Multimedia Tools and Applications (2023) 82:11461–11487
https://doi.org/10.1007/s11042-022-14052-y

1226: DEEP-PATTERNS EMOTION RECOGNITION IN THE WILD

A hybrid deep feature selection framework for emotion recognition from human speeches

Aritra Marik[1] · Soumitri Chattopadhyay[1] · Pawan Kumar Singh[1]

Received: 6 November 2021 / Revised: 1 June 2022 / Accepted: 6 October 2022 /
Published online: 27 October 2022

Abstract

Speech Emotion Recognition (SER) is an active area of signal processing research that aims at identifying emotional states from audio speech signals. Applications of SER range from psychological diagnosis to human-computer interaction and as such, a robust framework is needed for accurate classification. To this end, we propose a two-stage hybrid deep feature selection (HDFS) framework that combines deep learning with automated feature engineering for emotion recognition from human speeches, which shines both in terms of accuracy and computational efficiency. Our pipeline extracts self-learned features using a customized Wide-ResNet-50-2 deep learning model from mel-spectrograms of raw audio signals, whose dimensionality is reduced using a hybrid deep feature selection algorithm that comprises a fuzzy entropy and similarity-based feature ranking method, followed by Whale optimization algorithm, which is a popular meta-heuristic optimization algorithm in literature. A k-nearest neighbor classifier is used to classify the optimized feature subset into the respective emotion classes. The proposed pipeline is evaluated on three publicly available SER datasets using a 5-fold cross-validation scheme, where it is found to outperform several state-of-the-art existing works in literature by significant margins thus, justifying the superiority and reliability of the proposed research. The source codes of the proposed method can be found at: https://github.com/soumitri2001/Wrapper-Filter-Speech-Emotion-Recognition.

Keywords Speech emotion recognition · Deep learning · Feature selection · Fuzzy entropy & similarity measures · Whale optimization algorithm

✉ Pawan Kumar Singh
pawansingh.ju@gmail.com

Aritra Marik
aritramarik2002@gmail.com

Soumitri Chattopadhyay
soumitri.chattopadhyay@gmail.com

[1] Department of Information Technology, Jadavpur University, Jadavpur University Second Campus, Plot No. 8, Salt Lake Bypass, LB Block, Sector III, Salt Lake City, Kolkata, - 700106, West Bengal, India

1 Introduction

Speech signals are considered the most natural and intuitive means of social communication, an interactive episode that mostly comprise the conveying of different emotion states via conversation. Therefore, the effective transfer of emotion concepts from the speaker to the listener is of utmost importance so as to interpret and analyze the actual instinct behind an individual's communication. This has where speech emotion recognition (SER) [5, 29] comes into play, the task being to investigate and accurately predict the emotion class a speech sample belongs to. SER has been instrumental to the success of human-computer interaction (HCI) [19, 54], detection of clinical depression [7, 45] and therapy [39, 57] through verbal interpretation. Thus, it is crucial to develop a dependable and automated SER system for high confidence emotion classification from speech audio clips.

SER has mostly been tackled using conventional machine learning (ML) approaches [14, 66] whose pipeline comprises extraction of handcrafted features followed by classification into respective emotion classes. The key to success for such techniques lies in the choice of the best suited feature descriptor (such as LPC, MFCC or RAASTA features), which requires manual feature engineering and therefore, is subject to several rounds of trial-and-error. Furthermore, situations may arise when traditional feature extractors vary in performance across different speech datasets, and thus a combination of multiple descriptors [17] is required to obtain a more optimal feature set, which demand more storage requirements and also increase the number of trial combinations.

On the contrary, deep learning methods [29] alleviate the troubles of handcrafted feature extraction by providing a self-learning paradigm that can automatically generate the most informative features describing the raw data. Further, they also provide an end-to-end pipeline [31], removing the need for explicit feature engineering. In the context of SER, deep learning has been leveraged predominantly in two directions – modelling upon sequential raw audio data [26, 49] and using vision-based models on audio mel-spectrograms [31, 44]. While the former approach is computationally expensive, conversion of raw signals to spectograms map the temporal audio sequence to a *frequency-based spatial spectrum*, allowing the use of state-of-the-art vision models [25, 60, 67] for robust classification. However, a limitation of deep learning models is that they require huge amounts of data for achieving desirable performance, which is a bottleneck in this case of datasets curated for SER tasks [11, 33]. Transfer learning is one of the solutions to this problem, where a model trained on a large corpus (such as ImageNet [16]) is reused on the problem at hand. In this study, we have used the mel-spectrogram transformations of the raw speech signals for emotion detection, employing a customized Wide-ResNet-50-2 [67] network pre-trained on the ImageNet database as the CNN feature extractor backbone.

Feature selection (FS) [3, 50] aims at selecting the most optimal subset from a given feature set with the objective of enhancing discriminatory performance as well as reducing storage requirements and thereby making the pipeline computationally efficient. The number of features extracted by the CNN backbone is quite large and as such, may contain redundant information that limit the performance of the model, bringing forth the need for FS. Two prominent approaches have been used for FS, one being to rank features based on intrinsic properties among them [37, 53], and the other being to select the most optimal subset for a heuristic objective [12, 23]. In this study, we have leveraged a two-tier FS approach that uses both the intrinsic property-based feature ranking as well as a heuristic-based algorithm for dimensionality reduction of the feature space and enhanced classification performance. Specifically, we use a *fuzzy entropy and similarity based metric* [37] for ranking features from which a top-$q\%$ subset is chosen, which is further optimized using Whale Optimization Algorithm (WOA) [47], a nature-inspired meta-heuristic inspired from the social behaviour of humpback whales. The final feature subset selected by WOA is fed into a k-nearest neighbors (KNN) [8] classifier to make the final predictions.

The main contributions of the present research are as follows:

1. A bi-stage wrapper-filter hybrid deep feature selection (HDFS) framework has been proposed for dimensionality reduction of feature space and robust classification of emotions from speech data.
2. A customized pre-trained Wide-ResNet-50-2 CNN network [67] has been fine-tuned to extract features from mel-spectrogram transformations of raw audio clips, after which a fuzzy entropy and similarity measure based FS strategy [37] has been employed to rank features based on metric scores. A top-$q\%$ subset is selected from the ranked features, the value of 'q' being set experimentally.
3. WOA [47] has been used to further refine the top-$q\%$ feature subset and select the most discriminative features for enhanced performance.
4. The proposed two-tier HDFS approach is evaluated on three publicly available benchmark SER datasets [11, 33, 36] and compared with several existing works in literature. The proposed approach achieves classification accuracies of 93.64%, 96.25%, and 89.72% on the respective datasets, outperforming many existing state-of-the-art techniques by significant margins.

The rest of the paper is organized as follows: Section 2 reviews some of the recent developments in the relevant areas of speech emotion detection and feature selection; Section 3 elaborately describes the proposed SER framework; Section 4 discusses the results obtained upon evaluation of the proposed pipeline on three publicly available SER datasets along with a comparative study against several state-of-the-art works on SER in literature; and Section 5 concludes the findings of the present study.

2 Related work

SER [1, 5] has been a field of active research for over two decades, primarily due to its application in healthcare, social robotics and understanding human behaviour. Mostly, researchers have leveraged traditional ML methods [14, 17, 51, 61, 66] for classification of handcrafted features extracted from audio signals. Danisman et al. [14] fused MFCC and energy-based features and trained an ensemble of support vector machine (SVM) classifiers for SER. Albornoz et al. [6] extracted accoustic and prosodic features from speech samples and employed a hierarchical classification scheme for emotion recognition. The authors of [51] proposed a handcrafted feature fusion framework followed by dimensionality reduction of the feature space, while Song et al. [61] introduced feature selection (FS) based transfer subspace for cross-corpus SER. More recent works that leverage handcrafted feature engineering include a hybrid meta-heuristic based FS framework [17]; a quantum-modified swarm-based algorithm [13] for dimensionality reduction of fused handcrafted features; and a clustering-based genetic algorithm (GA) [27] for optimization of raw audio features.

Deep learning-based methods [18, 24, 31, 43, 44], on the other hand, can learn relevant informative features automatically from the raw data, thereby alleviating the explicit need for handcrafted features to be extracted from data samples. Typically, SER has seen two prominent directions of research pertaining to deep learning-based approaches: (1) feeding raw audio samples (or features) into sequence-modelling neural networks such as LSTMs [26, 49], so as to learn temporal audio features; and (2) converting raw audio signals to mel-spectrograms and then passing them into 2D CNNs [18, 24, 44], so as to learn visual spectral features. Mao et al. [44] proposed a bi-stage pipeline comprising unsupervised feature learning from mel-spectrograms followed by disentangling affect-salient features to be fed into an SVM classifier. Mirsamadi et al. [49] used a local attention-guided deep RNN to model long-term contextual dependencies among emotionally salient parts of an audio clip for SER. Mansouri et al. [43] proposed a novel cross-modal enhancement approach using spiking neural networks for unsupervised SER, while the authors of [24] designed a complex architecture leveraging both handcrafted MFCC features and mel-spectrograms of audio signals and encapsulated them together to be fed into a 3D CNN for classification. Among recent works in SER, Ibrahim et al. [26] proposed a novel reservoir computing framework using bi-directional RNNs; Kwon et al. [31] employed a simple self-attention module in 2D CNNs for emotion classification from audio mel-spectrograms; Latif et al. [32] explored adversarial domain adaptation for cross-lingual SER. In a rather unique work, Liu et al. [35] introduced a GA-aided reinforcement learning-based approach for SER, mimicking the emotional processing mechanism of the limbic system in the brain.

The need for FS [3, 69] arises in order to alleviate potential redundancies captured by automated feature extraction pipelines, which limit the performance of a model at making accurate predictions. Broadly speaking, FS methods may be categorised as: (1) filter-based methods [30, 37, 53] which employ various scoring metrics based on intra-feature properties to rank features according to their discriminative (or regressive) importance; (2) wrapper-based approaches [4, 20, 50] which involve training a learning model with a subset of features followed by iterative inclusion/exclusion of features based on a heuristic objective; and (3) embedded methods [42, 63, 70] which are combinations of filter and wrapper methods having an intrinsic FS model implemented within itself. It is intuitive that filter methods are computationally cheap as they only explore intra-feature properties without calculating an explicit performance objective, although wrapper methods typically perform better compared to the former as they aim at optimizing a heuristic function modelled upon the task-dependent performance metric [69]. Researchers have also introduced hybrid wrapper-wrapper [4, 59] and wrapper-filter [56, 58] methods to obtain superior performance as compared to using a single algorithm.

FS has been leveraged in several tasks, especially in those where there is a possibility of redundancy among features [18, 21, 34, 66]. For SER tasks, various FS approaches have been employed primarily in association with traditional ML-based methods [13, 27, 34, 66]. Liu et al. [34] proposed a filter-based FS method followed by final classification using a decision tree-like classifier. Yildirim et al. [66] leveraged Cuckoo Search [65] and NSGA-II [15] to perform FS on extracted accoustic features for SER. FS has also been used with a deep learning-based approach by Farooq et al. [18] where the authors optimized the feature space obtained by training CNN models on mel-spectrograms obtained from raw audio signals for emotion classification.

3 Proposed method

In this section, we elaborately describe each stage of the proposed HDFS framework for emotion detection from speech data. Figure 1 shows the overall workflow of the proposed pipeline, where the sequential stages are:

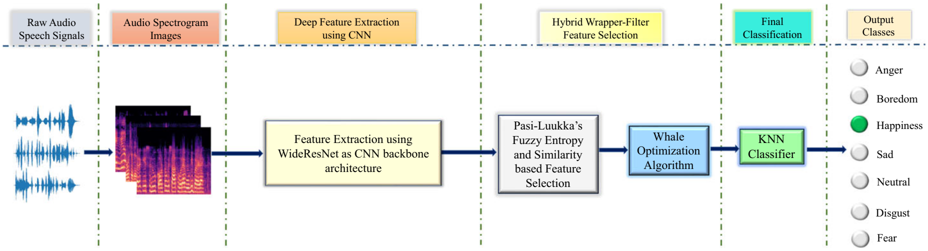

Raw Audio Speech Signals
Audio Spectrogram Images
Deep Feature Extraction using CNN
Hybrid Wrapper-Filter Feature Selection
Final Classification
Output Classes
Feature Extraction using WideResNet as CNN backbone architecture
Pasi-Luukka's Fuzzy Entropy and Similarity based Feature Selection
Whale Optimization Algorithm
KNN Classifier
Anger
Boredom
Happiness
Sad
Neutral
Disgust
Fear

Fig. 1 Overall workflow of the proposed hybrid deep feature selection framework for SER from raw audio signals

∗ Feature extraction using Wide-ResNet-50-2
∗ Filter-based FS using fuzzy entropy and similarity based feature ranking
∗ Wrapper-based FS using WOA
∗ Classification using KNN classifier on the optimal feature subset.

The stages have been explained in detail in the subsequent sections.

3.1 Deep feature extraction using wide-ResNet-50-2

The first stage of our proposed pipeline involves feature extraction from speech mel-spectrograms representing the spatial time-frequency distribution of the audio signal, obtained upon applying fast Fourier transformation (FFT) on the raw speech samples. In this study, a customized Wide-ResNet-50-2 [67] CNN model has been employed to capture a rich feature representation of the mel-spectrogram images for further engineering. The architecture of Wide-ResNet-50-2 is shallower and wider compared to its predecessor, the ResNet [25] family, thereby reducing the training time as well as parameters without compromising on performance and increasing computational efficiency. Further, to ensure the information is extracted effectively from the speech mel-spectrograms, a fully-connected (FC) layer is added after flattening the final pooling layer. The FC layer comprises 512 neurons and is associated with the Rectified Linear Unit (ReLU) activation function, and this is the layer from which the deep features ($dim = 512$) have been extracted. The FC layer reduces loss of information when the feature representation is compressed from 2048 units (i.e. after flattening the output from the last pooling layer) to the classification layer (i.e. having number of neurons equal to the emotion classes in the dataset i.e. N). The final classification layer is associated with the softmax activation function, which maps the outputs to a probability distribution (i.e. values between 0 and 1). A schematic diagram representing the customized CNN architecture has been provided in Fig. 2.

3.2 Filter-based FS: fuzzy entropy and similarity measures

FS is used to improve classification performance in situations when the number of training examples is less than the number of measured features to choose from. Filter methods are used to filter out the undesirable features through checking the consistency of the data present and eliminating repetitive features. The primary objective is to select optimum features subset for an input for a learning algorithm. In this article, we have used a filter based feature selection technique [37] based on similarity and fuzzy entropy measures for feature selection.

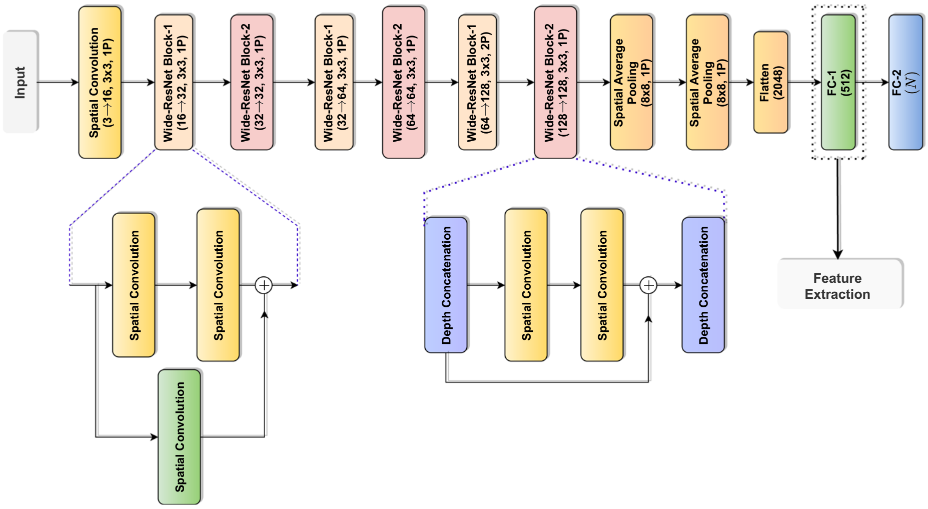

Input
Spatial Convolution (3→16, 3x3, 1P)
Wide-ResNet Block-1 (16→32, 3x3, 1P)
Wide-ResNet Block-2 (32→32, 3x3, 1P)
Wide-ResNet Block-1 (32→64, 3x3, 1P)
Wide-ResNet Block-2 (64→64, 3x3, 1P)
Wide-ResNet Block-1 (64→128, 3x3, 2P)
Wide-ResNet Block-2 (128→128, 3x3, 1P)
Spatial Average Pooling (8x8, 1P)
Spatial Average Pooling (8x8, 1P)
Flatten (2048)
FC-1 (512)
FC-2 (N)
Spatial Convolution
Spatial Convolution
Spatial Convolution
Depth Concatenation
Spatial Convolution
Spatial Convolution
Depth Concatenation
Feature Extraction

Fig. 2 Schematic representation of the customized Wide-ResNet-50-2 CNN architecture used in this study

Fuzziness measures, viz., measure of impreciseness and vagueness, imply how far a given fuzzy set is from a reference set. In this work, we have used the measure of probabilistic entropy (discussed later in this section). While using these fuzzy entropy measures with the similarity classifier, we first define the ideal vectors $v_i = (v_i(f_1), ...v_i(f_t))$ which represent the class i having t features. The calculation of the ideal vector is done based on the generalised mean. Then the similarity between each sample x and ideal vector v is measured by (1). The class of the sample is decided in accordance with the similarity value calculated. If a sample is from class i then the similarity value for the sample S(x, y) will be 1, else it will be 0.

$$S = (1 - |I_{k,i}{}^{p} - x_{j,i}{}^{p}|)^{(1/p)} \tag{1}$$

where, $I_{k,i}$ is the ideal vector for the i^{th} feature of the j^{th} individual sample $x_{j,i}$ belonging to the k^{th} class and p is a parameter from the Łukasiewicz structure [38]. We take $p = 1$ for a normal Łukasiewicz structure.

We have used the following equation to measure the probabilistic entropy,

$$H_1(A) = -\sum_{j=1}^{n}(\mu_A(x_j)\log\mu_A(x_j) + (1-\mu_A(x_j))\log(1-\mu_A(x_j))) \tag{2}$$

where $\mu_A x_j$ are the fuzzy values. We have used this fuzziness measure to evaluate the global deviation from the ordinary sets to see if any crisp set A_0 leads to $h(A_0) = 0$.

Now for a sample, while calculating the similarity values with the ideal vector, we can get j similarity values for j features, which is where we have used fuzzy entropy measures in order to calculate the relevance of the feature. The underlying idea is, while calculating fuzzy entropy values (2) in which $\mu_A(x_j)$ is the similarity value, for higher similarity values, we get lower entropy and if the similarity values are near 0.5 the entropy values are higher. Based on this, we have calculated the entropy values of the similarity values derived from the samples which we want to classify and the ideal vectors calculated initially. We have finally found t entropy values for t features of each sample. After calculating the entropy value for each feature, we use the above mentioned idea to rank the features based on their entropy scores. The primary idea behind ranking is that the feature having the highest entropy would be having the least amount of contribution in the deviation between classes and for more informative features, the entropy values are lower. In the present study, this algorithm has been used to rank features and select a top-$q\%$ subset from the entire set. Experimentally, we have set $q = 50\%$, implying that out of the 512 deep features, the top-ranked 256 features are chosen at this stage A graphical representation and the pseudo-code of the algorithm described above has been provided in Fig. 3 and Algorithm 1 respectively.

```
Input: $I[1, ..., l], x[1, ..., m]$
    for $j$ = 1 to $m$ do
        for $i$ = 1 to $t$ do
            for $k$ = 1 to $l$ do
                $S[j][i][k] = (1 - I[j][i][k]^p - x[i][j]^p)^{(1/p)}$ (using (1))
            end for
        end for
    end for
    Sort similarity values $S[i][j][k]$ according to feature set $U$
    for $i$ = 1 to $t$ do
        $H[i] = -\sum_{x \epsilon U} \mu_i(x) \log \mu_i(x) + (1 - \mu_i(x)) \log (1 - \mu_i(x))$ (using (2))
    end for
    Rank the features based on entropy scores.
    Return ranked feature values.
```

Algorithm 1 Pseudo-code of proposed algorithm for the fuzzy entropy and similarity based filter method for FS where m is the number of samples, t is the number of features and l is the number of classes.

3.3 Wrapper-based FS: whale optimization algorithm

WOA [47] is a meta-heuristic optimization algorithm that uses a spiral to simulate the bubble-net attacking mechanism of the humpback whales who dive into the water and form a bubble-net spiral around their prey and swim up towards the surface. The three stages of the WOA algorithm are as follows: (1) encircling the prey, (2) bubble-net attacking phase (exploitation) and (3) searching for prey (exploration).

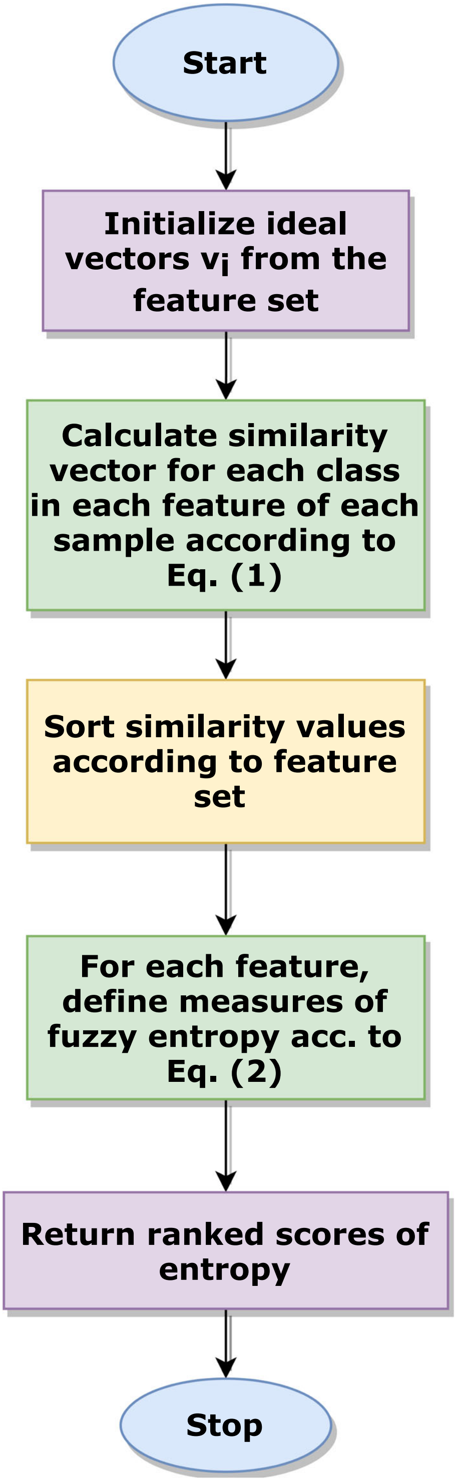
The first stage involves the identification of the best search agent using a fitness function and updating the distance of the other search agents towards the best search agent. The current prey or solution is considered to be closer to the global optimum. In the Bubble-net attacking stage, based on the values of certain constraints, there's a 50% chance between the approaches of shrinking encircling and spiral updating position. The stage of "searching for prey" is an exploration stage where the search agent can search for the prey randomly without opting for the spiral updating positioning.

Equations (3) and (4) are used to update the position of the search agent denoted by position vector X^i, where A and C are coefficient vectors in the t^{th} iteration and X^* is the position vector of the best search agent found.

$$D = |C \cdot X_t^* - X_t^i| \tag{3}$$

$$X_{t+1}^i = X_t^* - A \cdot D \tag{4}$$

Fig. 3 Flowchart for the fuzzy entropy and similarity based filter method for FS used in the present work


Start
Initialize ideal vectors vi from the feature set
Calculate similarity vector for each class in each feature of each sample according to Eq. (1)
Sort similarity values according to feature set
For each feature, define measures of fuzzy entropy acc. to Eq. (2)
Return ranked scores of entropy
Stop

The coefficient vectors A and C are calculated as follows:

$$A = 2a \cdot r - a \tag{5}$$

$$C = 2 \cdot r \tag{6}$$

where, a decreases linearly from 2 to 0 over the iterations in both exploration and exploitation phases and r is a random variable in [0,1]. After each iteration, we have updated the position of the best search agent (X^*) if there is a better solution available or the search agent goes beyond the search space. Declaring r as a random variable allows the search agent to achieve a position in the vicinity of the best search agent and implement encircling the prey.

The shrinking encircling approach in the exploitation phase is implemented by decreasing the value of a in (5). For the spiral updating position approach, we have first calculated the distance (7) between the i^{th} whale and the prey (best solution obtained till current iteration). Then, a spiral is created between the position of the whale and the prey to imitate the movement by the humpback whale. The equations are as follows,

$$D' = |X_t^* - X_t^i| \tag{7}$$

$$X_{t+1}^i = D' \cdot e^{b \times l} \cdot \cos(2\pi l) + X_t^* \tag{8}$$

where, b is a constant for defining the shape of the logarithmic spiral, l is a random number $\in [-1, 1]$, and $(\cdot)$ is element-by-element multiplication.

The approach is decided based on the value of p, a random variable in $[0, 1]$. We have assumed that there is a probability of 50% to choose between the above mentioned approaches of exploitation to update the position of the whale during optimization. The mathematical model is as follows,

$$X_{t+1}^i = \begin{cases} X_t^* - A \cdot D & \text{if} p < 0.5 \\ D^* \cdot e^{b \times l} \times \cos(2\pi l) + X_t^* & \text{if} p \geq 0.5 \end{cases} \tag{9}$$

where, p is a random number $\in [0, 1]$.

The algorithm also involves an exploration phase to allow the agent for a randomised search in the search space. It is used to emphasize on the random search according to the relative position of the agents. We have used the random values of the coefficient vector A to decide for the approach in case of encircling. Random values of A greater than -1 and less than 1 are used to force the search agent to move farther away from the reference whale. For exploration, we use values for which $|A| > 1$. For the randomised search, we have modified (3) and (4) and use a random agent instead of the best search agent,

$$D = |C \cdot X_t^j - X_t^i| \tag{10}$$

$$X_{t+1}^i = X_t^j - A \cdot D \tag{11}$$

where, X_t^j is a random position vector chosen from the current population, C is a coefficient vector (as in (3)), and t is the current iteration.

We have modified the WOA to map the continuous space search of WOA to a binary one in accordance with our problem of feature selection. We have used the S-shaped sigmoid transfer function to do the needful as shown in the following equation:

$$\mathcal{S}(z) = \frac{1}{1 + e^{-z}} \tag{12}$$

The position of destination points X_{t+1} for the i^{th} whale will be updated according to (13).

$$X^i_{t+1} = \begin{cases} 1, & rand() < \mathcal{S}(X^i_t) \\ 0, & otherwise \end{cases} \tag{13}$$

where, $rand()$ yields a random number $\in [0, 1]$.

FS is a multi-objective paradigm with its objectives being: (1) *maximization* of classification accuracy and (2) *minimization* of number of features. Thus, it is evident that the two objectives are opposing to each other. To alleviate this contradiction, we combine them to formulate a heuristic fitness function $\mathcal{F}(\cdot)$ using a weight factor α for the ensemble, thereby reducing the problem to a single-objective optimization task. The expression for the fitness function is shown in (14).

$$\uparrow \mathcal{F} = \alpha \times \eta + (1 - \alpha) \times \Delta \tag{14}$$

where η is the classification accuracy of the feature subset (obtained by KNN classifier), Δ is the feature reduction given by (15), and $\alpha \in [0, 1]$ signifies the relative weight of the classification accuracy and feature reduction. For the present work, we have considered $\alpha = 0.99$ for all experimentation.

$$\Delta = \left(\frac{|F| - |f|}{|F|} \right) \tag{15}$$

where, $|F|$ is the original feature dimension and $|f|$ is the number of features selected. In our work, $|F|$ is the cardinality of the top-$q\%$ feature subset ($= 256$) obtained by filter-guided FS in the previous step.

The main advantages of WOA include wide range of exploration over the search space because of randomised parameters and constraints which improve random agents of the population and search for prey guided by both the best search agent and randomised agents of the population because of randomised parameters which help in the task of exploration. The flowchart for WOA is shown in Fig. 4.

3.4 Classification using KNN classifier

KNN [8] is a simple non-parametric classification algorithm that rely on distance computation as the sole classification criterion. In this algorithm, the training samples of the dataset (i.e. the feature subset selected by WOA) are treated as data points in the embedding space and divided into several distinct classes. Among n data points $\{p_i : 1 \leq i \leq n\}$ in the proposed embedding space, to predict the class of a new instance point p_j, the distances between p_j and all its k-nearest neighbors are computed. Finally, a majority vote of the k points considered decide the class allotment to p_j. Following the recommendations by [40, 41], in this study we have set the value of $k = 5$ for all experimentation.

3.5 Analysis of computational complexity

Although fuzzy entropy and similarity measure based FS algorithm is known to be computationally cost reducing, the general notion in literature regarding wrapper-based FS algorithms is that they are computationally more expensive in comparison with other feature selection algorithms. So it is necessary to be aware of the computational complexities of the algorithms implemented in the proposed method. In this section, we shall discuss the computational complexity of both the implemented algorithms for feature selection and feature optimisation.

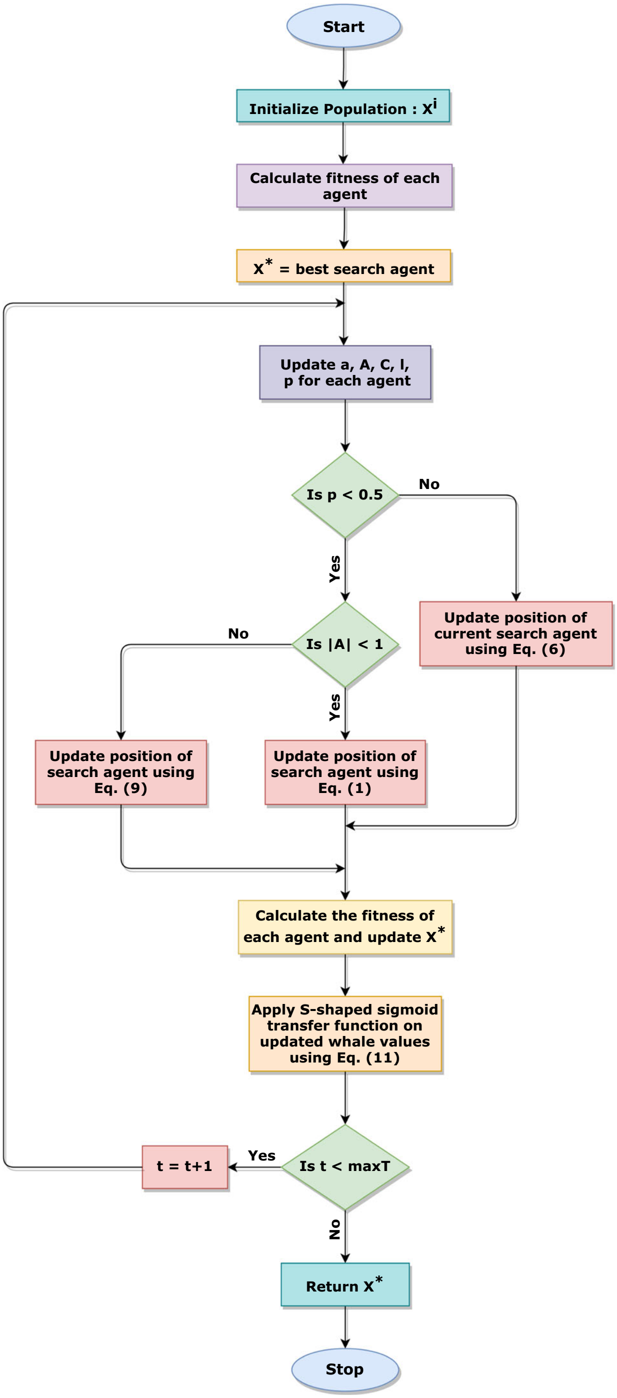

Start
Initialize Population : X^i
Calculate fitness of each agent
X* = best search agent
Update a, A, C, l, p for each agent
Is p < 0.5
No
Yes
Update position of current search agent using Eq. (6)
Is |A| < 1
No
Yes
Update position of search agent using Eq. (9)
Update position of search agent using Eq. (1)
Calculate the fitness of each agent and update X*
Apply S-shaped sigmoid transfer function on updated whale values using Eq. (11)
t = t+1
Yes
Is t < maxT
No
Return X*
Stop

Fig. 4 Schematic diagram of WOA used in the present work. Here, X^i represents the i^{th} member of the population, X^* is the best search agent found, and a, p, l, A and C are random variables and coefficient vectors respectively

3.5.1 Pasi-Luukka's algorithm:

In Pasi-Luukka's algorithm, the parameters which are to be considered for the calculation of computational complexity are: number of classes(l), number of samples(m) and number of features(t). In a single iteration of the algorithm, the similarity values of the features are calculated, using (1), by iterating over the entire feature set and the classes for each feature. Therefore, complexity of one iteration($iter$) is,

$$O(iter) = O(t).O(l) \tag{16}$$

After calculation of similarity values for each sample, the entropy values are calculated using (2), by iterating through the feature set.Thus, the computational complexity of the algorithm is,

$$O(P) = O(m).O(iter) + O(t) \tag{17}$$

that is,

$$O(P) = \Omega(m).O(t).O(l) + O(t) \tag{18}$$

3.5.2 WOA:

In WOA, there are three parameters to be considered for the calculation of computational complexity of the algorithm. They are: number of iterations(t), population size(i), and number of features(j). Now, in one iteration, algorithm iterates over the entire population and each agent is updated through an iteration over the feature set. Therefore, the complexity of one iteration($iter$) shall be,

$$O(iter) = O(i).O(j) \tag{19}$$

Therefore, for a total t iterations, the computational complexity of the WOA algorithm would be,

$$O(W) = O(t).O(iter) \tag{20}$$

or,

$$O(W) = \Omega(t).O(i).O(j) \tag{21}$$

4 Results and discussion

In this section, we describe the SER datasets used to evaluate the proposed framework, as well as report the results obtained on each dataset, using a 5-fold cross-validation scheme. We also compare the proposed approach with some existing methods in literature, to justify the superiority and reliability of the proposed method.

4.1 Datasets used

The proposed framework has been robustly evaluated on three publicly available SER datasets using a five-fold cross-validation scheme:

1. RAVDESS database by Livingstone et al. [36]
2. URDU speech database by Latif et al. [33]
3. EmoDB database by Burkhardt et al. [11]

For each of the aforementioned datasets, a train and validation split of 80% and 20% respectively have been taken to evaluate the proposed pipeline. A brief description of the datasets is provided in the following subsections.

4.1.1 RAVDESS database

The RAVDESS [36] database was originally a multi-modal emotion recognition dataset comprising facial expressions as well as audio samples for speech and music. The dataset was recorded with a North American accent by 24 professional actors (12 females and 12 males) with eight emotions: calm, happy, sadness, angry, fearful, surprise, neutral, and disgust expressions. Overall, RAVDESS contains 1440 speech files for SER, which have been used in the present study. The class-wise distribution of the dataset is given in Table 1.

4.1.2 URDU speech database

The Urdu-language speech emotion database (URDU) was originally proposed in the context of cross-lingual SER [33] which comprises 400 audio samples covering four basic emotions: angry, happy, neutral and sad. The corpus was created using video clips collected from YouTube based on the discussion and situations going on in the talk shows. It is a class-balanced dataset, the distribution given in Table 1.

4.1.3 EmoDB database

The Berlin database of emotional speech (EmoDB) [11] is a German SER database produced by the Technical University of Berlin. It was recorded by 10 actors (5 females and

Table 1 Class-wise distribution of samples in each of the publicly available SER datasets used in this study

Dataset	Class	Emotion label	Number of samples
RAVDESS [36]	0	Angry	192
	1	Calm	192
	2	Disgust	192
	3	Fearful	192
	4	Happy	192
	5	Neutral	96
	6	Sad	192
	7	Surprised	192
URDU [33]	0	Angry	100
	1	Happy	100
	2	Neutral	100
	3	Sad	100
EmoDB [11]	0	Anger	127
	1	Boredom	81
	2	Disgust	79
	3	Fear	46
	4	Happiness	69
	5	Neutral	71
	6	Sadness	62

5 males, between the age of 20 and 35) and covers seven emotion classes: anger, boredom, neutral, disgust, fear, happiness, and sadness, with a total of 535 audio samples. The class-wise distribution of the dataset is given in Table 1.

4.2 Implementation details

The proposed framework has been implemented in Python3 using the PyTorch Toolbox [52] on a 12GB K80 Nvidia GPU. The CNN feature extractor was trained for 100 epochs using the stochastic gradient descent (SGD) [62] optimizer with a learning rate of 0.0005 and momentum = 0.9. All mel-spectrogram images were resized to 224 × 224 before being passed into the CNN backbone, the training batch size being set to 4.

4.3 Evaluation metrics

Four commonly used evaluation measures have been considered in this study to evaluate the proposed framework on the aforementioned multi-class SER datasets, namely, *Accuracy, Precision, Recall* and *F1-Score*. The formulae of these metrics are given in (22), (23), (24) and (25), all of which having derived from a confusion matrix C.

$$Accuracy = \frac{\sum_{i=1}^{N} C_{ii}}{\sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij}} \tag{22}$$

$$Precision_i = \frac{C_{ii}}{\sum_{j=1}^{N} C_{ji}} \tag{23}$$

$$Recall_i = \frac{C_{\imath}ii}{\sum_{j=1}^{N} C_{ij}} \tag{24}$$

$$F1 - Score_i = \frac{2}{\frac{1}{Precision_i} + \frac{1}{Recall_i}} \tag{25}$$

Here, N denotes the number of emotion classes in a given dataset.

4.4 Results

A five-fold cross-validation scheme has been employed for robust and consistent evaluation of the proposed framework on each of the publicly available SER datasets described in Section 4.1. The results obtained on each dataset are discussed in the following sections.

4.4.1 Results on RAVDESS dataset

Table 2 tabulates the results of each evaluation metric, along with their mean and standard deviation (SD) values, obtained by the proposed framework across each fold of the 5-fold cross-validation scheme. Further, the accuracy scores and number of features selected at each stage of our multi-stage pipeline have been shown in Table 3. It can be seen that the method obtains consistent results across the 5-folds of cross-validation, thereby depicting robustness in the approach.

For the feature extraction phase, the Wide-ResNet-50-2 [67] CNN backbone training curve has been shown in Fig. 5, which shows a moderately satisfactory convergence behaviour.

Further, the confusion matrices obtained by the proposed method on each fold of the cross-validation scheme on RAVDESS have been shown in Fig. 6, which essentially

Table 2 Results obtained by the proposed method on each fold of 5-fold cross-validation scheme on RAVDESS dataset

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FS
1	89.24	90.17	88.49	88.82	165
2	90.28	91.01	89.46	89.81	159
3	89.58	90.43	88.82	89.14	172
4	89.24	90.12	88.49	88.79	163
5	90.28	90.33	88.82	89.08	161
Avg±SD	**89.72±0.53**	**90.41±0.36**	**88.82±0.40**	**89.13±0.41**	**164±5**

(Here, FS denotes number of features selected.)

Table 3 Accuracies and number of features obtained at each stage of the proposed framework on five folds of cross-validation on the RAVDESS dataset

Fold	CNN Feature Extraction		Filter Method		WOA	
	Accuracy (%)	No. of Features	Accuracy (%)	No. of Features	Accuracy (%)	No. of Features
1	86.82	512	87.85	256	89.24	165
2	87.15	512	87.50	256	90.28	159
3	86.45	512	87.15	256	89.58	172
4	85.76	512	86.81	256	89.24	163
5	87.15	512	87.50	256	90.28	161
Avg±SD	**86.67±0.58**	**512±0**	**87.36±0.40**	**256±0**	**89.72±0.53**	**164±5**

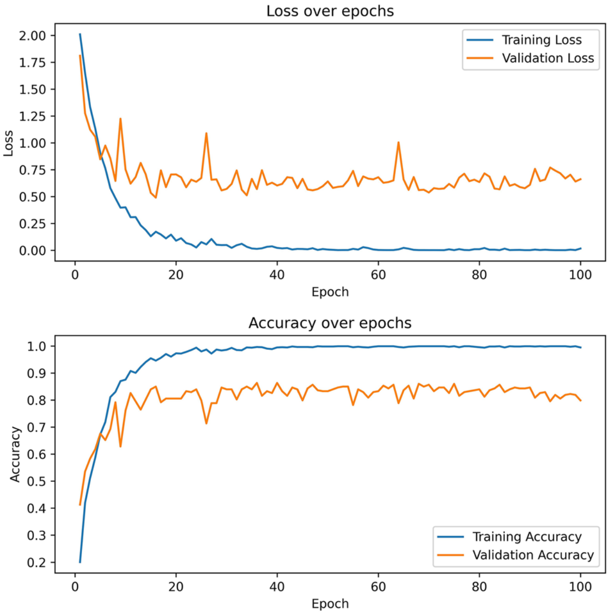
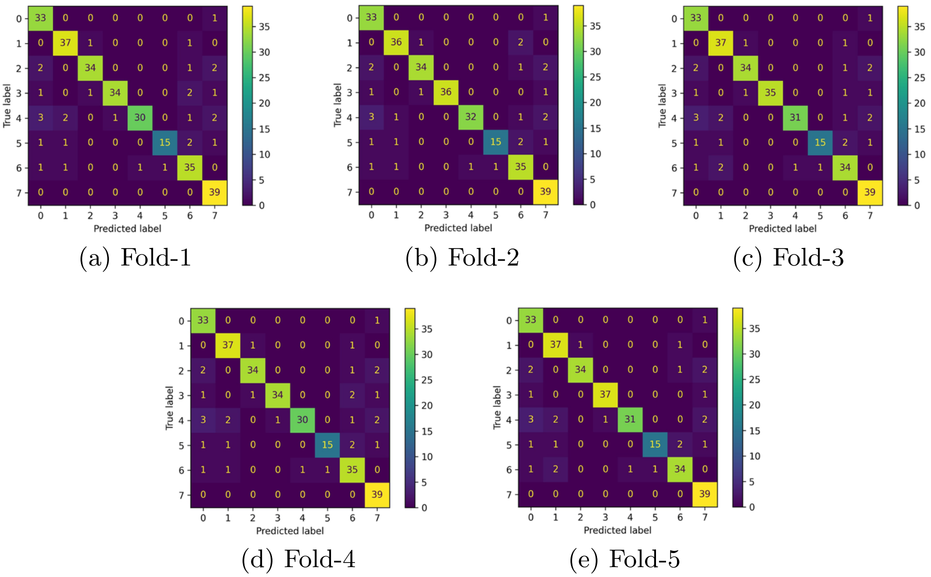


Fig. 5 Learning curves obtained during CNN training for feature extraction on the RAVDESS dataset



(a) Fold-1 (b) Fold-2 (c) Fold-3

(d) Fold-4 (e) Fold-5

Fig. 6 Confusion matrices obtained by the proposed method using 5-fold cross-validation procedure on the RAVDESS dataset

describes the model's performance on each class of the dataset. It can be observed that the proposed pipeline achieves high true positive values on most of the emotion classes consistently across each of the folds of cross-validation, which justifies the robust performance of the model. For a more concise depiction of class-wise performances, Fig. 7 provides the class-wise metric scores averaged over five folds of cross-validation, where the model is found to achieve a perfect accuracy for the emotion "Surprised", in addition to showing a cent per cent precision scores for two emotion classes (i.e. "Fearful" and "Surprised"). The

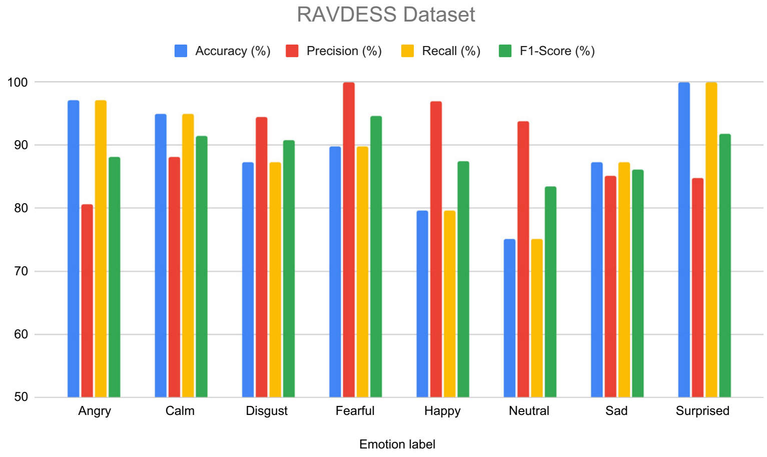


Fig. 7 Class-wise results obtained by the proposed method on the RAVDESS dataset

Table 4 Results obtained by the proposed method on each fold of 5-fold cross-validation scheme on URDU speech dataset

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FS
1	96.25	96.37	96.25	96.31	97
2	96.25	96.45	96.05	96.25	95
3	96.25	96.45	96.05	96.25	101
4	96.25	96.37	96.25	96.31	92
5	96.25	96.37	96.25	96.31	107
Avg±SD	**96.25±0.00**	**96.40±0.04**	**96.17±0.11**	**96.29±0.03**	**98±6**

(Here, FS denotes number of features selected.)

performance of the proposed framework on such a challenging dataset highlight its potential in robust detection of emotion from human speeches.

4.4.2 Results on URDU speech dataset

The performance of the proposed method on the URDU dataset across each fold has been tabulated in Table 4, while the accuracy values obtained at each stage of the pipeline are shown in Table 5. Furthermore, the learning curves obtained during CNN backbone training in Fig. 8 shows a commendable convergence behaviour without any signs of overfitting, something small datasets are highly prone to. A high classification accuracy of 96.25% is obtained by the proposed approach, justifying its effectiveness and suitability for SER.

The class-wise performance of the proposed approach on the URDU speech corpus have been illustrated by the confusion matrices in Fig. 9 and the class-wise metric scores in Fig. 10. Exemplary performance of the proposed framework across the emotion classes can be inferred from the aforementioned figures, including high true positive values and perfect accuracy scores for two emotion classes.

4.4.3 Results on EmoDB dataset

The evaluation metric scores, along with their mean and SD values over five folds of cross-validation obtained by the proposed study on EmoDB [11] have been tabulated in Table 6. It

Table 5 Accuracies and number of features obtained at each stage of the proposed framework on five folds of cross-validation on the URDU speech dataset

Fold	CNN Feature Extraction		Filter Method		WOA	
	Accuracy (%)	No. of Features	Accuracy (%)	No. of Features	Accuracy (%)	No. of Features
1	95.00	512	95.00	256	96.25	97
2	95.00	512	96.25	256	96.25	95
3	96.25	512	96.25	256	96.25	101
4	95.00	512	95.00	256	96.25	92
5	95.00	512	96.25	256	96.25	107
Avg±SD	**95.25±0.56**	**512±0**	**95.75±0.68**	**256±0**	**96.25±0.00**	**98±6**

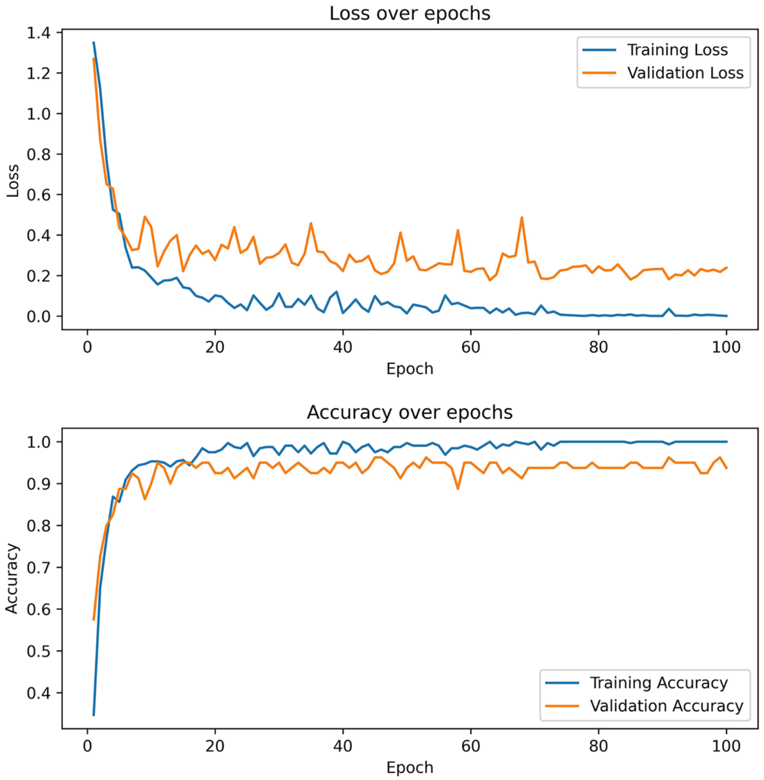
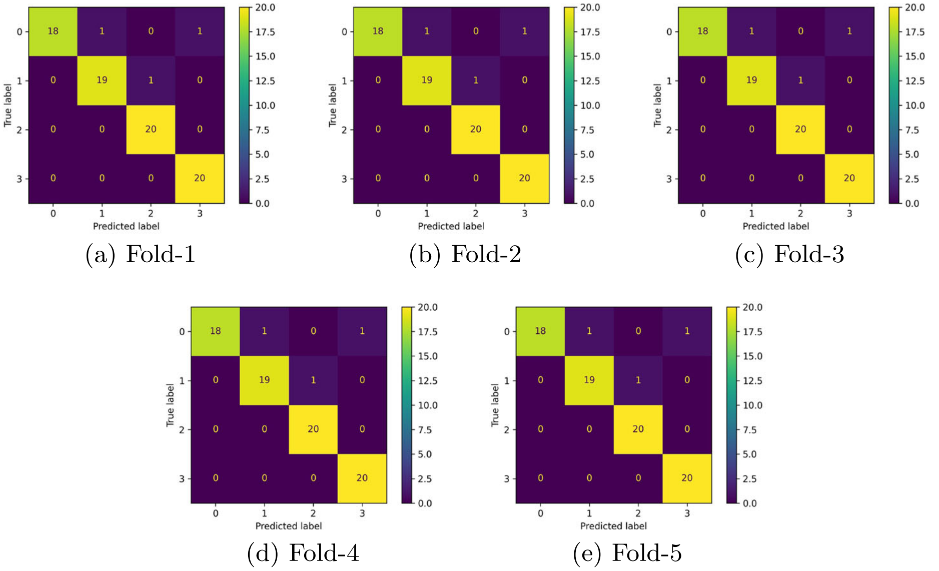


Fig. 8 Learning curves obtained during CNN training for feature extraction on the URDU speech dataset



(a) Fold-1 (b) Fold-2 (c) Fold-3

(d) Fold-4 (e) Fold-5

Fig. 9 Confusion matrices obtained by the proposed method using 5-fold cross-validation procedure on the URDU speech dataset

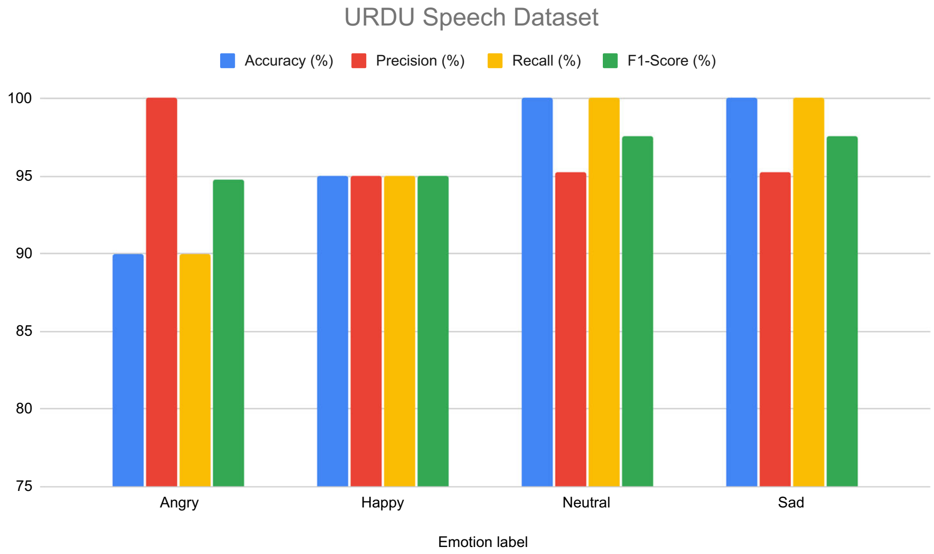


Fig. 10 Class-wise results obtained by the proposed method on the URDU speech dataset

Table 6 Results obtained by the proposed method on each fold of 5-fold cross-validation scheme on EmoDB dataset

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FS
1	93.46	94.32	92.55	93.04	121
2	92.52	93.39	91.99	92.52	116
3	94.39	95.10	93.98	94.28	122
4	93.46	94.32	92.55	93.04	133
5	94.39	95.10	93.98	94.28	129
Avg±SD	**93.64±0.78**	**94.45±0.71**	**93.01±0.91**	**93.43±0.80**	**124±7**

(Here, FS denotes number of features selected.)

Table 7 Accuracies and number of features obtained at each stage of the proposed framework on five folds of cross-validation on EmoDB dataset

Fold	CNN Feature Extraction		Filter Method		WOA	
	Accuracy (%)	No. of Features	Accuracy (%)	No. of Features	Accuracy (%)	No. of Features
1	89.72	512	91.89	256	93.46	121
2	87.81	512	90.28	256	92.52	116
3	90.28	512	92.52	256	94.39	122
4	88.89	512	92.55	256	93.46	133
5	89.72	512	92.52	256	94.39	129
Avg±SD	**89.28±0.96**	**512±0**	**91.95±0.97**	**256±0**	**93.64±0.78**	**124±7**

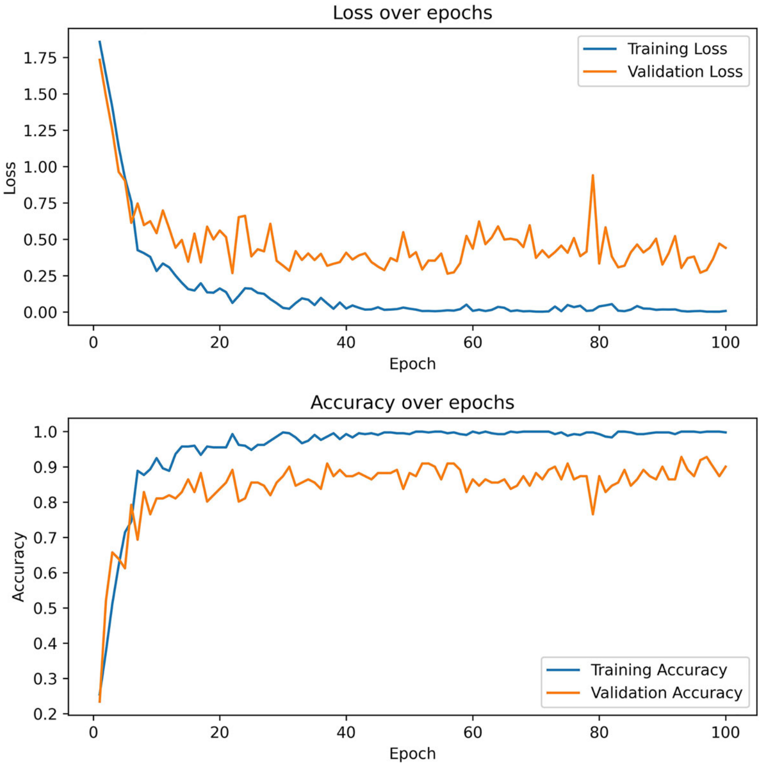


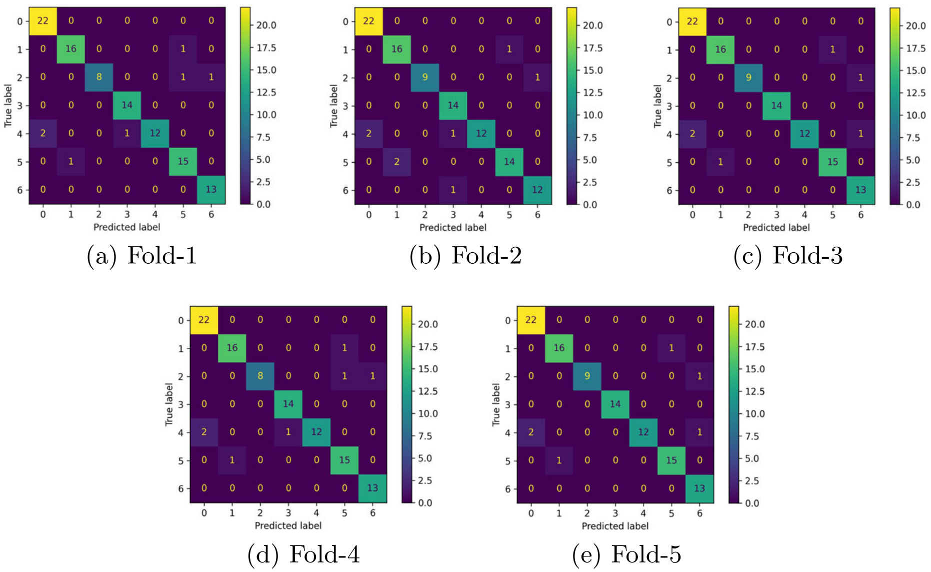
Fig. 11 Learning curves obtained during CNN training for feature extraction on the EmoDB dataset

is observed that our approach achieves a promising mean classification accuracy of 93.64% along with a precision of 94.45%. The accuracies as well as number of features obtained after each stage of the pipeline have been listed in Table 7. The learning curves obtained during training of the Wide-ResNet-50-2 [67] CNN feature extractor is shown in Fig. 11, the convergence behaviour being quite stable, showing very little tendency to overfit the dataset. The results are a testimony of the faithful performance of our method.

Finally, the confusion matrices depicting class-wise performance of the proposed pipeline obtained on each fold of the cross-validation procedure on EmoDB are shown in Fig. 12, while Fig. 13 depicts the average metric values of each emotion class. It is observed that our approach achieves a perfect classification accuracy on two emotion classes (“Angry” and “Sadness”), as well as obtaining a perfect precision score on two other emotions (“Disgust” and “Happiness”). The results further validate the robustness of the proposed approach for SER tasks.

4.5 Comparison with state-of-the-art SER methods

Table 8 compares the proposed method against several works in literature pertaining to SER on the publicly available datasets used in this study based on the evaluation measures described in Section 4.3. It can be observed that the proposed framework outperforms all of the existing works on RAVDESS [36] and URDU [33] datasets by significant margins. On the EmoDB [11] dataset, the proposed pipeline achieves a performance equivalent to



(a) Fold-1 (b) Fold-2 (c) Fold-3

(d) Fold-4 (e) Fold-5

Fig. 12 Confusion matrices obtained by the proposed method using 5-fold cross-validation procedure on the EmoDB dataset

state-of-the-art, outperforming several existing works in literature. It may also be noted that several previous works have reported accuracy as the sole metric, which does not give any insights regarding the false positives (or true negatives) and hence is insufficient as well as unreliable on a multi-class classification task such as SER. On the other hand, our results justify that the proposed study is a highly effective approach for detecting emotions from speech signals.

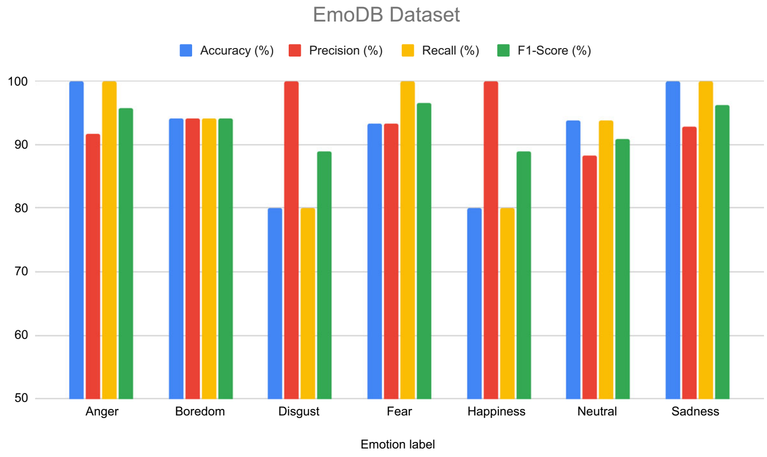


Fig. 13 Class-wise results obtained by the proposed method on the EmoDB dataset

Table 8 Comparison of the proposed method with existing works in literature

Dataset	Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RAVDESS	Mansouri et al. [43]	83.60	–	–	–
	Bhavan et al. [10]	75.69	–	–	–
	Farooq et al. [18]	81.30	–	–	–
	Tuncer et al. [64]	87.43	87.22	87.43	87.30
	Kwon et al. [31]	80.00	–	–	–
	Kanwal et al. [27]	82.50	–	–	–
	Ibrahim et al. [26]	85.68	86.05	85.68	85.79
	Proposed HDFS	**89.72**	**90.41**	**88.82**	**89.13**
URDU	Latif et al. [33]	83.40	–	–	–
	Latif et al. [32]	67.30	–	–	–
	Zehra et al. [68]	60.00	60.00	60.00	60.00
	Ancilin et al. [9]	95.25	–	95.25	–
	Proposed HDFS	**96.25**	**96.40**	**96.17**	**96.29**
EmoDB	Dansehfar et al. [13]	82.82	–	–	–
	Bhavan et al. [10]	92.45	–	–	–
	Farooq et al. [18]	95.10	–	–	–
	Tuncer et al. [64]	90.09	91.05	89.47	90.17
	Kwon et al. [31]	93.00	–	–	–
	Kanwal et al. [27]	89.65	–	–	–
	Ibrahim et al. [26]	91.64	93.38	91.64	92.34
	Proposed HDFS	**93.64**	**94.45**	**93.01**	**93.43**

The FS algorithm used in this study, WOA [47], has been compared to the following state-of-the-art meta-heuristics in literature:

1. Particle Swarm Optimization (PSO) by Kennedy et al. [28]
2. Arithmetic Optimization Algorithm (AOA) by Abualigah et al. [2]

Table 9 Comparison of accuracies and number of features selected (FS) among state-of-the-art optimization algorithms on each of the SER datasets

Optimization	RAVDESS		URDU		EmoDB	
Algorithm	Accuracy (%)	FS	Accuracy (%)	FS	Accuracy (%)	FS
PSO [28]	87.85	185	95.00	167	90.28	138
AOA [2]	87.50	197	95.00	173	89.72	201
GWO [48]	88.89	124	96.25	105	90.27	101
GSA [55]	87.85	176	96.25	108	92.52	99
CSA [65]	87.50	138	95.00	141	91.89	149
SCA [46]	88.89	119	96.25	127	92.52	107
WOA [47]	**89.72**	**164**	**96.25**	**98**	**93.64**	**124**

The results reported are aggregated over 10 independent runs of each algorithm, averaged over five folds of cross-validation

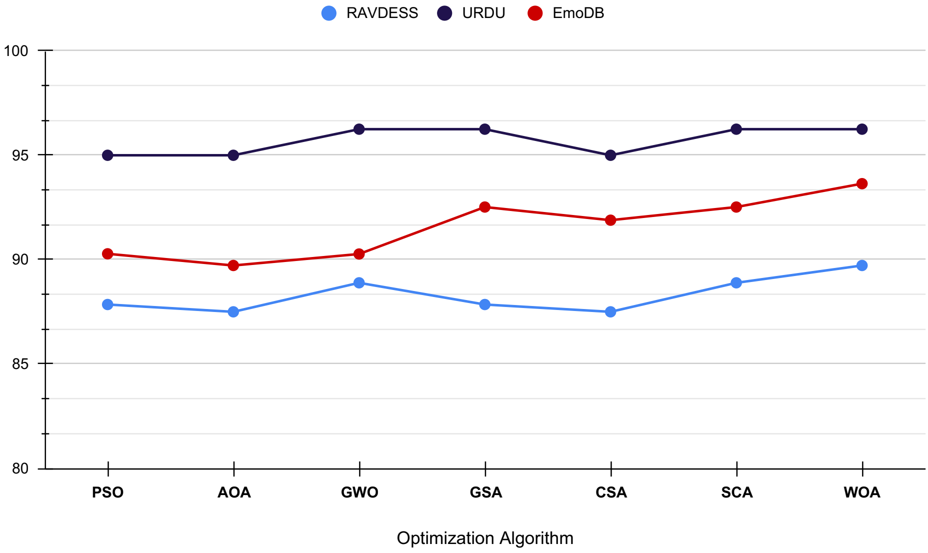


Fig. 14 Comparison of accuracies obtained by state-of-the-art optimization algorithms on each of the SER datasets. The results reported are aggregated over 10 independent runs of each algorithm, averaged over five folds of cross-validation

3. Grey Wolf Optimizer (GWO) by Mirjalili et al. [48]
4. Gravitational Search Algorithm (GSA) by Rashedi et al. [55]
5. Cuckoo Search Algorithm (CSA) by Yang et al. [65]
6. Sine Cosine Algorithm (SCA) by Mirjalili et al. [46]

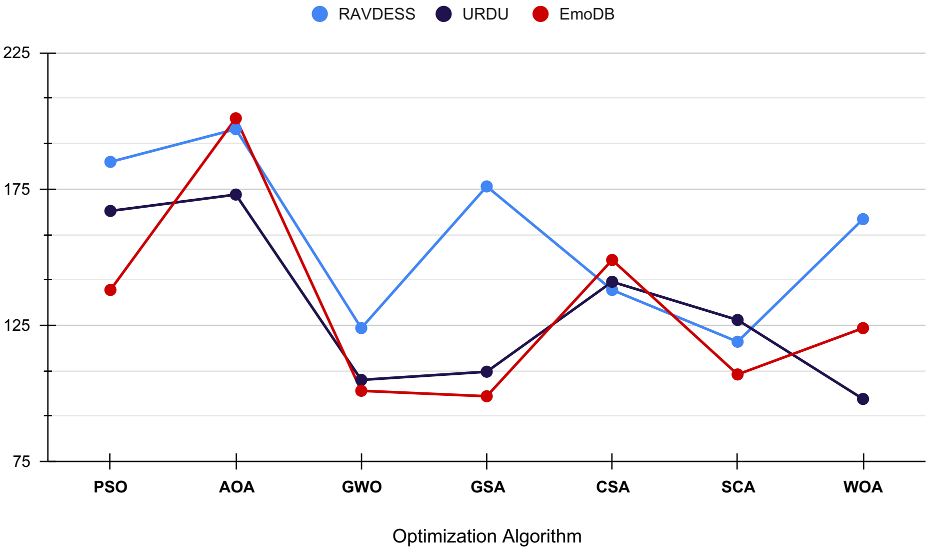


Fig. 15 Comparison of number of features selected by state-of-the-art optimization algorithms on each of the SER datasets. The results reported are aggregated over 10 independent runs of each algorithm, averaged over five folds of cross-validation

For each optimization algorithm, a population size of 40 is chosen and the maximum number of iterations is set to 100. The average values over 10 independent runs on a given fold of a dataset, aggregated by running over five folds of cross-validation, have been reported in Table 9. In each of the datasets, it can be observed that WOA achieves the highest classification accuracy and also shows competitive performance in terms of feature space reduction. On RAVDESS [36], SCA and GWO rank second in terms of the classification metric, while SCA is found to select the minimal number of features. On the URDU dataset [33], GWO, GSA and SCA perform equally as WOA in terms of accuracy, although the latter selects the minimal feature subset. On the EmoDB dataset [11], GSA and SCA rank second in terms of accuracy, with the former showing the greatest feature reduction. Note that all of these experiments have been conducted on the top-$q\%$ feature subset ($q = 50\%$) obtained by the filter-based FS method [37] as described in Section 3.

A graphical view of the aforementioned comparison in terms of classification accuracy and number of features selected have been depicted in Figs. 14 and 15 respectively. The plots show that WOA has shown robust performance in optimising both of the aforesaid objectives for each of the SER datasets, justifying the use of the same in our proposed study.

5 Conclusion and future work

The present study proposes a computationally efficient two-tier hybrid wrapper-filter FS pipeline for dimensionality reduction of the feature representation extracted by a CNN backbone from mel-spectrograms of speech audio clips, as well as robust classification of speech signals into respective emotion classes. Our approach alleviates the cumbersome process of handcrafted feature extraction, providing an end-to-end framework for SER. The proposed method has been evaluated on three publicly available standard speech datasets, where it has been found to outperform several existing works in literature, justifying the reliability of the framework. The hybrid dimensionality reduction approach used in this study is a new addition to FS literature and thus, can be used as a stand-alone algorithm for traditional ML-based approaches requiring feature engineering. Further, the proposed pipeline is domain-independent and hence may be applied off-the-shelf to different facets of image classification, such as disease detection [12] or human action recognition [23], to name a few.

In order to contribute to the research on SER, we intend to explore other speech datasets available in the public domain for greater generalization and reliability so as to be used in real-world applications. We may also try various other approaches to meta-heuristic algorithm-based FS, such as initialization using clustering-guided population [22], hybrid of wrapper-based approaches [17] and local search-embedded optimization algorithms [12]. Last but not the least, we also intend to explore temporal features of raw audio signals using deep learning-based architectures to investigate deeper into emotion classification and further the community.

Data Availability Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of Interests All the authors declare that there are no conflicts of interest in the publication of this paper.

References

1. Abbaschian BJ, Sierra-Sosa D, Elmaghraby A (2021) Deep learning techniques for speech emotion recognition, from databases to models. Sensors
2. Abualigah L, Diabat A, Mirjalili S, Abd Elaziz M, Gandomi AH (2021) The arithmetic optimization algorithm. Comput Methods Appl Mech Eng 376:113609
3. Agrawal P, Abutarboush HF, Ganesh T, Mohamed AW (2021) Metaheuristic algorithms on feature selection: a survey of one decade of research (2009-2019). IEEE Access 9:26766–26791
4. Ahmed S, Ghosh KK, Garcia-Hernandez L, Abraham A, Sarkar R (2021) Improved coral reefs optimization with adaptive β-hill climbing for feature selection. Neural Comput & Applic 33(12):6467–6486
5. Akçay MB, Oğuz K (2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Comm 116:56–76
6. Albornoz EM, Milone DH, Rufiner HL (2011) Spoken emotion recognition using hierarchical classifiers. Comput Speech & Lang 25(3):556–570
7. Alghowinem S, Goecke R, Wagner M, Epps J, Gedeon T, Breakspear M, Parker G (2013) A comparative study of different classifiers for detecting depression from spontaneous speech. In: ICASSP. IEEE
8. Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician
9. Ancilin J, Milton A (2021) Improved speech emotion recognition with mel frequency magnitude coefficient. Appl Acoust 179:108046
10. Bhavan A, Chauhan P, Shah RR et al (2019) Bagged support vector machines for emotion recognition from speech. Knowl-Based Syst 184:104886
11. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B et al (2005) A database of german emotional speech. In: Interspeech, vol 5, pp 1517–1520
12. Chattopadhyay S, Kundu R, Singh PK, Mirjalili S, Sarkar R (2021) Pneumonia detection from lung x-ray images using local search aided sine cosine algorithm based deep feature selection method. International Journal of Intelligent Systems, pp 1–38
13. Daneshfar F, Kabudian SJ (2020) Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. Multimed Tools Appl 79(1):1261–1289
14. Danisman T, Alpkocak A (2008) Emotion classification of audio signals using ensemble of support vector machines. In: International tutorial and research workshop on perception and interactive technologies for speech-based systems. pp 205–216. Springer
15. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: nsga-ii. IEEE Trans Evol Comput 6(2):182–197
16. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on computer vision and pattern recognition. pp 248–255. IEEE
17. Dey A, Chattopadhyay S, Singh PK, Ahmadian A, Ferrara M, Sarkar R (2020) A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition. IEEE Access 8:200953–200970
18. Farooq M, Hussain F, Baloch NK, Raja FR, Yu H, Zikria YB (2020) Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. Sensors 20(21):6008
19. Fragopanagos N, Taylor JG (2005) Emotion recognition in human–computer interaction. Neural Netw 18(4):389–405
20. Ghosh KK, Ahmed S, Singh PK, Geem ZW, Sarkar R (2020) Improved binary sailfish optimizer based on adaptive β-hill climbing for feature selection. IEEE Access 8:83548–83560
21. Ghosh S, Hassan S, Khan AH, Manna A, Bhowmik S, Sarkar R (2021) Application of texture-based features for text non-text classification in printed document images with novel feature selection algorithm. Soft Computing, pp 1–19
22. Guha R, Ghosh M, Chakrabarti A, Sarkar R, Mirjalili S (2020) Introducing clustering based population in binary gravitational search algorithm for feature selection. Appl Soft Comput 93:106341
23. Guha R, Khan AH, Singh PK, Sarkar R, Bhattacharjee D (2021) Cga: a new feature selection model for visual human action recognition. Neural Comput & Applic 33(10):5267–5286
24. Hajarolasvadi N (2019) Demirel, h.: 3d cnn-based speech emotion recognition using k-means clustering and spectrograms. Entropy 21(5):479
25. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778
26. Ibrahim H, Loo CK, Alnajjar F (2021) Speech emotion recognition by late fusion for bidirectional reservoir computing with random projection. IEEE Access 9:122855–122871

27. Kanwal S, Asghar S (2021) Speech emotion recognition using clustering based ga-optimized feature set. IEEE Access 9:125830–125842
28. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks. vol 4, pp 1942–1948. IEEE
29. Khalil RA, Jones E, Babar MI, Jan T, Zafar MH, Alhussain T (2019) Speech emotion recognition using deep learning techniques: A review. IEEE Access 7:117327–117345
30. Kononenko I (1994) Estimating attributes: analysis and extensions of relief. In: European conference on machine learning. pp 171–182. Springer
31. Kwon S et al (2021) Att-net: enhanced emotion recognition system using lightweight self-attention module. Appl Soft Comput 102:107101
32. Latif S, Qadir J, Bilal M (2019) Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition. In: 2019 8Th international conference on affective computing and intelligent interaction (ACII). pp 732–737. IEEE
33. Latif S, Qayyum A, Usman M, Qadir J (2018) Cross lingual speech emotion recognition: urdu vs. western languages. In: 2018 International conference on frontiers of information technology (FIT). pp 88–93. IEEE
34. Liu ZT, Wu M, Cao WH, Mao JW, Xu JP, Tan GZ (2018) Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing 273:271–280
35. Liu ZT, Xie Q, Wu M, Cao WH, Mei Y, Mao JW (2018) Speech emotion recognition based on an improved brain emotion learning model. Neurocomputing 309:145–156
36. Livingstone SR, Russo FA (2018) The ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in north american english. PloS one
37. Luukka P (2011) Feature selection using fuzzy entropy measures with similarity classifier. Expert Syst Appl 38(4):4600–4607
38. Luukka P, Saastamoinen K, Kononen V (2001) A classifier based on the maximal fuzzy similarity in the generalized lukasiewicz-structure. In: 10Th IEEE international conference on fuzzy systems. pp 195–198. IEEE
39. Machado PP, Beutler LE, Greenberg LS (1999) Emotion recognition in psychotherapy: impact of therapist level of experience and emotional awareness. Journal of Clinical Psychology
40. Mafarja MM, Mirjalili S (2017) Hybrid whale optimization algorithm with simulated annealing for feature selection. Neurocomputing
41. Mafarja M, Qasem A, Heidari AA, Aljarah I, Faris H, Mirjalili S (2020) Efficient hybrid nature-inspired binary optimizers for feature selection. Cognitive Computation
42. Maldonado S, López J. (2018) Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for svm classification. Applied Soft Computing
43. Mansouri-Benssassi E, Ye J (2019) Speech emotion recognition with early visual cross-modal enhancement using spiking neural networks. In: 2019 International joint conference on neural networks (IJCNN). pp 1–8. IEEE
44. Mao Q, Dong M, Huang Z, Zhan Y (2014) Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans Multimed 16(8):2203–2213
45. Meftah IT, Le Thanh N, Amar CB (2012) Detecting depression using multimodal approach of emotion recognition. In: 2012 IEEE International conference on complex systems (ICCS). IEEE
46. Mirjalili S (2016) Sca: a sine cosine algorithm for solving optimization problems. Knowledge-based systems 96:120–133
47. Mirjalili S, Lewis A (2016) The whale optimization algorithm. Adv Eng Softw 95:51–67
48. Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. Adv Eng Softw 69:46–61
49. Mirsamadi S, Barsoum E, Zhang C (2017) Automatic speech emotion recognition using recurrent neural networks with local attention. In: ICASSP. IEEE
50. Nguyen BH, Xue B, Zhang M (2020) A survey on swarm intelligence approaches to feature selection in data mining. Swarm Evol Comput 54:100663
51. Ooi CS, Seng KP, Ang LM, Chew LW (2014) A new approach of audio emotion recognition. Expert Syst Appl 41(13):5858–5869
52. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:8026–8037
53. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions in Pattern Analyis and Machine Intelligence
54. Ramakrishnan S, El Emary IM (2013) Speech emotion recognition approaches in human computer interaction. Telecommun Syst 52(3):1467–1478

55. Rashedi E, Nezamabadi-Pour H, Saryazdi S (2009) Gsa: a gravitational search algorithm. Inf Sci 179(13):2232–2248
56. Sarkar SS, Sheikh KH, Mahanty A, Mali K, Ghosh A, Sarkar R (2021) A harmony search-based wrapper-filter feature selection approach for microstructural image classification. Integr Mater Manuf Innov 10(1):1–19
57. Schipor OA, Pentiuc SG, Schipor MD (2011) Towards a multimodal emotion recognition framework to be integrated in a computer based speech therapy system. In: 2011 6Th conference on speech technology and human-computer dialogue (sped). IEEE
58. Sen S, Saha S, Chatterjee S, Mirjalili S, Sarkar R (2021) A bi-stage feature selection approach for covid-19 prediction using chest ct images. Applied Intelligence, pp 1–16
59. Sheikh KH, Ahmed S, Mukhopadhyay K, Singh PK, Yoon JH, Geem ZW, Sarkar R (2020) Ehhm: electrical harmony based hybrid meta-heuristic for feature selection. IEEE Access 8:158125–158141
60. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
61. Song P, Zheng W (2018) Feature selection based transfer subspace learning for speech emotion recognition. IEEE Trans Affect Comput 11(3):373–382
62. Sutskever I, Martens J, Dahl G, Hinton G (2013) On the importance of initialization and momentum in deep learning. In: International conference on machine learning. pp 1139–1147. PMLR
63. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol
64. Tuncer T, Dogan S, Acharya UR (2021) Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. Knowledge-Based Systems
65. Yang XS, Deb S (2009) Cuckoo search via lévy flights. In: 2009 World congress on nature & biologically inspired computing (naBIC). IEEE
66. Yildirim S, Kaya Y, Kılıç F (2021) A modified feature selection method based on metaheuristic algorithms for speech emotion recognition. Appl Acoust 173:107721
67. Zagoruyko S, Komodakis N (2016) Wide residual networks. arXiv:1605.07146.
68. Zehra W, Javed AR, Jalil Z, Khan HU, Gadekallu TR (2021) Cross corpus multi-lingual speech emotion recognition using ensemble learning. Complex & Intelligent Systems, pp 1–10
69. Zhang R, Nie F, Li X, Wei X (2019) Feature selection with multi-view data: a survey. Inf Fusion 50:158–167
70. Zhang H, Zhang R, Nie F, Li X (2018) A generalized uncorrelated ridge regression with nonnegative labels for unsupervised feature selection. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp 2781–2785. IEEE.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.