=1}^{n} \sum_{h=1}^{n} D(L_l, L_h, x) \tag{4}$$

where the diversity between two AdaBoost algorithms L_l and L_h is measured as,

$$D(L_l, L_h, x) = \frac{1}{d}\sum_{k=1}^{d}\left(L_l(y_k, x) - L_h(y_k, x)\right)^2 \tag{5}$$

where L represented the AdaBoost ensemble, L_l represents the lth AdaBoost in the ensemble, x is the $n \times m$ matrix representing the selected features for the AdaBoosts in the ensemble, and d is the number of data records and y_k is the kth data record in the validation set. The evolutionary algorithm should maximise two objectives: the performance and the diversity of the AdaBoost algorithms. Thus, the fitness of the uth solution in the population, x^u is measured as the total number of solutions in the population dominated (outperformed) by the x^u,

$$\mathcal{F}(x^u) = \sum_{v=1, v\neq u}^{p}\left[\!\left[D(L, x^u) > D(L, x^v)\right]\!\right] + \sum_{v=1, v\neq u}^{p}\left[\!\left[\mathcal{P}(L, x^u) > \mathcal{P}(L, x^v)\right]\!\right], \tag{6}$$

where $[\![statement]\!]$ returns 1 if *statement* is true and returns 0 otherwise, and $\mathcal{P}(L, x^u)$ shows the average performance of the AdaBoost algorithms in the ensemble, where the solution x^u determines the features selected for each AdaBoost algorithm.

Similar to PSO, the QEA algorithm allows particles to evaluate the fitness of their neighbouring particles and adjust their positions accordingly. In step 10, each q-individual assesses the fitness of its neighbouring solutions, selects the best one, and updates its state using the update operator based on this selected value.

4.4. The proposed pruning algorithm

For each feature type (OpenSmile, Fourier, Wavelet, etc.), one ensemble of n AdaBoosts is generated. Thus, if there are r feature types (9 in this paper), a total of $n \times r$ AdaBoosts are generated. During the evolutionary optimisation in algorithm 1, these AdaBoost algorithms have been optimised to be diverse and perform the best on each specific type of features. However, these algorithms should perform the best when their decisions are aggregated in the final ensemble. The best combination of these AdaBoost algorithms should be found before they form the final ensemble. This is because many of these AdaBoosts may be redundant, adding unnecessary complexity to the algorithm, which results in reduced performance and higher time complexity. To manage this, an evolutionary pruning algorithm (see Fig. 3) is presented in this paper that searches through all possible combinations of the generated AdaBoosts and finds the optimal subset of algorithms that delivers the best performance at reduced computational cost.

This paper adopts the Quantum Evolutionary Algorithm (QEA) (Tayarani-N & Akbarzadeh-T, 2014) as presented in algorithm 1 to prune the set of $n \times r$ AdaBoosts and find the optimal subset of base learners. Similar to the feature selection problem, the pruning problem is a binary combinatorial problem for which QEA is well-suited. The pruning algorithm follows the same fundamental structure as the evolutionary AdaBoost algorithm 1, differing primarily in how each step is executed.

In step 3, the population is initialised. In this algorithm, each solution, denoted as z, is represented as an $n \times r$ matrix of zeros and ones, where $z_{ls} = 1$ indicates that the base learner L_l^s is selected for inclusion in the ensemble, while $z_{ls} = 0$ signifies that the base learner is pruned. In evolutionary algorithms, initialisation is performed randomly to ensure that solutions are uniformly distributed across the search space. In QEA, this is performed using Eq. (3). In the pruning problem in this paper, prior knowledge suggests that base learners with higher performance and greater contribution to diversity should be retained, while those with lower performance and minimal diversity contribution should be removed. Therefore, this paper proposes an initialisation method that generates quantum individuals with a higher probability of selecting high-performing base learners while pruning the less effective ones.

The *leave-one-out* cross-validation scheme is used to measure the performance of a base learner. The diversity brought to the ensemble by the base learner L_l^s is defined as the sum of the pairwise diversity between L_l^s and all other base learners in the ensemble,

$$\mathbb{D}(L_l^s) = \sum_{g=1}^{n}\sum_{h=1}^{r} D(L_l^s, L_g^h), \tag{7}$$

where $D(.,.)$ is the diversity between two base learners and is measured as Eq. (5). The quality of a base learner is measured as,

$$\mathcal{G}(L_l^s) = \sum_{g=1}^{r}\sum_{h=1}^{n}\left[\!\left[P(L_l^s) > P(L_h^g)\right]\!\right] + \sum_{g=1}^{r}\sum_{h=1}^{n}\left[\!\left[\mathbb{D}(L_l^s) > \mathbb{D}(L_h^g)\right]\!\right], \tag{8}$$

where $P(L_l^s)$ shows the performance of the lth AdaBoost algorithm trained on the sth feature type and $[\![statement]\!] = 1$ if *statement* is true and it is $[\![statement]\!] = 0$ otherwise. The base learners are then ranked according to their quality, and this ranking is used to determine the probability of each base learner being retained in the ensemble or pruned. In the initialisation step, the q-individuals are initialised to represent higher-quality AdaBoost algorithms with higher probability.

The rank of a base learner L_l^s is found as,

$$\mathcal{R}(L_l^s) = \sum_{g=1}^{r}\sum_{h=1}^{n}\left[\!\left[\mathcal{G}(L_l^s) < \mathcal{G}(L_h^g)\right]\!\right]. \tag{9}$$

The rank varies in the range $[0, r \times n - 1]$ where the highest quality AdaBoost has the rank 0 and the worst one has the rank equal to $r \times n - 1$. The following formula is proposed to initialise the q-individuals,

$$\begin{bmatrix}\alpha_{ls}^{i0}\\ \beta_{ls}^{i0}\end{bmatrix} = \begin{bmatrix}\sqrt{\frac{\mathcal{R}(L_l^s)}{r\times n-1}}\\ \sqrt{\frac{r\times n-\mathcal{R}(L_l^s)-1}{r\times n-1}}\end{bmatrix}. \tag{10}$$

With this initialisation scheme, q-individuals represent solutions where the highest-quality base learners have a probability of one of being retained in the ensemble, while the lowest-quality base learners have a probability of zero. The participation probability of the remaining base learners is assigned based on their quality.

The binary solutions z are evaluated by training the ensemble learning algorithm using the *leave-one-out* scheme and measuring its performance. To evaluate the solutions, the AdaBoosts are trained based on the features suggested by the evolutionary AdaBoost algorithm 1 and the selected AdaBoosts suggested by z are aggregated via a weighted voting scheme. The output of the ensemble algorithm for a data record y_k is found as,

$$L(y_k, z) = \sum_{l=1}^{n}\sum_{s=1}^{r} z_{ls} w_{ls} L_l^s(y_k), \tag{11}$$

where z_{ls} decides if the AdaBoost L_l^s participates in the ensemble, $L(y_k, z)$ is the output of the ensemble learning algorithm for the data record y_k, $L_l^s(y_k)$ is the output of the AdaBoost L_l^s for the data record y_k, and w_{ls} is the weight of the base learner L_l^s in the voting scheme. The learning process in the ensemble involves optimising the weights of the AdaBoosts in the voting system. These weights are optimised via a gradient descent algorithm, where the cost function is defined as,

$$C(L, z) = \frac{1}{|T|}\sum_{\forall y_k \in T}\left(L(y_k, z) - o_k\right)^2, \tag{12}$$

where T is the training set and $|T|$ denotes the number of data records in the set. The gradient of the cost function with respect to the voting weights is found as,

$$\frac{\partial C}{\partial w_{ls}} = \frac{1}{|T|}\frac{\partial \sum_{\forall y_k \in T}\left(L_l^s(y_k, z_{ls}) - o_k\right)^2}{\partial w_{ls}} = \tag{13}$$

$$\frac{1}{|T|}\frac{\partial \sum_{\forall y_k \in T}(L_l^s(y_k, z_{ls}) - o_k)^2}{\partial L_l^s(y_k, z_{ls})}\frac{\partial L(y_k, z_{ls})}{\partial w_{ls}} = \tag{14}$$

$$\frac{2}{T}\sum_{\forall y_k \in T}\left((L_l^s(y_k, z_{ls}) - o_k) z_{ls} L_l^s(y_k)\right). \tag{15}$$

In the gradient descent process, the weights are updated as,

$$\delta w_{ls} = \frac{2\eta}{T}\sum_{\forall y_u \in T}\left((z_{ls} L_l^s(y_u) - o_k) z_{ls} L_l^s(y_k)\right), \tag{16}$$

where η is the learning rate in the gradient descent algorithm. Once the ensemble algorithm is trained and the voting scheme weights w_{ls} are optimised, the ensemble's performance is evaluated to determine the fitness of the binary solution z. Fitness is assessed based on the algorithm's accuracy in predicting the target classes. The pruning algorithm aims to identify the optimal subset of base learners. After completing the pruning algorithm and selecting the best subset, the gradient descent algorithm in Eq. (16) is applied to optimise the weights of the base learners in the voting scheme.

Note that to find the solution's fitness, only the ensemble algorithm's voting weights are optimised via Eq. (16). Training the base learners L_l^s on the training data is performed only once, and is stored in a lookup table. The outputs of the base learners are not computed and are fetched from the table to calculate the fitness function.

5. Experimental results

This section performs experimental studies on the proposed algorithm and some state-of-the-art and classic machine learning algorithms. The experiments perform a comparison between the algorithms when different types of features are used to train the models, when all the features are used together to train the model and when the proposed ensemble method is used. The leave-one-out cross-validation is performed, so all the data appear at least once in the test set. All results are averaged over 30 runs. The classic learning algorithms used in this paper are described as follows. Logistic Regression (LR) is a linear model used for binary classification, estimating class probabilities via the logistic function and predicting outcomes based on a threshold. Random Forest (RF) is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes of the individual trees. Support Vector Machine (SVM) is a supervised learning algorithm that identifies the hyperplane maximising the margin between classes. K-Nearest Neighbours (KNN) is a simple, non-parametric, instance-based algorithm that classifies data based on the majority class among its 'k' nearest neighbours. Deep Feedforward Neural Network (DFFN) is a neural architecture where information flows in one direction through multiple hidden layers. Radial Basis Network (RBN) uses radial basis functions as activation functions. Learning Vector Quantisation (LVQ) is a supervised, prototype-based algorithm that adjusts prototype positions to approximate class boundaries. Probabilistic Neural Network (PNN) estimates class probabilities using kernel functions and classifies based on Bayes' rule. Radial Basis Function Network (RBE) is similar to RBN, using radial basis functions. Cascading Feedforward Neural Network (CFNN) dynamically adds layers based on data, adapting model complexity for potentially improved performance. Pattern Recognition Network (PRN) is optimised for recognising patterns in labelled data. Function Fitting Neural Network (FFNN) is designed to approximate functions by learning from data through weight adjustments. Feedforward Neural Network (FNN) is the most basic neural network type, where data flows forward through hidden layers.

The set of existing ensemble learning algorithms includes Stacking LR (STLR) (Jurek et al., 2014), Random Forest (RF) (Jurek et al., 2014), AdaBoost J48 (ABJ48) (Jurek et al., 2014), Oblique Random Forest (oRF) (Menze et al., 2011), (Multisurface Proximal Random Forest) MPRoF-P (Zhang & Suganthan, 2015), Majority Voting (MV) (Jurek et al., 2014), RoF (Zhang & Suganthan, 2015) and Classification by Cluster Analysis (CBCA) (Jurek et al., 2014). All experiments are averaged over 30 runs.

To measure the performance of different learning algorithms several metrics are used in this paper. For detecting depressed individuals, evaluation metrics derived from the confusion matrix help assess how well the model identifies those with and without depression. The confusion matrix includes True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Each of the following metrics offers insight into a different aspect of the model's behaviour:

Accuracy (ACC) measures the overall proportion of correctly classified individuals, both depressed and non-depressed:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

This gives a general sense of model performance, but can be misleading if the number of non-depressed individuals far exceeds the number of depressed ones.

True Positive Rate (TPR), also known as recall or sensitivity, indicates the proportion of truly depressed individuals that the model correctly identifies:

$$\text{TPR} = \frac{TP}{TP + FN}$$

This is especially important in mental health screening, where missing a depressed individual (false negative) can have serious consequences.

True Negative Rate (TNR), or specificity, measures the proportion of non-depressed individuals that are correctly classified:

$$\text{TNR} = \frac{TN}{TN + FP}$$

This is useful for understanding how well the model avoids incorrectly labelling healthy individuals as depressed.

Positive Predictive Value (PPV), also called precision, shows the proportion of individuals predicted to be depressed who actually are:

$$\text{PPV} = \frac{TP}{TP + FP}$$

This is important when the cost of falsely diagnosing someone as depressed is high, such as unnecessary stress or clinical follow-up.

False Positive Rate (FPR) represents the proportion of non-depressed individuals who are incorrectly predicted to be depressed:

$$\text{FPR} = \frac{FP}{FP + TN}$$

A lower FPR means the model is better at minimising false alarms among healthy individuals.

F1 score for the positive class (F1P) is the harmonic mean of precision and recall, balancing the need to detect depression while avoiding false positives:

$$\text{F1P} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

It is particularly useful when the dataset is imbalanced and both false negatives and false positives are critical.

F1 score for the negative class (F1N) is the harmonic mean of specificity and the negative predictive value, measuring performance on the non-depressed group:

$$\text{F1N} = \frac{2 \cdot TN}{2 \cdot TN + FN + FP}$$

This metric is relevant if the focus is on correctly identifying healthy individuals.

F1 score (F1) is commonly taken to be equal to F1P, focusing on the model's ability to correctly identify depressed individuals:

$$\text{F1} = \text{F1P}$$

Together, these metrics help evaluate the model's effectiveness in identifying depression while considering the trade-offs between different types of classification errors.

Table 1
The ACC of the proposed Evolutionary AdaBoost algorithm and some machine learning and ensemble algorithms for each type of features. The results are everaged over 30 runs.

	OpenSmile	Fourier	Wavelet	Walsh	HHT	1D to 2D	SFTF CNN	Wavelet CNN	HHT CNN	Rank
LR	69.16	67.8	67.13	65.5	65.21	65.74	64.44	65.31	62.83	9.69
RF	69.15	68.29	68.19	66.69	65.76	66.99	63.89	66.19	62.82	9.86
SVM	69.27	67.53	66.66	66.48	65.22	66.54	64.61	65.86	63.98	9.88
KNN	68.62	68.06	67.36	65.41	63.77	65.7	63.21	64.24	62.52	9.45
DFFN	68.19	67.88	66.52	66.31	65.08	66.69	64.2	65.73	62.24	9.65
RBN	68.07	67.8	67	64.89	63.94	65.28	63.34	64.37	62.88	9.48
LVQ	69.09	68.48	67.9	64.94	63.76	66.77	63.71	64.84	63.31	9.81
PNN	68.46	68.02	67.77	66.77	64.78	66.93	63.92	66.43	62.99	10.16
RBE	68.1	67.92	67.43	64.66	64.02	66.84	63.77	64.62	62.73	9.92
CFNN	68.44	67.13	65.29	64.85	63.42	64.97	62.79	63.6	63.12	9.43
PRN	68.48	67.39	66.81	64.76	64.2	65.32	63.16	64.07	62.91	9.42
FFNN	68.12	67.77	66.89	65.96	64.98	66.71	63.33	66.12	62.77	9.87
FNN	69.16	68.66	67.93	64.64	63.78	67.36	63.21	64.95	63.11	9.85
AdaBoost	75.98	74.76	74.47	73.16	72.59	73.33	71.57	72.7	70.74	13.41
MPRoF	76.63	76.11	75.59	74.83	72.85	74.88	71.97	73.99	71.8	13.86
STLR	76.33	75.45	74.82	73.87	71.9	73.71	71.44	72.89	71.31	13.88
CBCA	75.27	75.28	74.45	73.27	71.67	73.97	71.68	73.24	70.99	13.6
ABJ48	75.67	75.3	75.04	73.63	72.87	74.43	72.85	73.2	71.83	14.08
oRF	76.2	74.97	73.55	72.26	71.53	73.21	70.89	72.22	70.66	13.45
RoF	76.16	76.05	75.52	73.97	72.49	74.27	72.11	73.91	71.42	13.74
MV	76.46	76.09	74.48	73.24	72.35	73.75	71.96	72.63	70.56	13.62
EVAdaBoost	80.91	80.67	80.51	80.06	79.48	80.53	79	79.74	78.69	16.88
Rank	5.69	5.55	5.34	4.94	4.68	5.08	4.46	4.87	4.4	–

Table 1 shows the accuracy of the proposed Evolutionary AdaBoost (EVadaBoost) and some machine learning and ensemble learning algorithms for each type of feature. The results in this table demonstrate that the proposed EVAdaBoost algorithm consistently outperforms all other methods across all feature extraction techniques. Regardless of the input representation – whether handcrafted features such as OpenSmile and Fourier, or more complex representations like 1D-to −2D transformations and CNN-based features – EVAdaBoost achieves the highest accuracy, with values exceeding 80% in every case. This indicates that the integration of evolutionary strategies within the AdaBoost framework significantly enhances the model's ability to generalise and accurately detect depression from varied feature domains. Importantly, this performance edge is not marginal; EVAdaBoost improves upon even the best-performing ensemble methods (e.g., MPRoF, STLR, ABJ48) by several percentage points.

The data in Table 1 suggest that CNN-derived features achieve lower performance than handcrafted features such as Fourier, Wavelet, HHT, and OpenSmile. Several factors likely explain this observation. First, the dataset size in this paper may be insufficient for CNN-based representations, which typically require large-scale data. Second, handcrafted features embed domain knowledge and are tailored to speech characteristics—prosody, spectral balance, and perturbations, which are known markers of depression, providing a stronger inductive bias. Third, CNN-based features are often high-dimensional and noisy, making classical classifiers prone to overfitting or failing to exploit their richness. Finally, while raw CNN features underperform, our results with EVAdaBoost demonstrate that evolutionary feature subset selection and ensemble pruning can suppress irrelevant components, allowing CNN-based features to approach the effectiveness of handcrafted ones.

While EVAdaBoost is more complex than standard baselines, its components are motivated by specific challenges in depression detection from speech. Table 1 demonstrates that a wide range of simpler classifiers (e.g., LR, RF, SVM, KNN) combined with both handcrafted and CNN-derived features were thoroughly evaluated, yet none of these combinations consistently matched the performance of our proposed framework. The design choices justify the added complexity, as they translate into improvements across multiple evaluation metrics.

When examining the Friedman ranks, which reflect relative performance across all datasets and feature types, EVAdaBoost stands out with a rank of 16.88—higher than the next best method, ABJ48, which scores 14.08. Notably, traditional machine learning algorithms such as Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM) cluster around much lower rank values (approximately 9.4–9.9), highlighting the limitations of standard classifiers when applied to the complex task of depression detection.

Feature-wise, EVAdaBoost maintains a clear lead regardless of the extraction method used, though the performance is slightly higher for conventional signal-based features like OpenSmile, Fourier, and Wavelet. However, even in more challenging or abstract feature domains, such as SFTF CNN or HHT CNN, where other algorithms typically degrade in performance, EVAdaBoost remains strong. This implies that EVAdaBoost not only adapts well to high-dimensional and nonlinear data but also scales effectively across different types of feature representations.

When comparing the different feature extraction methods across all algorithms, it is evident that traditional handcrafted features like OpenSmile, Fourier, and Wavelet generally yield higher accuracies compared to more complex or transformed representations such as HHT, 1D-to-2D, or CNN-based features (SFTF CNN, Wavelet CNN, HHT CNN). OpenSmile consistently provides the highest performance among feature types, suggesting that it captures emotionally and acoustically relevant patterns effectively for depression detection. Fourier and Wavelet also show strong results, especially with top-performing models like EVAdaBoost and MPRoF, indicating their suitability for frequency-domain analysis of speech signals. In contrast, CNN-based features tend to result in lower accuracies, particularly for simpler classifiers like LR and KNN, which may not fully leverage the high-dimensional, structured representations extracted by deep models. This suggests that while CNN-based features have potential, their effectiveness depends heavily on the strength of the classifier, and simpler, well-engineered features still offer a reliable foundation for robust performance across models.

In depression detection, a high TPR is critical because it reflects the system's ability to successfully identify those suffering from depression—missing these cases (false negatives) can delay access to treatment and exacerbate mental health outcomes. Table 2 presents the True Positive Rate (TPR) results for the classifiers across the feature extraction methods. The proposed Evolutionary AdaBoost (EVAdaBoost) algorithm consistently achieves the highest TPR values across all feature types, with notable margins over both traditional machine learning models and competitive ensemble techniques. EVAdaBoost reaches a

Table 2
The TPR of the proposed Evolutionary AdaBoost algorithm and some machine learning and ensemble algorithms for each type of features. The results are everaged over 30 runs.

	OpenSmile	Fourier	Wavelet	Walsh	HHT	1D to 2D	SFTF CNN	Wavelet CNN	HHT CNN	Rank
LR	16.73	16.13	13.81	13.69	13.1	13.81	12.98	13.63	12.86	9.29
RF	16.25	16.61	15.89	13.27	13.51	13.15	13.04	13.1	12.98	9.26
SVM	15.42	15.77	14.4	15.6	14.7	15.77	14.11	15.24	13.15	10.11
KNN	16.25	16.13	15.6	14.11	13.57	14.64	12.92	13.63	13.27	9.12
DFFN	14.58	14.52	13.63	13.15	13.39	14.58	13.45	13.45	12.8	8.57
RBN	15.65	15.36	14.76	13.99	13.15	14.17	13.57	14.35	13.39	9.2
LVQ	16.43	15.71	14.11	13.51	13.27	14.4	12.98	13.04	13.27	8.67
PNN	14.7	15.42	13.69	14.35	13.1	14.23	13.57	14.29	13.27	9.14
RBE	15.65	15.71	15.06	14.05	13.75	14.76	13.27	13.04	12.8	8.75
CFNN	15.42	15.3	15.77	15.06	14.52	15.18	13.57	13.63	13.1	9.62
PRN	15.77	15.24	14.76	15.12	13.63	15.12	13.04	14.05	13.27	9.27
FFNN	16.61	15.3	14.64	15	13.99	14.64	12.86	14.64	12.86	9.18
FNN	15	16.55	15.24	13.93	13.93	14.05	13.51	14.05	13.45	9.13
AdaBoost	28.75	25.83	24.46	23.21	19.94	22.02	17.26	20.83	15.3	13.77
MPRoF	30.42	30.71	30.18	27.86	22.32	28.33	20.12	25.83	18.04	15.47
STLR	30.83	26.37	23.69	24.23	18.27	23.87	17.32	20.77	17.56	14.01
CBCA	28.15	27.02	25.89	21.61	17.44	23.99	16.96	22.92	15.89	14.46
ABJ48	27.38	28.69	27.32	22.02	21.37	26.79	21.37	21.01	18.45	14.96
oRF	31.19	25	24.11	20.42	16.61	21.9	16.19	18.99	15.71	13.49
RoF	30.36	31.13	27.44	22.44	18.63	24.23	19.7	22.98	16.43	14.84
MV	31.25	29.35	25.83	22.32	18.51	21.19	16.19	18.15	14.76	14.41
EVAdaBoost	37.98	36.19	36.13	36.01	32.32	36.55	30.42	32.8	30.24	18.27
Rank	5.86	5.64	5.44	5.04	4.56	5.23	4.33	4.85	4.05	–

peak TPR of 37.98% using OpenSmile features and maintains strong performance with other representations such as Fourier (36.19%), Wavelet (36.13%), and 1D-to-2D transformations (36.55%). This suggests that EVAdaBoost is especially effective in correctly identifying individuals with depression, making it highly valuable in screening contexts where false negatives must be minimised. Compared to other strong ensemble models like MPRoF, ABJ48, and RoF, EVAdaBoost provides a clear improvement in sensitivity, which is critical for early intervention and treatment in mental health applications.

Furthermore, the Friedman rank values reinforce EVAdaBoost's overall superiority, showing a leading average rank of 18.27, with a noticeable margin above the second-best model, MPRoF, with 15.47. While traditional classifiers like SVM, KNN, and RF cluster around much lower TPR scores and ranks (typically below 10), EVAdaBoost consistently outperforms them regardless of the feature type. Interestingly, while CNN-derived features generally yield lower TPRs for most classifiers, EVAdaBoost remains robust in these settings, suggesting its capacity to adapt well to high-dimensional, non-linear representations.

The results in Table 2 reveal notable differences in TPR of various feature types. Overall, CNN-based features—particularly those derived from HHT (HHT CNN), wavelet (Wavelet CNN), and SFTF (SFTF CNN)—consistently yield superior true positive rates (TPR), as reflected in their lower average ranks, with HHT CNN achieving the best average rank of 4.05. Traditional signal processing features such as Fourier, Wavelet, Walsh, and HHT also show competitive results, with Wavelet features outperforming Fourier and Walsh in most classifiers. OpenSmile, despite being a widely used handcrafted feature set, lags behind the deep learning-derived features, indicating that learned representations capture more discriminative patterns relevant to the classification task. Additionally, performance gaps become more pronounced when paired with complex ensemble classifiers such as EVAdaBoost, where CNN-based features consistently enhance TPR. This trend suggests that advanced feature representations, especially those leveraging both time–frequency transforms and deep learning, offer a clear advantage for this classification problem.

The results in Table 3 present the True Negative Rate (TNR) performance. TNR, also known as specificity, measures the proportion of actual negative cases (i.e., individuals not experiencing depression) that are correctly identified by the model. A high TNR is crucial in mental health applications because it ensures that healthy individuals are not misclassified as depressed, which could otherwise lead to unnecessary psychological concern, misallocation of clinical resources, or stigmatisation. Across the table, the Evolutionary AdaBoost (EVAdaBoost) classifier consistently delivers the highest TNR values, reaching 91.79% with OpenSmile and Fourier features, indicating exceptional reliability in excluding non-depressed cases.

A clear performance difference is evident between standard machine learning classifiers and ensemble-based methods. Conventional models such as Logistic Regression (LR), Support Vector Machines (SVM), and Feedforward Neural Networks (FFNN) typically achieve TNRs in the low 80% range. While adequate, this level of performance may still result in a concerning rate of false positives in real-world screening scenarios. Ensemble classifiers like AdaBoost, MPRoF, and RoF, by contrast, achieve significantly higher TNRs across nearly all feature sets, surpassing 85% and often approaching or exceeding 87%. These results highlight the ability of ensemble techniques to reduce false alarms.

When comparing the different feature types, traditional handcrafted features such as OpenSmile, Fourier, and Wavelet consistently outperform deep learning-based CNN features in terms of TNR. OpenSmile features attain the highest average rank (5.49), with Fourier and Wavelet close behind. These features offer stable and interpretable representations that appear well-suited for distinguishing non-depressed individuals. In contrast, features derived from spectrogram-based CNN approaches—like SFTF CNN, Wavelet CNN, and HHT CNN—generally rank lower. While these deep features may capture subtle emotional cues, their complexity may introduce noise or overfitting, particularly when used with simpler models. However, their TNR improves substantially when paired with advanced ensemble methods, indicating that with the right classifiers, CNN features can still contribute meaningfully to reducing false positives. Overall, these findings suggest that in depression detection tasks, both the choice of feature and the model architecture play a critical role in minimising the risk of misclassifying healthy individuals.

Table 4 displays the Positive Predictive Value (PPV) results. PPV, or precision, reflects the proportion of predicted positive cases (i.e., individuals classified as depressed) that are truly positive. In the context of depression detection, a high PPV is crucial for ensuring that those flagged by the model as experiencing depression are indeed likely to be suffering from it, thereby minimising unnecessary psychological distress and resource misallocation caused by false positives. The results indicate a broad performance spectrum, with ensemble models consistently outperforming traditional classifiers in terms of PPV.

Table 3
The TNR of the proposed Evolutionary AdaBoost algorithm and some machine learning and ensemble algorithms for each type of features. The results are everaged over 30 runs.

	OpenSmile	Fourier	Wavelet	Walsh	HHT	1D to 2D	SFTF CNN	Wavelet CNN	HHT CNN	Rank
LR	83.21	81.79	81.43	79.52	79.4	79.64	78.57	79.4	76.67	10.12
RF	83.33	82.26	82.02	81.07	80.12	81.67	77.86	80.71	76.55	10.26
SVM	83.57	81.43	80.6	80.36	78.93	80.12	78.45	79.52	77.86	10.2
KNN	82.74	82.14	81.43	79.4	77.5	79.52	77.14	77.98	76.31	9.82
DFFN	82.74	82.5	80.95	80.71	79.4	80.83	78.21	80.12	76.07	10.13
RBN	82.26	81.79	81.07	78.81	77.74	79.17	76.9	78.21	76.55	9.86
LVQ	82.98	82.74	82.5	79.05	77.86	81.31	77.62	79.05	77.26	10.31
PNN	82.98	82.5	82.38	80.83	78.81	81.07	77.62	80.48	76.9	10.61
RBE	82.26	81.9	81.67	78.57	78.1	80.83	77.86	78.81	76.67	10.33
CFNN	82.5	80.95	78.57	78.45	76.79	78.45	76.31	77.26	76.67	9.78
PRN	82.62	81.43	80.71	78.21	77.98	78.93	76.9	77.86	76.79	9.86
FFNN	82.02	81.79	81.07	79.88	79.05	80.71	77.38	80	76.67	10.25
FNN	83.45	82.86	82.02	78.69	77.5	81.9	76.9	78.81	76.79	10.33
AdaBoost	88.1	87.38	87.26	86.31	86.31	86.55	86.07	86.31	85.6	12.98
MPRoF	88.21	87.74	87.26	86.79	86.07	86.79	85.83	86.43	85.83	13.05
STLR	87.86	87.98	87.74	86.9	86.19	86.67	85.83	86.67	85.83	13.37
CBCA	87.5	87.5	86.9	86.55	85.95	86.79	86.07	86.31	85.6	13.16
ABJ48	87.98	87.38	87.26	86.9	86.31	86.79	86.31	86.67	85.83	13.45
oRF	87.74	87.74	86.67	85.83	85.83	86.55	85.6	86.07	85.48	13.1
RoF	87.74	87.62	87.74	87.26	86.43	87.38	86.07	87.02	85.95	13.14
MV	87.98	87.98	87.02	86.55	86.43	87.14	86.43	86.67	85.6	13.28
EVAdaBoost	91.79	91.79	91.67	91.31	91.31	91.67	91.19	91.55	91.07	15.61
Rank	5.49	5.36	5.28	4.93	4.79	5.07	4.63	4.88	4.57	–

Table 4
The PPV of the proposed Evolutionary AdaBoost algorithm and some machine learning and ensemble algorithms for each type of features. The results are everaged over 30 runs.

	OpenSmile	Fourier	Wavelet	Walsh	HHT	1D to 2D	SFTF CNN	Wavelet CNN	HHT CNN	Rank
LR	36.42	30.72	32.13	27.32	26.29	29.01	26.42	24.18	25.53	9.52
RF	35.51	35.73	38.73	32.02	28.15	29.77	26.81	29.48	22.55	9.75
SVM	36.21	31.23	31.89	31.31	27.12	37.02	27.94	29.55	21.24	9.9
KNN	35.18	33.28	32.45	25.27	18.97	26.47	20.91	25.44	20.02	9.36
DFFN	37.36	31.61	26.29	29.15	26.29	34.29	25.1	26.14	18.97	9.46
RBN	30.86	34.72	31.88	29.55	28.2	27.07	28.83	26.06	24.53	9.41
LVQ	42.61	35.57	31.79	22.92	21.88	23.88	25.27	25.37	24.22	9.68
PNN	33.57	34.68	35.27	36.71	33.64	36.87	28.8	37.04	21.4	10.24
RBE	33.56	35.94	34.92	31.66	23.26	36.92	26.53	27.83	25.66	10
CFNN	33.24	31.73	28.41	21.45	27.75	25.52	25.9	21.78	33.83	9.28
PRN	34.56	32.71	35.05	30.06	27.88	31.52	25.4	22.43	18.81	9.41
FFNN	38.73	38.37	34.28	27.51	27.21	38.46	23.1	33.17	22.39	9.86
FNN	38.25	31.31	39.12	23.58	26.18	34.83	27.46	31.46	25.64	9.94
AdaBoost	56.63	50.64	52.97	46.14	48.62	47.11	42.25	44.27	39.49	13.46
MPRoF	60.41	53.35	54.3	54.77	45.55	49.11	41.42	47.06	43.8	13.94
STLR	54.97	59.56	53.34	52.69	42.56	51.01	39.63	44.18	38.84	14.01
CBCA	50.31	52.48	48.15	47.2	41.23	53.05	46.41	51.21	44.67	13.61
ABJ48	63.82	61.05	62.23	47.92	42.9	54.89	45.49	45.77	49.67	14.3
oRF	53.75	50.52	44.35	50.74	50.65	49.91	42.74	52.7	37.74	13.59
RoF	59.34	49.41	53.49	48.52	45.41	48.36	43.1	50.04	46.48	13.91
MV	53.91	62.81	58.99	48.93	43.88	49.71	41.9	40.74	33.38	13.67
EVAdaBoost	69.72	67	67.12	65.62	74.64	69.35	68.3	69.24	66.39	16.73
Rank	5.67	5.53	5.31	4.96	4.72	5.14	4.47	4.83	4.38	–

Among all classifiers, the proposed Evolutionary AdaBoost (EVAdaBoost) algorithm stands out as the top-performing model, achieving the highest PPV scores across nearly all feature types. EVAdaBoost reaches a peak PPV of 74.64% using HHT features and maintains consistently strong performance across other feature sets, such as OpenSmile (69.72%), Fourier (67%), and SFTF CNN (68.3%). These results indicate EVAdaBoost's robustness and reliability in making accurate positive predictions, which is especially valuable in clinical screening scenarios where false positives can lead to undue anxiety, overdiagnosis, and misdirected interventions. Other strong ensemble methods, including ABJ48, MPRoF, and CBCA, also achieve respectable PPV values, though they lag behind EVAdaBoost both in absolute performance and average rank.

When comparing feature types, it becomes evident that certain representations are more conducive to higher PPV. Handcrafted features like HHT and OpenSmile deliver strong precision scores when paired with top-performing models, with HHT achieving the best PPV in EVAdaBoost. Interestingly, time–frequency representations and handcrafted features tend to outperform CNN-derived features in terms of PPV, with CNN-based features like SFTF CNN and HHT CNN yielding relatively lower PPVs across most classifiers. This suggests that while CNN-derived features may capture complex patterns, they may introduce noise or overfitting risks in certain classifiers, leading to less precise positive predictions. Overall, the results highlight the importance of both robust model architectures and well-suited feature representations for maximising PPV in depression detection tasks.

Table 5 reports the F1 scores – a harmonic mean of precision and recall that balances the trade-off between false positives and false negatives – for various classification models across different feature extraction methods, based on 30 experimental runs. The results highlight the superior performance of the proposed Evolutionary AdaBoost (EVAdaBoost) algorithm, which achieves the highest F1 scores for all

Table 5
The F1 of the proposed Evolutionary AdaBoost algorithm and some machine learning and ensemble algorithms for each type of features. The results are everaged over 30 runs.

	OpenSmile	Fourier	Wavelet	Walsh	HHT	1D to 2D	SFTF CNN	Wavelet CNN	HHT CNN	Rank
LR	19.69	18.7	16.6	15.87	15.08	16.06	14.89	15.54	14.49	8.99
RF	18.77	19.25	19.11	15.87	15.92	15.81	14.77	15.57	14.2	9.1
SVM	19.16	18.99	17	17.35	16.42	17.68	15.67	16.73	14.47	9.48
KNN	18.74	18.8	18.7	15.84	14.41	16.46	14.02	15.24	14.19	8.92
DFFN	18.19	17.64	15.71	15.54	15.55	17.28	15.05	15.7	13.79	8.81
RBN	18.29	17.96	17.67	16.1	15.08	15.8	15.18	16.17	14.59	8.92
LVQ	20.19	18.44	17	15.02	14.64	16.05	14.67	14.88	14.8	8.99
PNN	17.91	19.14	16.7	16.85	15.53	17.09	15.42	16.85	14.33	9.4
RBE	18.64	18.92	17.6	16.08	15.38	17.57	14.95	15.16	14.32	9.28
CFNN	18.84	18.04	17.27	15.72	16.15	16	15.14	14.66	15.24	8.94
PRN	18.84	17.59	17.49	16.58	15.37	17.62	14.48	15.15	14.31	8.94
FFNN	19.66	18.81	17.47	16.97	15.9	18.22	14.41	17.14	14.26	9.3
FNN	18.22	19.27	18.65	15.33	15.44	17.42	15.07	16.15	15.02	9.34
AdaBoost	33.6	31.27	27.97	27.87	22.74	27.23	21.51	24.97	19.46	13.95
MPRoF	36.28	35.32	33.64	31.89	26.33	32.26	23.33	29.88	22.38	15.15
STLR	35.47	31.4	28.97	27.63	22.24	27.95	21.58	24.99	22.03	14.51
CBCA	32.64	32.07	29.75	26.6	22.02	28.4	22.03	25.66	20.45	14.37
ABJ48	32.87	33.6	32.56	26.2	25.48	29.69	24.75	25.9	23.28	15.17
oRF	33.69	29.36	27.49	24.6	21.37	26.46	20.17	24.45	19.79	13.9
RoF	33.2	34.55	32.44	27.47	22.42	28.93	23.74	26.82	20.9	14.99
MV	35.9	33.82	30.49	27.04	22.54	26.2	20.32	22.92	18.3	14.23
EVAdaBoost	45.06	43.11	43.43	42.2	41.27	41.71	37.8	39.59	36.86	18.33
Rank	5.96	5.74	5.46	5.01	4.51	5.18	4.28	4.79	4.07	–

feature types, with scores ranging from 36.86% (HHT-CNN features) to 45.06% (OpenSmile features). Among the feature types, traditional handcrafted features like OpenSmile, Fourier, and Wavelet tend to yield higher F1 scores overall, especially when used with strong ensemble classifiers. In contrast, features derived from deep learning transformations (e.g., SFTF CNN, Wavelet CNN, and HHT CNN) generally show lower F1 performance across models, indicating a possible challenge in generalising from these high-dimensional representations without specialised training. The ranking row further confirms this trend, with handcrafted features like OpenSmile and Fourier consistently ranking higher in model performance than CNN-based features, suggesting they offer more robust information for F1-optimised depression detection.

Table 6 reports the results of a Friedman statistical analysis applied to evaluate the performance of different algorithms using the EVAdaBoost method across five key classification metrics. In this context, the *Sum of Squares (SS)* quantifies the total variability observed in the performance metric, divided into variability between algorithm groups (*Columns*) and within-group variability (*Error*). The *Degrees of Freedom (df)* represent the number of independent pieces of information used to calculate each SS value—21 for the number of algorithms and 609 for the residual error (based on the total number of observations minus the number of groups). The *Mean Square (MS)* is obtained by dividing each SS by its respective df, providing an average measure of variance. The *F-statistic* is then calculated as the ratio of the MS for the algorithm groups to the MS of the error, assessing whether the variance between groups is significantly greater than within groups. Finally, the *p-value (Prob > F)* indicates the probability that such a result could occur by random chance; smaller values (typically <0.05) suggest statistically significant differences. In this table, the p-values for ACC, TPR, PPV, and F1 are all extremely small (e.g., 4.66×10^{-7} for ACC and 5.43×10^{-30} for TPR), strongly indicating that the differences in algorithm performance are statistically significant for these metrics. However, the *p*-value for TNR (0.06905) exceeds the conventional 0.05 threshold, implying that differences among algorithms in terms of their ability to correctly identify negative cases are not statistically significant. Overall, the Friedman test confirms that the choice of algorithm has a substantial effect on most classification outcomes when using EVAdaBoost.

Table 7 shows the performance of different algorithms in terms of different metrics. The performance comparison across ensemble and non-ensemble algorithms reveals clear advantages of advanced ensemble methods—particularly EVAdaBoost—in the context of depression detection from speech data. EVAdaBoost achieves the highest accuracy (ACC = 86.38%) and the top true positive rate (TPR = 80%), indicating its superior ability to correctly identify depressed individuals while maintaining a high true negative rate (TNR = 87.98%). It also delivers the highest F1 scores for both positive (F1P = 71.84) and negative (F1N = 90.89) classes, reflecting a well-balanced model across sensitivity and specificity. Other strong performers include CBCA, MV, and ABJ48, all of which exhibit high precision (PPV) and robust F1 scores, though they fall slightly behind EVAdaBoost in overall performance. Traditional classifiers like Logistic Regression, SVM, and KNN trail significantly in

Table 6
The Friedman statistics of the results of EVAdaBoost algorithm presented in Tables 1, 2, 3, 4, and 5. This is for 30 independent runs and compares different algorithms.

ACC					
Source	SS	df	MS	F	Prob > F
Columns	2884.6	21	137.3619	69.233	4.6615e−07
Error	23 364.4	609	38.3652		
Total	26 249	659			
TPR					
Source	SS	df	MS	F	Prob > F
Columns	7331.5	21	349.119	194.0161	5.4263e−30
Error	16 475	609	27.0525		
Total	23 806.5	659			
TNR					
Source	SS	df	MS	F	Prob > F
Columns	1280.4833	21	60.9754	31.2846	0.06905
Error	24 505.5167	609	40.2389		
Total	25 786	659			
PPV					
Source	SS	df	MS	F	Prob > F
Columns	3009.7333	21	143.3206	73.3045	1.0279e−07
Error	22 856.7667	609	37.5316		
Total	25 866.5	659			
F1					
Source	SS	df	MS	F	Prob > F
Columns	6756.2667	21	321.727	161.9129	9.3011e−24
Error	19 532.2333	609	32.0726		
Total	26 288.5	659			

Table 7
The performance of different algorithms in terms of different metrics. These results are averaged over 30 runs. The last row shows the Friedman rank of each algorithm. A higher rank indicates better performance.

	ACC	TPR	TNR	PPV	FPR	F1P	F1N	F1	Rank
LR	78.03	38.51	87.98	60.06	12.02	41.82	86.11	41.82	11
RF	74.17	21.07	87.86	49.27	12.14	25.78	83.98	25.78	3.63
SVM	77.09	33.75	87.98	58.34	12.02	38.89	85.61	38.89	8.63
KNN	75.79	29.64	87.74	54.88	12.26	34.71	84.85	34.71	6.63
DFFN	76.47	31.73	87.86	57.51	12.14	35.41	85.22	35.41	7.56
RBN	74.72	23.51	87.86	51.57	12.14	28.06	84.32	28.06	4.44
LVQ	75.17	25.77	87.98	51.69	12.02	31.09	84.58	31.09	5
PNN	77.61	37.8	87.86	56.75	12.14	41.02	85.91	41.02	9.19
RBE	78.28	40.36	87.86	59.91	12.14	44.54	86	44.54	11.94
CFNN	75.49	26.85	87.98	51.76	12.02	30.94	84.73	30.94	6.13
PRN	77.08	35.42	87.86	61.16	12.14	39.31	85.24	39.31	9.25
FFNN	78.03	38.51	87.98	60.41	12.02	41.02	86.05	41.02	11.19
FNN	78.03	38.51	87.98	58.84	12.02	41.86	85.98	41.86	11.25
AdaBoost	82.24	74.17	84.4	62.15	15.6	63.71	87.82	63.71	14.5
MPRoF	82.57	76.19	84.17	64.73	15.83	65.12	87.99	65.12	15.75
STLR	80.1	60.48	85	62.12	15	55.87	86.7	55.87	13.38
CBCA	85.33	74.76	87.98	69.7	12.02	68.39	90.18	68.39	18.31
ABJ48	83.52	79.52	84.52	62.33	15.48	67.89	88.73	67.89	16.63
oRF	84.38	74.76	86.79	66.54	13.21	67.46	89.46	67.46	16.69
RoF	83.43	65.24	87.98	66.3	12.02	61.91	89.1	61.91	15.13
MV	85.08	75.65	87.5	68.42	12.5	68.3	89.95	68.3	17.75
EVAdaBoost	86.38	80	87.98	69.61	12.02	71.84	90.89	71.84	19.06

recall (TPR ranging from 21.07% to 40.36%), indicating a limitation in detecting depressed cases despite decent specificity. In terms of overall ranking based on Friedman rank, EVAdaBoost holds the top position (Rank = 19.06), outperforming both classic ensemble techniques (e.g., AdaBoost, Random Forest) and modern stacking or clustering approaches. These results collectively highlight the effectiveness of EVAdaBoost's evolutionary optimisation and feature selection strategies, offering a powerful and reliable tool for voice-based depression screening.

The proposed EVAdaBoost algorithm demonstrates exceptional consistency across all key metrics, making it particularly well-suited for clinical and large-scale mental health screening applications. With the highest accuracy (86.38%) among all tested models, EVAdaBoost achieves a strong baseline of overall correctness. However, what truly distinguishes it is its true positive rate (TPR) of 80%, which reflects its ability to correctly identify individuals who are actually depressed. In real-world mental health settings, this is critical, as failing to detect depression (false negatives) can result in untreated suffering and increased risk of deterioration. EVAdaBoost's high sensitivity indicates that it minimises this risk better than any competing model.

At the same time, EVAdaBoost maintains a very high true negative rate (TNR = 87.98%) and low false positive rate (FPR = 12.02%), meaning it is also reliable in avoiding misclassification of healthy individuals, which is important for reducing stigma and avoiding unnecessary interventions. Its positive predictive value (PPV = 69.61%) shows that the majority of individuals flagged as depressed by the model are indeed likely to be suffering from the condition—an essential quality for ensuring that limited clinical resources are directed towards those most in need. The high F1 scores for both the positive class (F1P = 71.84%) and negative class (F1N = 90.89%) indicate that the model strikes an effective balance between precision and recall, offering both thorough coverage and confidence in predictions. Taken together, these results suggest that EVAdaBoost is not only the most accurate model but also the most balanced and clinically practical, offering reliable detection without overwhelming healthcare systems or misclassifying healthy individuals.

Fig. 4 shows the Friedman statistics of the results of EVAdaBoost algorithm presented in Tables 1, 2, 3, 4, and 5. This is for 30 independent runs and compares different feature extraction methods. The box plots in the figure compares the proposed EVAdaBoost algorithm across different feature types (e.g., OpenSmile, Fourier, Wavelet, CNN-based representations) and performance metrics. Each box plot visualises the distribution of ranks for a given metric – such as accuracy (ACC), F1 score, true positive rate (TPR), true negative rate (TNR), positive predictive value (PPV), false positive rate (FPR), and F1 scores for positive (F1P) and negative (F1N) classes – across 30 experimental runs. Notably, OpenSmile, Fourier, and Wavelet features consistently result in the highest and most stable performance, with compact box plots centred near the top ranks, especially for ACC, TNR, and F1. Conversely, CNN-based features (SFTF, Wavelet, HHT CNN) exhibit slightly more variability and lower median ranks in some metrics like TNR and PPV, suggesting less consistent performance. Importantly, the tight distribution and high rank of EVAdaBoost across most metrics confirm its robustness, stability, and superior performance regardless of the feature type, validating its effectiveness as a versatile depression detection model.

Table 8
The comparison between the performance of the EVAdaBoost before and after the pruning algorithm is applied. The *p*-value shows the result of the Wilcoxon rank sum test between the first and second columns. These results are averaged over 30 runs.

Metric	Pruning	No pruning	*p*-value
ACC	86.38	86.19	0.89
TPR	80	79.52	0.96
TNR	87.98	87.86	0.96
PPV	69.61	68.68	0.9
FPR	12.02	12.14	0.96
F1P	71.84	70.87	0.81
F1N	90.89	90.72	0.87
F1	71.84	70.87	0.81

Table 8 presents a comparative analysis of the proposed EVAdaBoost algorithm before and after applying the evolutionary pruning mechanism. The pruning process is designed to reduce the number of base AdaBoost learners in the final ensemble by eliminating redundant or less informative models while maintaining or improving overall performance. The results in the table demonstrate that there is no statistically significant difference in performance between the pruned and unpruned ensembles across all metrics. For example, the accuracy of the pruned model is 86.38%, nearly identical to the unpruned model's 86.19%, with a *p*-value of 0.89, indicating no significant difference. Similarly, true positive rate (TPR) remains high in both configurations (80% vs. 79.52%), and true negative rate (TNR) and false positive rate (FPR) remain exactly the same, confirming that pruning does not

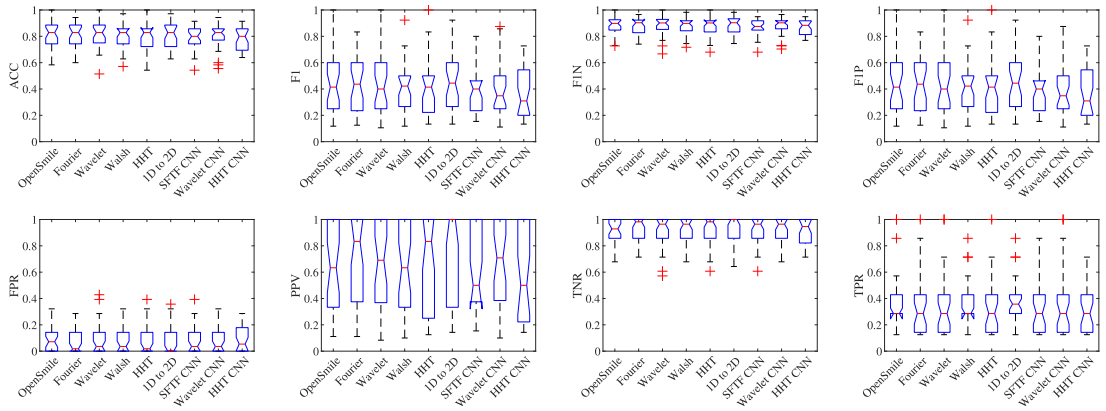

Fig. 4. The Friedman statistics of the results of EVAdaBoost algorithm presented in Tables 1, 2, 3, 4, and 5. This is for 30 independent runs and compares different feature extraction methods.

Table 9
The confusion matrices when all the features are used for different algorithms. In this table, PP is predictive positive, PN is Predicted Negative, AP is Actual Positive, and AN is Actual Negative. All results are averaged over 30 runs.

	RF		SVM		KNN		DFFN		RBN		LVQ		PNN	
	PP	PN	PP	PN	PP	PN	PP	PN	PP	PN	PP	PN	PP	PN
AP	1.5	5.7	2.37	4.67	2.1	5.1	2.23	4.87	1.67	5.5	1.83	5.4	2.67	4.5
AN	3.4	24.6	3.37	24.63	3.43	24.57	3.4	24.6	3.4	24.6	3.37	24.63	3.4	24.6
	RBE		CFNN		PRN		FFNN		FNN		AdaBoost		MPRoF	
	PP	PN	PP	PN	PP	PN	PP	PN	PP	PN	PP	PN	PP	PN
AP	2.83	4.23	1.9	5.27	2.5	4.67	2.7	4.33	2.7	4.33	5.2	1.87	5.33	1.67
AN	3.4	24.6	3.37	24.63	3.4	24.6	3.37	24.63	3.37	24.63	4.37	23.63	4.43	23.57
	STLR		CBCA		ABJ48		oRF		RoF		MV		EVAdaBoost	
	PP	PN	PP	PN	PP	PN	PP	PN	PP	PN	PP	PN	PP	PN
AP	4.23	2.77	5.23	1.77	5.57	1.43	5.23	1.77	4.57	2.43	5.3	1.73	5.6	1.4
AN	4.2	23.8	3.37	24.63	4.33	23.67	3.7	24.3	3.37	24.63	3.5	24.5	3.37	24.63

harm the model's ability to distinguish between depressed and non-depressed individuals. Metrics like F1 score for the positive class (F1P) and positive predictive value (PPV) also show negligible differences (p-values > 0.88).

These findings indicate that the pruning mechanism successfully reduces model complexity and computational cost by discarding redundant base learners without sacrificing predictive performance. This makes EVAdaBoost not only accurate and robust but also highly efficient and scalable, a critical property for real-world deployment in large-scale or resource-constrained mental health screening systems.

Table 9 presents the averaged confusion matrices over 30 runs for the learning algorithms when the full feature set is used. Most classical classifiers, such as RF, SVM, KNN, and standard neural architectures (DFFN, RBN, LVQ, PNN, etc.), show a systematic bias towards predicting the negative class. This is reflected in their relatively high counts of predicted negatives (PN ≈ 24–25) and comparatively low predicted positives (PP ≈ 1.5–2.7), regardless of whether the true label was positive (AP) or negative (AN). This imbalance suggests that these models are conservative in assigning the positive label, likely due to class imbalance or overlapping distributions in the feature space.

In contrast, ensemble-based methods, particularly AdaBoost, ABJ48, STLR, RoF, and the proposed EVAdaBoost, demonstrate a more balanced detection of actual positives. For example, ABJ48 achieves, on average, 5.57 correctly identified positives (APPP), while EVAdaBoost achieves 5.6, the highest among all algorithms. At the same time, these models present low predicted negatives for actual positives (PN ≈ 1.4–1.7), indicating stronger sensitivity. This improvement does not come at the cost of a significant increase in false positives for the actual negatives: the AN-PP values remain moderate (e.g., 4.33 for ABJ48 and 3.37 for EVAdaBoost).

Table 10
The training and inference time required for each algorithm in seconds. The data are averaged over 30 runs.

Algorithm	Training		Training	
	Mean	STD	Mean	STD
LR	0.08	5.48e−02	0.01	4.03e−03
RF	0.13	1.15e−02	0.01	8.42e−04
SVM	1.00	1.75e−01	0.08	1.29e−02
KNN	0.68	1.75e−01	0.06	1.29e−02
DFFN	0.98	1.15e−02	0.08	8.42e−04
RBN	3.40	7.50e−01	0.28	5.51e−02
LVQ	16.62	7.99e−02	1.38	5.87e−03
PNN	0.15	6.24e−02	0.01	4.59e−03
RBE	0.08	4.29e−02	0.01	3.15e−03
CFNN	2.03	5.48e−02	0.17	4.03e−03
PRN	3.25	1.17e−02	0.27	8.61e−04
FFNN	0.98	1.15e−02	0.08	8.42e−04
FNN	1.00	1.61e−02	0.08	1.19e−03
AdaBoost	30.36	1.94e+00	1.64	7.03e−02
MPRoF	37.69	1.80e+00	1.37	1.00e−01
STLR	39.67	1.90e+00	1.69	5.81e−02
CBCA	50.19	1.48e+00	1.93	6.26e−02
ABJ48	51.31	1.90e+00	2.12	5.88e−02
oRF	40.58	1.67e+00	2.36	1.05e−01
RoF	31.72	1.72e+00	2.10	9.06e−02
MV	51.22	1.61e+00	1.96	9.03e−02
EVAdaBoost	161.45	6.20e+00	1.50	5.45e−02

Table 10 presents the training and inference times for all classifiers. As expected, base learner algorithms such as LR, RF, or KNN require less training time, while more sophisticated ensemble approaches are

computationally more expensive. The proposed EVAdaBoost demonstrates the highest training time (161.45 s), due to the evolutionary optimisation process. However, the inference time is low (1.50 s), which is comparable to that of other ensemble algorithms. Since training is performed only once and deployment relies on inference, the added computational complexity is acceptable given the improved performance. Furthermore, the evolutionary nature of the algorithm allows parallel implementation, where the fitness of individuals can be evaluated simultaneously on different processing units, significantly reducing training overhead.

6. Conclusion

This paper proposes an evolutionary AdaBoost algorithm for automated depression detection via voice signals. The proposed method employs a diverse set of nine signal processing feature extraction methods, including Fourier, Wavelet, Walsh, Hilbert–Huang, and OpenSmile, as well as time–frequency transformations for CNNs. Each of these feature sets is used to train a specialised AdaBoost algorithm with Broad Learning algorithms serving as base learners. A quantum evolutionary algorithm was then employed to optimise the feature subsets assigned to each AdaBoost model via a wrapper scheme. The evolutionary AdaBoost algorithm evolves to select the best subset of features for each AdaBoost base learner to reduce noise, redundancy and keep relevant information. This evolutionary selection improves the performance of the classifiers both in terms of accuracy and computational efficiency. Once a set of AdaBoost algorithms have been designed, an evolutionary pruning algorithm is adopted to find the optimal subset of AdaBoost algorithms to optimise the performance and computational complexity of the algorithms. Experimental studies were performed on the algorithm, and it was shown that the proposed algorithm offers better performance compared to state-of-the-art algorithms.

The approach presented in this work can handle overfitting in different ways. First, generating multiple AdaBoost models on diverse feature subsets ensures the ensemble does not over-rely on dataset-specific patterns from any single feature set. Second, during model training, a 5-fold cross-validation scheme is employed, thereby enforcing evaluation on unseen partitions and preventing the models from memorising the training data. Finally, an evolutionary pruning algorithm enhances generalisability by removing redundant or overfit-prone learners. This process adopts leave-one-out cross-validation to carefully assess each base learner's contribution, ensuring that only those improving ensemble diversity and validation performance are retained. The experiments in this study are performed on independent test sets using a leave-one-out cross-validation protocol. This scheme guarantees that every reported result reflects performance on unseen data. Moreover, all experiments are averaged over 30 runs.

Beyond its strong empirical performance, this research highlights the value of evolutionary learning in managing high-dimensional, multimodal data, a common challenge in affective computing and mental health diagnostics. As mental health diagnostics increasingly turn to passive, technology-driven solutions, this work contributes a powerful, interpretable, and efficient tool for early-stage, speech-based depression screening in both clinical and real-world settings.

From a broader perspective, this work supports the growing shift towards objective, data-driven tools in mental health care, offering an alternative to traditional self-report questionnaires and clinical interviews that are often limited by bias or accessibility. By leveraging passive audio data, EVAdaBoost enables unobtrusive and continuous monitoring, which could be integrated into telehealth systems, mobile applications, or virtual agents for early detection and ongoing assessment.

The flexible, modular architecture of EVAdaBoost opens up several avenues for future research and development. One promising direction is the integration of multimodal data, such as textual content from interviews, facial expressions from video, or physiological signals like EEG or heart rate, to capture a more comprehensive representation of depressive symptoms. This would enable the framework to model complex emotional and cognitive states with greater accuracy. Additionally, the algorithm could be adapted to detect other psychological conditions, such as anxiety disorders, bipolar disorder, or cognitive decline, by fine-tuning the feature extraction and classification components for different symptom profiles. Future work could also explore online and continual learning capabilities, enabling the model to update itself with new data over time, which is particularly valuable for tracking mental health changes in longitudinal settings.

A valuable direction for future work is enhancing the explainability of the EVAdaBoost framework. While the model demonstrates strong predictive performance, understanding why specific predictions are made is critical for clinical adoption. Future research could integrate techniques such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) to highlight which features – such as specific vocal characteristics – most influence the model's decisions. Additionally, exploring layer-wise relevance propagation in the CNN components could offer insights into which time–frequency patterns are associated with depressive speech. Improving interpretability not only builds trust among clinicians and users but also supports more transparent and ethical deployment in mental health contexts.

One notable limitation of the data used in this study is that the voice data used for training and evaluation was collected in controlled laboratory settings. While this ensures high-quality recordings and consistent conditions, it may not fully capture the variability and noise present in real-world environments where depression detection systems are likely to be deployed. Factors such as background noise, spontaneous speech, emotional variability, and device heterogeneity can significantly affect performance in practical scenarios. As a direction for future work, it is essential to validate and adapt the EVAdaBoost framework using real-world datasets collected from naturalistic settings, such as phone conversations, telehealth consultations, or mobile app interactions, to ensure robustness, generalisability, and ecological validity. This would help bridge the gap between experimental performance and actual field effectiveness.

A key limitation of the current dataset, and a broader issue across much of the literature, is the binary classification of subjects into "depressed" and "non-depressed" groups. This oversimplifies the complex and spectrum-based nature of depression, which includes varying levels of severity (e.g., mild, moderate, severe) and different subtypes (e.g., melancholic, atypical, seasonal). Such dichotomous labelling may obscure important patterns within the data and limit the ability of models to capture the nuanced manifestations of depressive behaviour. Future work should focus on collecting and annotating datasets that reflect the full continuum of depressive symptoms, ideally informed by clinical assessments such as structured interviews or standardised rating scales. Additionally, machine learning models, particularly unsupervised or semi-supervised approaches, could be used to discover latent subcategories or symptom clusters that may not align neatly with existing diagnostic labels but hold clinical relevance. This could lead to more personalised, granular, and data-driven diagnostic tools that better reflect the heterogeneity of depression in real-world populations.

Given the sensitive nature of mental health diagnosis from voice data, some ethical concerns may arise. First, data collection must be conducted under strict protocols of informed consent, ensuring that participants understand the purpose of the study, how their data will be used, and how their identity will be protected. Privacy and data security require robust anonymisation strategies and secure storage to prevent misuse or unauthorised access. Moreover, false positives may lead to unnecessary anxiety or unwarranted interventions, while false negatives may delay needed support or treatment. Therefore, while the proposed algorithm shows promise as a screening tool, it should be positioned as an aid to clinicians rather than a standalone diagnostic

system, ensuring that automated predictions are always interpreted within a broader clinical and ethical framework.

Another limitation of this study is its exclusive focus on depression and reliance solely on audio data for detection. While depression is a critical and widespread mental health condition, it often co-occurs with or shares overlapping symptoms with other disorders such as anxiety, bipolar disorder, post-traumatic stress disorder (PTSD), and schizophrenia. These conditions can exhibit similar vocal patterns, such as reduced speech variability or slower tempo, but may also present distinct multimodal markers that are not captured through audio alone. To fully understand the complex and often interconnected nature of mental health disorders, future research should aim to collect datasets that include a broader range of psychological conditions, allowing for multi-label or hierarchical classification approaches.

Moreover, relying exclusively on voice data may limit the richness of the information available for diagnosis. Future datasets should incorporate multimodal inputs, such as video (facial expressions, gaze, gestures), EEG (neural activity), ECG or PPG (heart rate and variability), skin conductance (stress response), textual data (language patterns), and even wearable sensor data (sleep, movement, activity levels). These complementary data streams can provide a more comprehensive understanding of emotional and cognitive states, enabling the development of models that can better capture subtle patterns, distinguish between comorbid conditions, and support early detection and personalised interventions. Expanding the scope of data collection and target conditions is essential to advancing the generalisability and clinical utility of machine learning approaches in mental health.

A promising direction for future work is the development of models that support longitudinal and continuous monitoring of depression. Instead of relying on isolated assessments, future systems could analyse voice and other signals over extended periods to track the progression of depressive symptoms or response to treatment. Integrating the proposed framework with mobile devices or wearable technologies would enable real-time, passive monitoring in naturalistic environments, allowing for timely interventions and personalised care. This shift towards continuous data collection would enhance the clinical relevance and real-world applicability of automated mental health assessment tools.

Future work should also focus on enhancing cross-population generalisation to ensure the model performs reliably across diverse user groups. This includes evaluating the algorithm's robustness across different demographics, such as age, gender, cultural background, and language, as well as accounting for variations in dialects and accents. Additionally, testing the model on both clinical and non-clinical populations is essential to validate its broader applicability. Achieving this goal will require the collection of more diverse and representative datasets to reduce bias and improve the fairness and inclusiveness of depression detection systems.

An important direction for future research is the development of adaptive and personalised models that can tailor predictions to individual baselines and behavioural norms. Since vocal and emotional expression varies widely between individuals, static thresholds may lead to misclassification. Leveraging techniques like transfer learning or meta-learning can enable the model to calibrate itself to each user, improving accuracy and reliability over time. Personalised models would be especially valuable in long-term monitoring, where subtle changes from a person's own baseline are more informative than population-wide comparisons.

Future work must also address the ethical, legal, and social implications (ELSI) of deploying automated depression detection systems. Key considerations include ensuring privacy and data security, particularly when collecting sensitive mental health data through personal devices. Models should be evaluated for bias and fairness to prevent discriminatory outcomes across demographic groups. Additionally, systems must incorporate mechanisms for informed consent and give users control over their data and participation. Studying clinical usability, user acceptance, and the potential impact on stigma will be crucial for responsible, equitable, and effective real-world implementation.

To ensure real-world impact, future research should prioritise clinical validation and deployment of the proposed system. Collaborations with clinicians will be essential for conducting clinical trials or field studies that assess the model's effectiveness in real healthcare settings. Additionally, exploring integration with electronic health records (EHRs) and telehealth platforms can facilitate seamless adoption into existing workflows. Ultimately, the goal is to design deployable systems, whether for use by mental health professionals or as part of mobile health apps, that are practical, scalable, and capable of supporting early detection and ongoing mental health care.

CRediT authorship contribution statement

Ruhollah Sayeri: Methodology, Software, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Behnam Barzegar:** Conceptualization, Investigation, Resources, Visualization, Supervision, Project administration. **Yaser Bozorgi rad:** Conceptualization, Resources. **Nasser Mikaeilvand:** Conceptualization, Resources. **Mohammad Hassan Tayarani Najaran:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Alghamdi, N. S., Mahmoud, H. A. H., Abraham, A., Alanazi, S. A., & García-Hernández, L. (2020). Predicting depression symptoms in an Arabic psychological forum. *IEEE Access*, *8*, 57317–57334.

Almas, A., Forsell, Y., Iqbal, R., Janszky, I., & Moller, J. (2015). Severity of depression, anxious distress and the risk of cardiovascular disease in a Swedish population-based cohort. *PLoS One*, *10*(10), Article e0140742.

Ansari, L., Ji, S., Chen, Q., & Cambria, E. (2022). Ensemble hybrid learning methods for automated depression detection. *IEEE Transactions on Computational Social Systems*, *10*(1), 211–219.

Ben Ali, J., Fnaiech, N., Saidi, L., Chebel-Morello, B., & Fnaiech, F. (2015). Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals. *Applied Acoustics*, *89*, 16–27.

Bhadra, S., & Kumar, C. J. (2022). An insight into diagnosis of depression using machine learning techniques: A systematic review. *Current Medical Research and Opinion*, *38*(5), 749–771.

Cacheda, F., Fernandez, D., Novoa, F. J., & Carneiro, V. (2019). Early detection of depression: Social network analysis and random forest techniques. *Journal of Medical Internet Research*, *21*(6), Article e12554.

Cellini, P., Pigoni, A., Delvecchio, G., Moltrasio, C., & Brambilla, P. (2022). Machine learning in the prediction of postpartum depression: A review. *Journal of Affective Disorders*, *309*, 350–357.

Chen, C. P., & Liu, Z. (2017). Broad learning system: A new learning paradigm and system without going deep. In *2017 32nd youth academic annual conference of Chinese association of automation* (pp. 1271–1276). IEEE.

Chen, C. P., Liu, Z., & Feng, S. (2018). Universal approximation capability of broad learning system and its structural variations. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(4), 1191–1204.

Chen, K., Xue, B., Zhang, M., & Zhou, F. (2022). An evolutionary multitasking-based feature selection method for high-dimensional classification. *IEEE Transactions on Cybernetics*, *52*(7), 7172–7186.

Chikersal, P., Doryab, A., Tumminia, M., Villalba, D. K., Dutcher, J. M., Liu, X., Cohen, S., Creswell, K. G., Mankoff, J., Creswell, J. D., Goel, M., & Dey, A. K. (2021). Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection. *ACM Transactions on Computer-Human Interaction*, *28*(1), http://dx.doi.org/10.1145/3422821.

Darby, J. K., Simmons, N., & Berger, P. A. (1984). Speech and voice parameters of depression: A pilot study. *Journal of Communication Disorders*, *17*(2), 75–85.

Deshpande, M., & Rao, V. (2017). Depression detection using emotion artificial intelligence. In *2017 international conference on intelligent sustainable systems* (pp. 858–862). IEEE.

Desu, V., Komati, N., Lingamaneni, S., & Shaik, F. (2022). Suicide and depression detection in social media forums. In S. C. Satapathy, V. Bhateja, M. N. Favorskaya, & T. Adilakshmi (Eds.), *vol. 2, Smart intelligent computing and applications* (pp. 263–270). Singapore: Springer Nature Singapore.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*, 241–258.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on multimedia* (pp. 1459–1462).

Garcia-Toro, M., Talavera, J. A., Saiz-Ruiz, J., & Gonzalez, A. (2000). Prosody impairment in depression measured through acoustic analysis. *The Journal of Nervous and Mental Disease*, *188*(12), 824–829.

Ghosal, S., & Jain, A. (2023). Depression and suicide risk detection on social media using fasttext embedding and XGBoost classifier. *Procedia Computer Science*, *218*, 1631–1639. http://dx.doi.org/10.1016/j.procs.2023.01.141, URL: https://www.sciencedirect.com/science/article/pii/S1877050923001412. International Conference on Machine Learning and Data Engineering.

Gong, X., Zhang, T., Chen, C. P., & Liu, Z. (2021). Research review for broad learning system: Algorithms, theory, and applications. *IEEE Transactions on Cybernetics*, *52*(9), 8922–8950.

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., et al. (2014). The distress analysis interview corpus of human and computer interviews.. In *LREC* (pp. 3123–3128). Reykjavik.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43–49.

Guo, Z., Ding, N., Zhai, M., Zhang, Z., & Li, Z. (2023). Leveraging domain knowledge to improve depression detection on Chinese social media. *IEEE Transactions on Computational Social Systems*, *10*(4), 1528–1536. http://dx.doi.org/10.1109/TCSS.2023.3267183.

Hasin, D. S., Sarvet, A. L., Meyers, J. L., Saha, T. D., Ruan, W. J., Stohl, M., & Grant, B. F. (2018). Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. *JAMA Psychiatry*, *75*(4), 336–346.

Hassan, M., & Kaabouch, N. (2024). Impact of feature selection techniques on the performance of machine learning models for depression detection using EEG data. *Applied Sciences*, *14*(22), http://dx.doi.org/10.3390/app142210532, URL: https://www.mdpi.com/2076-3417/14/22/10532.

Helmy, A., Nassar, R., & Ramdan, N. (2024). Depression detection for twitter users using sentiment analysis in english and Arabic tweets. *Artificial Intelligence in Medicine*, *147*, Article 102716. http://dx.doi.org/10.1016/j.artmed.2023.102716, URL: https://www.sciencedirect.com/science/article/pii/S0933365723002300.

Hu, Q., He, Z., Zhang, Z., & Zi, Y. (2007). Fault diagnosis of rotating machinery based on improved wavelet package transform and SVMs ensemble. *Mechanical Systems and Signal Processing*, *21*(2), 688–705.

Huang, N. E., Wu, M.-L., Qu, W., Long, S. R., & Shen, S. S. (2003). Applications of Hilbert–huang transform to non-stationary financial time series analysis. *Applied Stochastic Models in Business and Industry*, *19*(3), 245–268.

Islam, M. M., Hassan, S., Akter, S., Jibon, F. A., & Sahidullah, M. (2024). A comprehensive review of predictive analytics models for mental illness using machine learning algorithms. *Healthcare Analytics*, Article 100350.

Jiang, H., Hu, B., Liu, Z., Yan, L., Wang, T., Liu, F., Kang, H., & Li, X. (2017). Investigation of different speech types and emotions for detecting depression using different classifiers. *Speech Communication*, *90*, 39–46.

Joshi, M. L., & Kanoongo, N. (2022). Depression detection using emotional artificial intelligence and machine learning: A closer review. *Materials Today: Proceedings*, *58*, 217–226.

Jurek, A., Bi, Y., Wu, S., & Nugent, C. D. (2014). Clustering-based ensembles as an alternative to stacking. *IEEE Transactions on Knowledge and Data Engineering*, *26*(9), 2120–2137.

Khadidos, A. O., Alyoubi, K. H., Mahato, S., Khadidos, A. O., & Nandan Mohanty, S. (2023). Machine learning and electroencephalogram signal based diagnosis of depression. *Neuroscience Letters*, *809*, Article 137313. http://dx.doi.org/10.1016/j.neulet.2023.137313, URL: https://www.sciencedirect.com/science/article/pii/S0304394023002720.

Konar, P., & Chattopadhyay, P. (2015). Multi-class fault diagnosis of induction motor using Hilbert and wavelet transform. *Applied Soft Computing*, *30*, 341–352.

König, A., Tröger, J., Mallick, E., Mina, M., Linz, N., Wagnon, C., Karbach, J., Kuhn, C., & Peter, J. (2022). Detecting subtle signs of depression with automated speech analysis in a non-clinical sample. *BMC Psychiatry*, *22*(1), 830.

Koops, S., Brederoo, S. G., de Boer, J. N., Nadema, F. G., Voppel, A. E., & Sommer, I. E. (2023). Speech as a biomarker for depression. *CNS & Neurological Disorders-Drug Targets-CNS & Neurological Disorders)*, *22*(2), 152–160.

Kour, H., & Gupta, M. K. (2022). An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. *Multimedia Tools and Applications*, *81*(17), 23649–23685.

Long, H., Guo, Z., Wu, X., Hu, B., Liu, Z., & Cai, H. (2017). Detecting depression in speech: Comparison and combination between different speech types. In *2017 IEEE international conference on bioinformatics and biomedicine* (pp. 1052–1058). IEEE.

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, *5*(1), 96–116.

Ma, W., Zhou, X., Zhu, H., Li, L., & Jiao, L. (2021). A two-stage hybrid ant colony optimization for high-dimensional feature selection. *Pattern Recognition*, *116*, Article 107933. http://dx.doi.org/10.1016/j.patcog.2021.107933, URL: https://www.sciencedirect.com/science/article/pii/S0031320321001205.

Menze, B. H., Kelm, B. M., Splitthoff, D. N., Koethe, U., & Hamprecht, F. A. (2011). On oblique random forests. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), *Machine learning and knowledge discovery in databases* (pp. 453–469). Berlin, Heidelberg: Springer Berlin Heidelberg.

Mohammed, H., & Diykh, M. (2023). Improving EEG major depression disorder classification using FBSE coupled with domain adaptation method based machine learning algorithms. *Biomedical Signal Processing and Control*, *85*, Article 104923. http://dx.doi.org/10.1016/j.bspc.2023.104923, URL: https://www.sciencedirect.com/science/article/pii/S1746809423003567.

Najaran, M. H. T. (2023). A genetic programming-based convolutional deep learning algorithm for identifying COVID-19 cases via X-ray images. *Artificial Intelligence in Medicine*, *142*, Article 102571.

Nguyen, B. H., Xue, B., Andreae, P., & Zhang, M. (2021). A new binary particle swarm optimization approach: Momentum and dynamic balance between exploration and exploitation. *IEEE Transactions on Cybernetics*, *51*(2), 589–603.

Pao, Y.-H., & Takefuji, Y. (1992). Functional-link net computing: Theory, system architecture, and functionalities. *Computer*, *25*(5), 76–79.

Philip Thekkekara, J., Yongchareon, S., & Liesaputra, V. (2024). An attention-based CNN-BiLSTM model for depression detection on social media text. *Expert Systems with Applications*, *249*, Article 123834. http://dx.doi.org/10.1016/j.eswa.2024.123834, URL: https://www.sciencedirect.com/science/article/pii/S0957417424007000.

Reece, A. G., Reagan, A. J., Lix, K. L., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Scientific Reports*, *7*(1), 13006.

Song, X., Yan, D., Zhao, L., & Yang, L. (2022). LSDD-EEGNet: An efficient end-to-end framework for EEG-based depression detection. *Biomedical Signal Processing and Control*, *75*, Article 103612. http://dx.doi.org/10.1016/j.bspc.2022.103612, URL: https://www.sciencedirect.com/science/article/pii/S1746809422001343.

fang Song, X., Zhang, Y., wei Gong, D., & yan Sun, X. (2021). Feature selection using bare-bones particle swarm optimization with mutual information. *Pattern Recognition*, *112*, Article 107804. http://dx.doi.org/10.1016/j.patcog.2020.107804, URL: https://www.sciencedirect.com/science/article/pii/S0031320320306075.

Song, X., Zhang, Y., Zhang, W., He, C., Hu, Y., Wang, J., & Gong, D. (2024). Evolutionary computation for feature selection in classification: A comprehensive survey of solutions, applications and challenges. *Swarm and Evolutionary Computation*, *90*, Article 101661. http://dx.doi.org/10.1016/j.swevo.2024.101661, URL: https://www.sciencedirect.com/science/article/pii/S2210650224001998.

Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Acharya, U. R., & Li, Y. (2023). Deep learning and machine learning in psychiatry: A survey of current progress in depression detection, diagnosis and treatment. *Brain Informatics*, *10*(1), 10.

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *Ieee Access*, *7*, 44883–44893.

Tayarani-N, M. H., & Akbarzadeh-T, M. (2014). Improvement of the performance of the quantum-inspired evolutionary algorithms: Structures, population, operators. *Evolutionary Intelligence*, *7*(4), 219–239.

Thom de Souza, R. C., de Macedo, C. A., dos Santos Coelho, L., Pierezan, J., & Mariani, V. C. (2020). Binary coyote optimization algorithm for feature selection. *Pattern Recognition*, *107*, Article 107470. http://dx.doi.org/10.1016/j.patcog.2020.107470, URL: https://www.sciencedirect.com/science/article/pii/S0031320320302739.

Tran, V. T., AlThobiani, F., Ball, A., & Choi, B.-K. (2013). An application to transient current signal based induction motor fault diagnosis of Fourier–Bessel expansion and simplified fuzzy ARTMAP. *Expert Systems with Applications*, *40*(13), 5372–5384.

Vandana, Marriwala, N., & Chaudhary, D. (2023). A hybrid model for depression detection using deep learning. *Measurement: Sensors*, *25*, Article 100587. http://dx.doi.org/10.1016/j.measen.2022.100587, URL: https://www.sciencedirect.com/science/article/pii/S2665917422002215.

Wang, X., Wang, Y., Wong, K.-C., & Li, X. (2022). A self-adaptive weighted differential evolution approach for large-scale feature selection. *Knowledge-Based Systems*, *235*, Article 107633. http://dx.doi.org/10.1016/j.knosys.2021.107633, URL: https://www.sciencedirect.com/science/article/pii/S0950705121008959.

Wang, P., Xue, B., Liang, J., & Zhang, M. (2023). Differential evolution-based feature selection: A niching-based multiobjective approach. *IEEE Transactions on Evolutionary Computation*, *27*(2), 296–310.

World Health Organization (2023). Depression. URL: https://www.who.int/news-room/fact-sheets/detail/depression. (Accessed 30 May 2025).

Wu, S., Wang, J., Sun, H., Zhang, K., & Pal, N. R. (2022). Fractional approximation of broad learning system. *IEEE Transactions on Cybernetics*, *54*(2), 811–824.

Xiang, X., Zhou, J., Li, C., Li, Q., & Luo, Z. (2009). Fault diagnosis based on walsh transform and rough sets. *Mechanical Systems and Signal Processing*, *23*(4), 1313–1326.

Xue, Y., Zhu, H., Liang, J., & Słowik, A. (2021). Adaptive crossover operator based multi-objective binary genetic algorithm for feature selection in classification. *Knowledge-Based Systems*, *227*, Article 107218. http://dx.doi.org/10.1016/j.knosys.2021.107218, URL: https://www.sciencedirect.com/science/article/pii/S0950705121004809.

Yang, K., Liu, Y., Yu, Z., & Chen, C. P. (2021). Extracting and composing robust features with broad learning system. *IEEE Transactions on Knowledge and Data Engineering*, *35*(4), 3885–3896.

Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. arXiv preprint arXiv:1709.01848.

Yazdavar, A. H., Al-Olimat, H. S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., Pathak, J., & Sheth, A. (2017). Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017* (pp. 1191–1198).

Ying, C., Qi-Guang, M., Jia-Chen, L., & Lin, G. (2013). Advance and prospects of AdaBoost algorithm. *Acta Automatica Sinica*, *39*(6), 745–758.

Yun, F., Yu, Z., Yang, K., & Chen, C. P. (2024). Adaboost-stacking based on incremental broad learning system. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, L., Li, J., Lu, G., Shen, P., Bennamoun, M., Shah, S. A. A., Miao, Q., Zhu, G., Li, P., & Lu, X. (2020). Analysis and variants of broad learning system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *52*(1), 334–344.

Zhang, F., Mei, Y., Nguyen, S., & Zhang, M. (2021). Evolving scheduling heuristics via genetic programming with feature selection in dynamic flexible job-shop scheduling. *IEEE Transactions on Cybernetics*, *51*(4), 1797–1811. http://dx.doi.org/10.1109/TCYB.2020.3024849.

Zhang, X., Shen, J., ud Din, Z., Liu, J., Wang, G., & Hu, B. (2019). Multimodal depression detection: Fusion of electroencephalography and paralinguistic behaviors using a novel strategy for classifier ensemble. *IEEE Journal of Biomedical and Health Informatics*, *23*(6), 2265–2275.

Zhang, L., & Suganthan, P. N. (2015). Oblique decision tree ensemble via multisurface proximal support vector machine. *IEEE Transactions on Cybernetics*, *45*(10), 2165–2176.

Zhang, H., Zuo, T., Chen, Z., Wang, X., & Sun, P. Z. (2024). Evolutionary ensemble learning for EEG-based cross-subject emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, *28*(7), 3872–3881.