applied sciences

MDPI

Article

A Feature Selection Algorithm Based on Differential Evolution for English Speech Emotion Recognition

Liya Yue [1], Pei Hu [2], Shu-Chuan Chu [3] and Jeng-Shyang Pan [3,4,*]

[1] Fanli Business School, Nanyang Institute of Technology, Nanyang 473004, China; 3172027@nyist.edu.cn
[2] School of Computer and Software, Nanyang Institute of Technology, Nanyang 473004, China; huxiaopei163@163.com
[3] College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China; scchu0803@gmail.com
[4] Department of Information Management, Chaoyang University of Technology, Taichung 413310, Taiwan
* Correspondence: jengshyangpan@gmail.com

Abstract: The automatic identification of emotions from speech holds significance in facilitating interactions between humans and machines. To improve the recognition accuracy of speech emotion, we extract mel-frequency cepstral coefficients (MFCCs) and pitch features from raw signals, and an improved differential evolution (DE) algorithm is utilized for feature selection based on K-nearest neighbor (KNN) and random forest (RF) classifiers. The proposed multivariate DE (MDE) adopts three mutation strategies to solve the slow convergence of the classical DE and maintain population diversity, and employs a jumping method to avoid falling into local traps. The simulations are conducted on four public English speech emotion datasets: eNTERFACE05, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Surrey Audio-Visual Expressed Emotion (SAEE), and Toronto Emotional Speech Set (TESS), and they cover a diverse range of emotions. The MDE algorithm is compared with PSO-assisted biogeography-based optimization (BBO_PSO), DE, and the sine cosine algorithm (SCA) on emotion recognition error, number of selected features, and running time. From the results obtained, MDE obtains the errors of 0.5270, 0.5044, 0.4490, and 0.0420 in eNTERFACE05, RAVDESS, SAVEE, and TESS based on the KNN classifier, and the errors of 0.4721, 0.4264, 0.3283 and 0.0114 based on the RF classifier. The proposed algorithm demonstrates excellent performance in emotion recognition accuracy, and it finds meaningful acoustic features from MFCCs and pitch.

Keywords: speech emotion recognition; feature selection; differential evolution; mutation

Citation: Yue, L.; Hu, P.; Chu, S.-C.; Pan, J.-S. A Feature Selection Algorithm Based on Differential Evolution for English Speech Emotion Recognition. *Appl. Sci.* **2023**, *13*, 12410. https://doi.org/10.3390/app132212410

Academic Editor: Mirka Saarela, Lilia Georgieva and Vili Podgorelec

Received: 12 October 2023
Revised: 13 November 2023
Accepted: 14 November 2023
Published: 16 November 2023

1. Introduction

Emotions play an important role in human interaction [1]. Speech is the most natural form of human expression and communication [2,3]. Therefore, the automatic recognition of speech signals by computing devices is considered a concern [4,5]. Words and messages are often combined to express a person's emotions [6,7]. There are two important sources of information in a speech signal: (a) an explicit source containing linguistic content, and (b) an implicit source carrying vocal cues and non-verbal elements about speakers [8,9].

Speech emotion recognition (SER) is an essential component of modern artificial intelligence-based systems [10,11]. For instance, identifying the emotions of customers or drivers can lead to adaptive responses. In healthcare and education, SER has the potential to monitor patients' and students' emotional states, and aids in diagnosing conditions such as depression, anxiety, or engagement. However, it is not an easy task in real life to categorize happiness, sadness, anger, fear, disgust, surprise, and neutral emotions from speech [12,13]. People express emotions differently across cultures and individuals, and they also convey mixed emotions. The main difficulty lies in extracting meaningful and optimal features from speech signals [14].

The characteristic of SERs is the high dimensionality of features, not all of which are correlated [15]. Many efforts have been made to improve the performance of emotional state recognition in speech through feature selection. The main aim of feature selection is to choose the most important acoustic features, which can reduce the computational cost of SERs and improve their recognition accuracy [16–18].

Researchers have applied various features to recognize emotional states, but emotion recognition is still a challenging issue. It is hard to connect speech features to specific emotions due to the lack of theoretical support, while the effectiveness of SERs is determined by the features extracted from speech signals that must be invariant to speakers and their languages. Over the years, people have utilized mel-frequency cepstral coefficients (MFCCs) to obtain acoustic features. MFCCs are essential because they capture the spectral characteristics of human speech and approximate the non-linear human auditory perception of sounds. MFCCs bridge the gap between speech's acoustic properties and its emotional content, and they provide a concise and informative representation of the spectral details in speech. Although these features carry important information about audio signals, it should be noted that the performance of recognition algorithms also subsequently decreases as the length and sampling rate of audio signals increase, requiring more calculations for analysis. DE is known for its robustness and simplicity in handling complex optimization problems [19,20], and it is a reliable choice for SER. In this study, we investigate DE to recognize speech emotion through feature selection, and the main contributions of this paper are summarized as follows:

(1) We introduce a system for extracting acoustic features.
(2) We propose an improved DE to implement feature selection.
(3) The proposed mutation strategies in DE are essential for enhancing exploration and exploitation. It is possible to achieve better convergence and exploration of the speech emotion space. The jumping method introduced in DE improves global search ability and escapes local optima.
(4) We validate the performance of the proposed algorithm with eNTERFACE05, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), the Surrey Audio-Visual Expressed Emotion (SAEE), and the Toronto Emotional Speech Set (TESS). The algorithm provides more accurate and efficient speech emotion recognition, and it extends applications in areas such as human–computer interaction, sentiment analysis, and emotional well-being assessment.

The structure of this paper is organized as follows. Section 2 introduces the related works of speech emotion recognition. Sections 3 and 4 describe the materials used and the proposed algorithm. Section 5 represents the experimental results with discussions, and Section 6 provides the conclusions.

2. Related Works

Yogesh et al. utilized a hybrid optimization algorithm, BBO_PSO, for emotion and stress recognition from natural speech [21]. Additionally, they employed higher-order spectral features in conjunction with the hybrid approach. These features capture the high statistical characteristics of speech signals, and they have been shown to be effective in obtaining subtle variations in speech related to emotions and stress.

Shahin et al. presented a novel approach to improve the performance of SER systems for both Arabic and English languages [22]. The research focuses on developing an efficient feature selection method using the grey wolf optimizer (GWO) algorithm, which aims to identify the most relevant features from speech signals for accurate emotion recognition. To develop an agent-independent speech emotion/stress recognition system, Yogesh et al. identified the speaker's emotion from speech where features are acquired from the OpenSmile toolbox and high-order spectral features [23]. They proposed a novel particle swarm optimization (PSO)-assisted biogeography for feature selection. Butta utilized an ensemble technique that combines multiple algorithms through cat swarm optimization

(CSO) [24]. This ensemble approach is designed to harness the collective intelligence of different algorithms to improve the accuracy and robustness of emotion classification.

Akinpelu and Viriri also utilized pre-trained deep neural networks to extract high-level features from speech data. Transfer learning allows the model to transfer knowledge learned from a source domain (general speech data) to the target domain (emotion-specific speech data), and enhances the model's ability to generalize to new and unseen emotion samples [25]. The study emphasizes the importance of robustness in speech emotion classification, so the model performs well under different speakers, noise levels, and recording environments. Feature selection and deep transfer learning techniques are intended to improve the model's robustness.

Depression is a prevalent mental health condition that is difficult to accurately diagnose. However, speech analysis has shown promise as a potential non-invasive and cost-effective method for depression detection. Kaur et al. proposed a novel approach that combines speech analysis with a quantum whale optimization algorithm (QWOA) for feature selection [26]. Gideon et al. recognized emotional expressions during natural phone conversations [27], and they specifically investigated individuals with recent suicidal ideation. By analyzing emotion patterns, the research examines if this vulnerable group has unique emotional expressions compared to individuals who have not recently had suicidal thoughts. Gharsellaoui et al. proposed a new algorithm combining DE and linear discriminant analysis (LDA) to design an efficient feature selection and classification model [28]. Auditory features are provided as input for a DE-LDA-based ESR system.

Although the aforementioned works have produced impressive recognition results, the binary optimization characteristics of SER are not considered when using evolutionary algorithms and feature selection. The global search and local search of evolutionary algorithms cannot be well balanced. This paper proposes a new SER model that utilizes multivariate DE to balance the exploration and exploitation in feature selection, and improves recognition accuracy.

3. Materials

In our SER system, we first pass audio through a pre-emphasis filter. Next, we extract mel-frequency cepstral coefficients (MFCCs) and pitch features with framing, windowing, and Fourier transform (FT) techniques. Finally, we utilize DE to select the most relevant features from acoustic features, and we also employ KNN and RF to perform classification tasks. The overview of the proposed system is given in Figure 1.

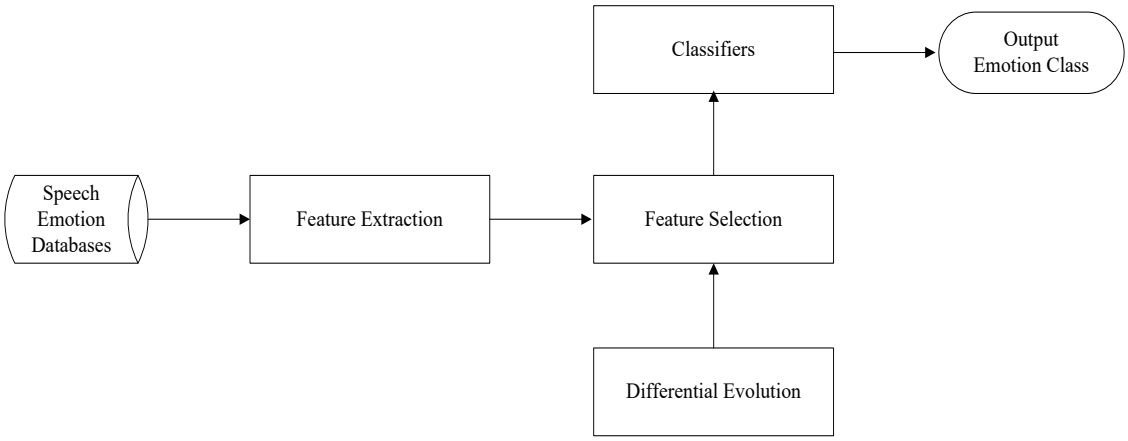
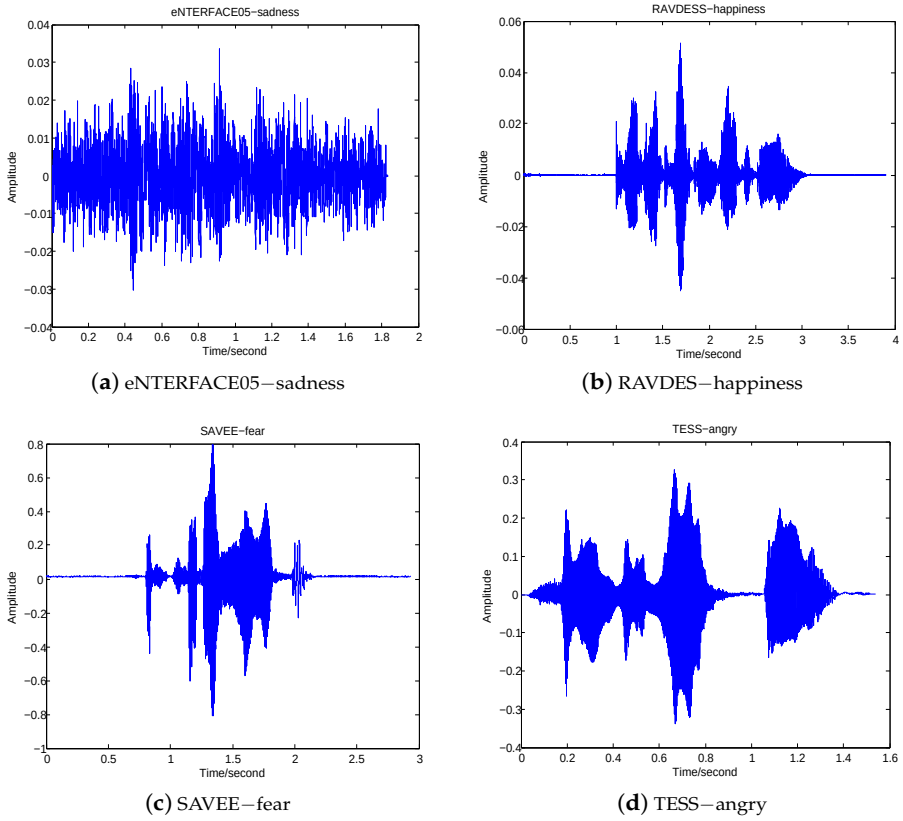


Figure 1. The scheme of the proposed system.

3.1. Dataset Description

In this paper, the eNTERFACE05, RAVDESS, SAEE, and TESS databases are utilized to evaluate feature selection algorithms, and Figure 2 presents the samples of signals.



(**a**) eNTERFACE05−sadness (**b**) RAVDES−happiness

(**c**) SAVEE−fear (**d**) TESS−angry

Figure 2. The samples of signals.

3.1.1. eNTERFACE05

The eNTERFACE05 database, developed as part of the eNTERFACE'05 project, comprises audio, video, and physiological signals recorded from participants. The emotions of these participants contain anger, happiness, sadness, surprise, disgust, and fear. Participants are asked to listen carefully to a short story and immerse themselves in the scene. They can read, memorize, and pronounce (one at a time) five utterances presented, which constitute different responses to a given situation.

3.1.2. Ryerson Audio-Visual Database of Emotional Speech and Song

In the RAVDESS database, there are 7356 audio and video clips, and each one lasts around 3 to 5 s. A total of 24 professional actors (12 male and 12 female) are present in these recordings, and each actor uses various language styles, including calm, angry, neutral, sad, and more.

3.1.3. Surrey Audio-Visual Expressed Emotion

SAVEE has 480 audio and video clips, with 60 recordings representing every emotion. Four male British English speakers who display seven different emotions, including anger, happiness, disgust, sadness, surprise, fear, and neutral.

3.1.4. Toronto Emotional Speech Set

TESS consists of audio recordings of emotional expressions acted by North American English speakers. This database includes 200 audio clips, and each represents distinct emotions: anger, disgust, fear, happiness, surprise, sadness, and neutrality. These emotions are conveyed through short sentences spoken in a neutral tone. In TESS, the emotional expressions are portrayed by actors. It proves particularly valuable for investigating acoustic features and patterns associated with different emotions in speech.

3.2. Feature Extraction

In this study, we extract MFCCs and pitch features from raw audios, and Figure 3 describes their steps. Pitch features contain 11 values (per sample window of 25 ms), including the maximum, minimum, median, mean and variance of each pitch, their corresponding derivatives, and spurt length. MFCC features have 130 values, including the maximum, minimum, median, mean and variance of each coefficient, and their corresponding derivatives.

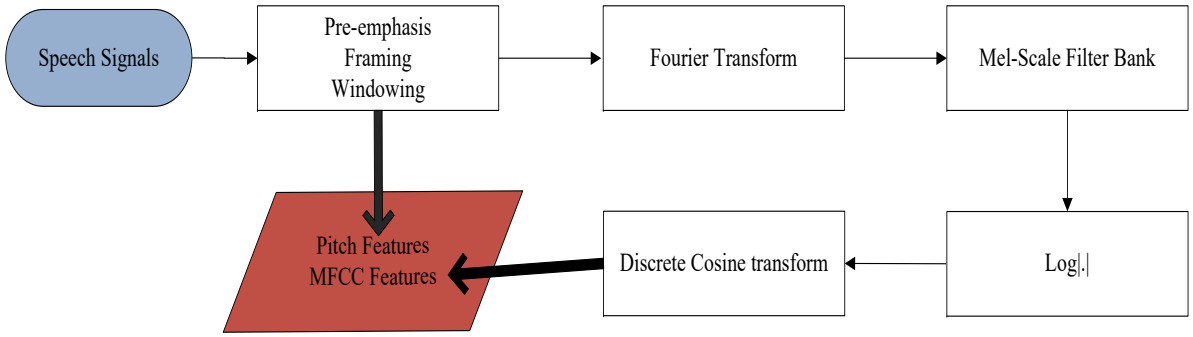


Figure 3. Steps involved in feature extraction.

3.2.1. Pre-Emphasis

In audio communication systems, pre-emphasis is often applied to audio signals before they are transmitted or recorded, and pre-emphasis signals are then de-emphasized on the receiving end to restore the original frequency. This technique enhances the clarity of speech signals and makes them easier to understand. Equation (1) demonstrates how to apply the pre-emphasis filter to a signal $x(t)$.

$$y(t) = x(t) - \alpha x(t-1) \tag{1}$$

where α is set to 0.97.

3.2.2. Framing

Framing divides a continuous audio signal into smaller segments (frames). Each frame typically consists of a fixed number of audio samples or time points, and the frames are usually overlapping to capture temporal information in signals. By breaking continuous signals into frames, we can extract useful information from each segment, and analyze it separately.

3.2.3. Windowing

Windowing is a key step in preparing audio signals for further analysis, and more accurate and meaningful results are obtained when using Fourier-transform-based methods. The most common window functions used in audio processing are the Hamming window, Hanning window, and Blackman window.

3.2.4. Pitch Features

Pitch features are important elements extracted from speech signals, and they provide information about the tonal characteristics of the human voice. Pitch can actually be defined as the repeat rate of complex signals in the autocorrelation function. The pitch is relatively stable when a person is calm. The pitch frequency increases when a person is happy or angry, while it decreases when a person is depressed.

3.2.5. Fourier Transform

Fourier transform is a mathematical technique used to analyze signals and data in the frequency domain. It transforms a signal from the time domain where it is represented as a sequence of amplitude values over time, into the frequency domain, where it is represented as a combination of sinusoidal waves with different frequencies.

3.2.6. Mel-Scale Filter Bank

The mel-scale filter bank extracts MFCCs from audio signals [29]. MFCCs represent the short-term power spectrum of sound. They are widely used in speech and audio processing tasks, because they capture important characteristics of the sound that are relevant to human perception.

3.2.7. Discrete Cosine Transform (DCT)

DCT converts a sequence of data points (such as audio samples or image pixels) from the time or spatial domain to the frequency domain. It achieves this by expressing data as a linear combination of cosine functions with different frequencies and amplitudes.

3.2.8. Mel-Frequency Cepstral Coefficients

MFCCs are the most popular features for recognizing human speech. In 1980, Davis and Mermelstein brought a representation of the approximate structure of the human vocal tract system in which MFCCs accurately describe the system's shape in the short-time power spectrum.

First, the Hamming window splits speech signals into frames of 25 ms with an overlap of 10 ms, and then a fast Fourier transform is utilized to acquire the power spectrum of each frame. Finally, DCT is applied to the logarithmically transformed spectrum to obtain MFCCs. The entire frequency range is divided into n mel filter banks, as shown in Equation (2).

$$c(n) = \sum_{k=1}^{K} (logS_k) cos[n(k - \frac{1}{2})\frac{\pi}{K}] \quad (2)$$

where s_k denotes the output of the k-channel filter bank, and n represents the index of mel cepstral coefficients.

4. Methodology

The problems of DE are premature convergence to local optima and fixed control parameters. It is necessary to make additional improvements to achieve better performance before using it in feature selection. An improved DE proposed in this study adopts three different mutation strategies to maintain population diversity during optimization, and thus balances exploration and exploitation. Figure 4 is the flowchart of the proposed multivariate DE (MDE).

In feature selection and SER, classification accuracy is the main indicator for evaluating algorithms. Consequently, it is used as the objective function in MDE, as shown in Equation (3):

$$fit = \frac{\sum_{i=1}^{10} error_i}{10} \quad (3)$$

where $error_i$ represents the classification error of the i-th cross validation, and we employ 10-fold cross validation in here.

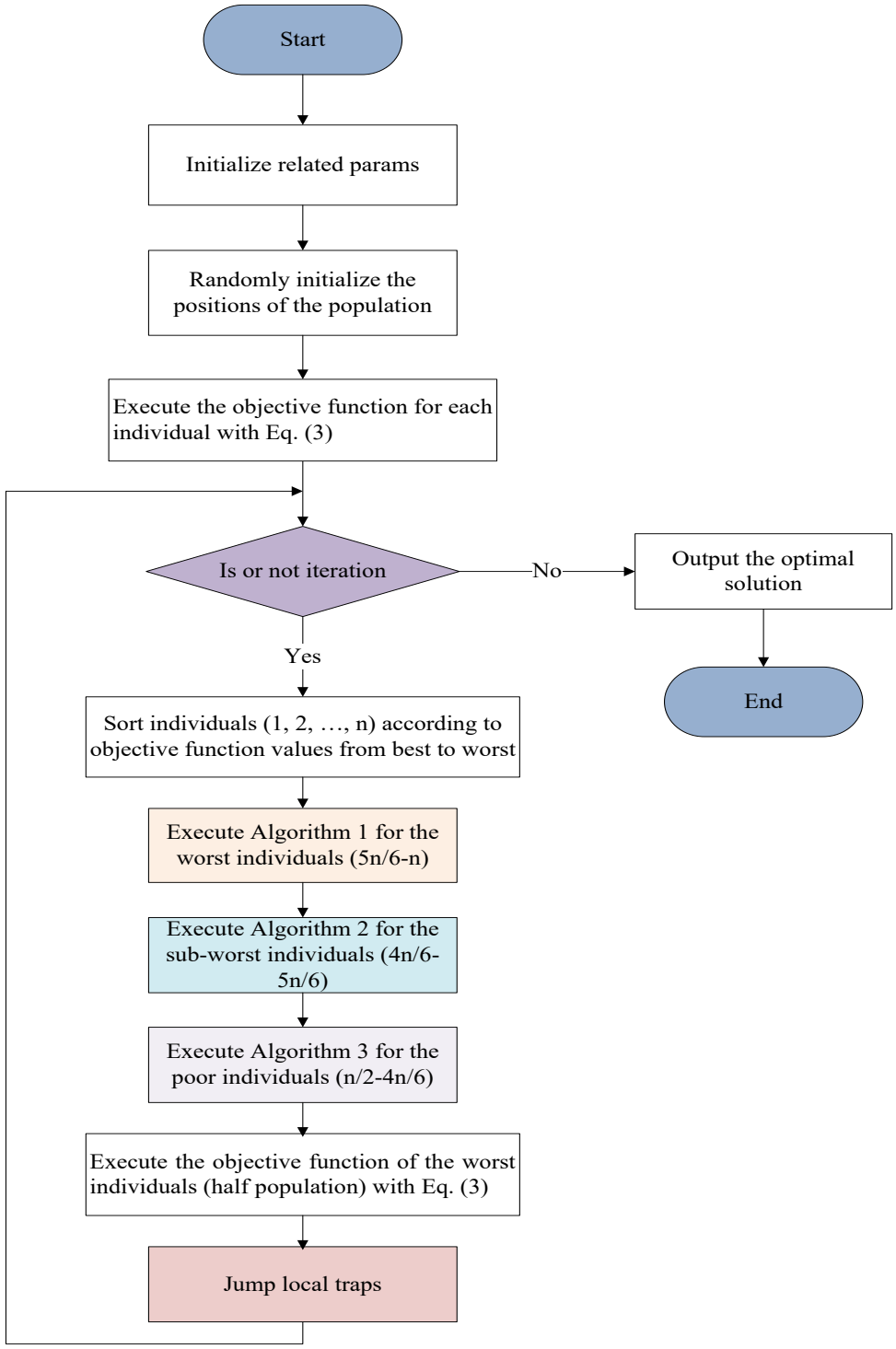


Figure 4. The flowchart of MDE.

The performance of DE is affected by both crossover and mutation, which generates a trial candidate solution. If the randomly selected learning solutions are not within the optimal region, they will mislead some individuals to approach them. MDE only allows individuals with poor objective function values (half of the population) to participate in position update. Individuals with great performance do not update their positions; instead, they serve as exemplars.

4.1. Mutation Strategies

The worst individuals (candidate solutions) learn from the optimal and sub-optimal solutions, and the newly generated solutions are mainly dominated by the optimal solution. The solutions participating in the selection are all superior to candidate solutions. To improve convergence, the new solutions do not execute crossover after mutation, but they directly compare with candidate solutions. Algorithm 1 describes the mutation scheme of the worst individuals. These individuals learn from excellent solutions, and their positions are mainly controlled by the global optimal solution a, which increases the convergence ability of the algorithm.

The mutation method of sub-worst solutions also randomly selects a group of distinct individuals. Unlike the random differential mutation approach, it employs the best individual from this group as the basis for differential mutation. Meanwhile, the other individuals with lower performance contribute to generating vector differences. The update method for sub-worst solutions is similar to Algorithm 1, but it will perform a crossover to enhance the population's diversity, as depicted in Algorithm 2.

Algorithm 1: The mutation method of the worst individuals

```
% k is the index of individual i after sorting.
% Position means the positions of individuals.
% a, b and c are randomly selected individuals for the crossover operator.
% if the fitness value of z is better than x, it will replace x.
k = index(n − i + 1) ;
x = Position(k,:) ;
A = randperm(nPop/2);
A(A == index(1)) = [];
a = index(1);
b = index(A(2)) ;
c = index(A(3));
z = Position(a,:)+beta.*(Position(b,:)-Position(c,:));
for j = 1 : dim do
    if z(j) > rand then
        z(j) = Position(a,j);
    end
    else
        if rand > 0.5 then
            z(j) = Position(b,j);
        end
        else
            z(j) = Position(c,j);
        end
    end
end
```

Algorithm 2: The mutation method of sub-worst individuals

```
k = index(n − i + 1) ;
x = Position(k,:) ;
A = randperm(nPop/2);
A(A == index(1)) = [];
a = index(1);
b = index(A(2));
c = index(A(3));
y = Position(a,:)+beta.*(Position(b,:)-Position(c,:));
for j = 1 : dim do
    if y(j) > rand then
        y(j) = Position(a,j);
    end
    else
        if rand > 0.5 then
            y(j) = Position(b,j);
        end
        else
            y(j) = Position(c,j);
        end
    end
end
z = zeros(1,dim);
j0 = randi([1 numel(x)]);
for j = 1:numel(x) do
    if j == j0 || rand <= pCR then
        z(j) = y(j);
    end
    else
        z(j) = x(j);
    end
end
```

The poor solutions learn from more exemplars to explore more space. They are not only controlled by the global optimal solution, but also affected by other solutions. The mutation increases the chance of learning from more solutions, and improves the exploration ability of the algorithm. In fact, the difference among them is not significant, so crossover is considered from expanding the diversity of the population, as described in Algorithm 3. It has excellent exploration.

Algorithm 3: The mutation method of poor individuals

```
k = index(n − i + 1) ;
x = Position(k,:) ;
A = randperm(nPop/2);
A(A == index(1)) = [];
a = index(1);
b = index(A(2));
c = index(A(3));
d = index(A(4)) ;
e = index(A(5)) ;
y = Position(a,:)+beta.*(Position(b,:)-Position(c,:))+beta.*(Position(d,:)-Position(e,:))
 ;
for j = 1 : dim do
    if y(j) > rand & rand > 0.75 then
        y(j) = Position(a,j);
    end
    else
        Execute the roulette strategy to determine the value of y(j) from b, c, d or e ;
    end
end
z = zeros(1,dim);
j0 = randi([1 numel(x)]);
for j = 1:numel(x) do
    if j == j0 || rand <= pCR then
        z(j) = y(j);
    end
    else
        z(j) = x(j);
    end
end
```

4.2. Jumping Method

It can be seen from MDE that elite individuals influence the search of the population. When a solution is too excellent, they will quickly converge to this position. If the solution is a local optimum, they may fall into a trap, leading the population to lose diversity. Elite individuals are forced to leave their positions and search for other space if the global optimal solution is not updated after ten iterations, as shown in Equation (4).

$$X_i^j = \begin{cases} 1 - X_i^j & if(rand <= 2 * i/nPop) \\ X_i^j & else \end{cases} \tag{4}$$

where $nPop$ is the population size, j is the dimension, and i represents the i-th elite individual according to the sorting order. This method allows most individuals to have the opportunity to escape local traps, and also allows several individuals to continue searching around their positions.

5. Experimental Results and Analysis

5.1. Approaches Used for Comparisons

To validate the superiority of the proposed MDE, the classification performance is compared with two previous works, DE [28] and BBO_PSO [21], and a metaheuristic algorithm SCA [30]. BBO_PSO is a new hybrid PSO-assisted biogeography-based optimization for emotion recognition, and SCA is a sine cosine algorithm for feature selection. Table 1 provides additional information concerning the algorithms. *beta_min* and *beta_max* repre-

sent the lower and upper bounds of the scaling factor, and the most popular strategy set *beta* in DE by using a Gaussian distribution with a mean of 0.5 and a standard deviation of 0.3. These values are consistent with the settings of SaDE [31], and they are beneficial to producing small and large search step sizes. *pCR* means the crossover probability, *KeepRate* is rate of kept habitats, *pMutation* is the mutation probability, *w* is the inertia weight, and *c*1 and *c*2 are learning factors. *thres* is a threshold value.

Table 1. The main parameters setting.

Algorithm	Main Parameters
DE	beta_min = 0.2; beta_max = 0.8; pCR = 0.2;
BBO_SCA	KeepRate = 0.2; pMutation = 0.1; w = 0.9; c1 = 2; c2 = 2;
SCA	thres = 0.5;
MDE	beta_min = 0.2; beta_max = 0.8; pCR = 0.2;

The algorithms adopt Equation (3) as their objective function. The maximum objective functions of the algorithms are set to 2000, and this process is repeated 20 times with a population size of 20. We apply the Wilcoxon rank-sum and Friedman tests to determine if there are significant differences in the experimental results in which a significance level of 0.05 is chosen.

5.2. Experimental Analysis

KNN and RF are adopted as classifiers, where *K* is 5, and the Euclidean distance is selected as the computational method for data points. The number of decision trees is set to 100, and the splitting criterion of decision trees is the Gini index (*gdi*), which reflects the influence of a certain feature on the classification results. All data serve as samples, and the data are randomly divided into 10 parts through 10-fold cross validation. One of them is used for testing, while the other parts are used for training. We obtain the final average recognition error after repeating ten times, and the bold font indicates that a algorithm has obtained the optimal solution.

5.2.1. Simulation Results on the K-Nearest Neighbor Classifier

Figure 5 displays the experimental results using the KNN classifier, and it shows the average, maximum, and minimum errors obtained from each independent run.

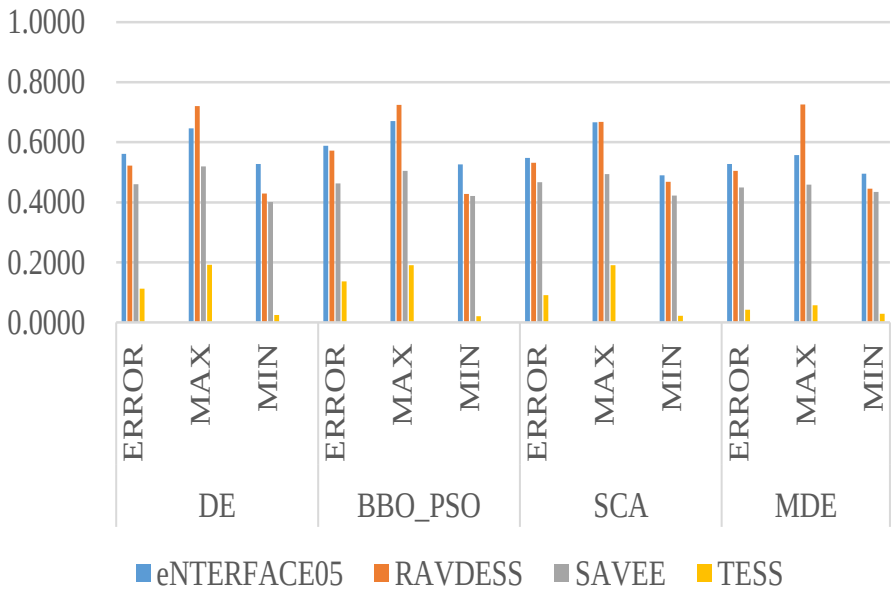


Figure 5. The classification errors of the compared algorithms based on KNN.

It is evident from the figure that MDE excels DE, BBO_PSO, and SCA by achieving errors of 0.5270, 0.5044, 0.4490, and 0.0420 in eNTERFACE05, RAVDESS, SAVEE, and TESS. Regarding the maximum error, MDE outperforms DE, BBO_PSO, and SCA in eNTERFACE05, SAVEE, and TESS, while SCA beats DE, BBO_PSO, and MDE in RAVDESS. In terms of the minimum error, DE and SCA perform well in SAVEE and eNTERFACE05, respectively, while BBO_PSO has the best performance in RAVDESS and TESS. It can be found that the data obtained by MDE have excellent stability, which is especially suitable for speech emotion recognition.

The Wilcoxon rank-sum test reveals that the algorithms have similar experimental results in SAVEE (as shown in Table 2), and it cannot distinguish the experimental results of DE and MDE in RAVDESS. DE, BBO_PSO, SCA, and MDE perform well on two, one, one, and four datasets, respectively. The Friedman test exhibits that their average ranks are 2.5, 3, 75, 2.75, and 1, with $p < 0.05$. Table 2 proves that MDE is superior to other algorithms.

Table 2. The non-parametric statistical analysis of the compared algorithms based on KNN.

	DE	BBO_PSO	SCA	MDE
>/=/<	0/2/2	0/1/3	0/1/3	4/0/0
Rank	2.5	3.75	2.75	1
p-Value	2.56×10^{-2}			

Table 3 illustrates the number of selected features and the running time of the algorithms. MDE obtains the least number of selected features and the shortest running time in eNTERFACE05 and TESS, while SCA outperforms DE, BBO_PSO, and MDE in RAVDESS and SAVEE. Their running time in eNTERFACE05 and SAVEE is the lowest, but they spend a lot of time in RAVDESS and TESS. MDE uses the fewest features to complete recognition in eNTERFACE05 and TESS, while it obtains more features than other algorithms in RAVDESS and SAVEE. As can be seen from Figure 5, the recognition accuracy of MDE is better than DE, BBO_PSO, and SCA. From the number of selected features and classification errors obtained by the algorithms, it can be concluded that using more or fewer features is not beneficial for emotion prediction.

Based on the above discussion, the proposed MDE exhibits the best performance in classification accuracy and running time, and it is suitable for English speech emotion recognition.

Table 3. The number of selected features and the running time (seconds) of the compared algorithms based on KNN.

Dataset	DE		BBO_PSO		SCA		MDE	
	Length	Time	Length	Time	Length	Time	Length	Time
eNTERFACE05	65.85	274.2602	65.05	259.1495	66.15	255.6974	**15.85**	**221.556**
RAVDESS	72.35	626.1439	67.75	660.178	**31.8**	**357.7027**	74.2	666.1531
SAVEE	66.35	294.514	66.9	267.3497	**28.2**	**238.9834**	68.55	374.099
TESS	70.5	1666.3398	68.55	1726.7875	73.45	1837.7258	**28.8**	**832.355**

5.2.2. Simulation Results on the Random Forest Classifier

Figure 6 displays the experimental results using the RF classifier, and it shows the average, maximum, and minimum errors acquired from each independent run.

The errors obtained with RF are superior to the values obtained by KNN. Figure 6 illustrates that MDE has the best results in eNTERFACE05, RAVDESS, SAVEE, and TESS. Its errors in the four emotion datasets are 0.4721, 0.4264, 0.3283, and 0.0114, and it outperforms DE, BBO_PSO, and SCA. In the maximum error, MDE performs well in RAVDESS, SAVEE, and TESS, while DE beats BBO_PSO, SCA, and MDE in eNTERFACE05. Concerning the minimum error, MDE exhibits the best performance in eNTERFACE05, SAVEE, and TESS,

and SCA outperforms DE, BBO_PSO, and MDE in RAVDESS. The data obtained by the RF classifier certify that the performance of MDE is stable, and it can be used for English speech emotion recognition.

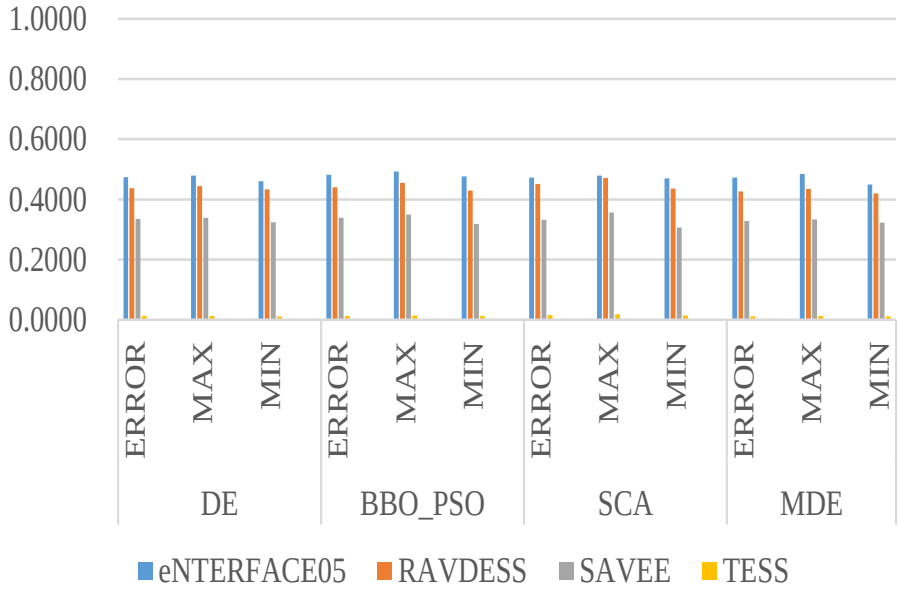


Figure 6. The classification errors of the compared algorithms.

Table 4 presents their non-parametric statistical analysis. Through the Wilcoxon rank-sum test, DE, BBO_PSO, SCA, and MDE perform well on three, two, two, and four datasets, respectively. MDE is superior to the other algorithms, while BBO_PSO and SCA exhibit comparable performance in RF. DE and MDE yield similar results in eNTERFACE05, SAVEE, and TESS, and the Wilcoxon rank-sum test cannot distinguish the experimental data of BBO_PSO, SCA, and MDE in eNTERFACE05 and TESS. Their average ranks are 2.5, 3.5, 3, and 1, and the Friedman test reveals that MDE wins first place, followed by DE, SCA, and BBO_PSO. Table 4 confirms the superiority of MDE in speech emotion recognition.

Table 4. The non-parametric statistical analysis of the compared algorithms.

	DE	BBO_PSO	SCA	MDE
>/=/<	0/3/1	0/2/2	0/2/2	4/0/0
Rank	2.5	3.5	3	1
p-Value	3.84×10^{-2}			

Table 5 illustrates the number of selected features and the running time of the algorithms. The RF classifier provides them with a greater number of features and a longer running time than the KNN classifier. SCA performs well in the number of selected features and running time, and it uses 28.8, 26.4, 20.4, and 38.6 features for classification in the four datasets, respectively. MDE, DE, and BBO_PSO employ approximately half of the features to accomplish emotion recognition. The number of features obtained by MDE in eNTERFACE05 and TESS is smaller than DE, while DE performs better than MDE and BBO_PSO in RAVDESS and SAVEE. The algorithms have more running time in eNTERFACE05 and SAVEE, while they have less time in RAVDESS and TESS.

From the experimental results, it can be noticed that although the algorithms obtain different results on KNN and RF classifiers, MDE consistently performs the best. RF has a higher time complexity than KNN, but it utilizes more features to achieve excellent recognition results.

Table 5. The number of selected features and the running time (seconds) of the compared algorithms.

Dataset	DE		BBO_PSO		SCA		MDE	
	Length	**Time**	**Length**	**Time**	**Length**	**Time**	**Length**	**Time**
eNTERFACE05	69.4	9999.6274	68	10,775.6739	**28.8**	**7119.838**	68.6	11,094.3004
RAVDESS	66.6	34,730.7615	69.6	32,983.0796	**26.4**	**22,223.1923**	68.6	36,696.1692
SAVEE	66.2	11,116.0668	66.6	12,783.4402	**20.4**	**9221.0034**	68.8	11,135.3497
TESS	71	34,910.0227	69.8	39,412.2563	**38.6**	**29,448.3799**	65.8	41,026.9059

6. Conclusions

In SER, researchers focus on identifying significant emotional features through feature selection; however, searching for the optimal features is impractical due to its high complexity. In this study, we use an improved differential evolution to classify English languages from speech signals based on MFCCs and pitch features. Compared with DE, BBO_PSO, and SCA, the experimental results and non-parametric statistical analysis in the four English speech emotion datasets illustrate that MDE achieves excellent recognition accuracy and reduces the number of selected features. The proposed algorithm works with three mutation strategies and a jumping method to balance global search and local search, and improves the accuracy of speech emotion by reconstructing input speech data with relevant and meaningful acoustic features. As speech emotion recognition becomes increasingly vital in various applications, including human–computer interaction, virtual assistants, and the ability to quickly and effectively process and analyze emotions. Our work provides a foundation for enhancing the applicability of such systems.

Emotion recognition is a multifaceted task. To further enhance the robustness and accuracy of our proposed algorithm, it is important to consider integrating other modalities, such as facial expression analysis. This multimodal approach can provide a more comprehensive understanding of users' emotional state. Additionally, our algorithm can be applied to different languages, which is an important avenue for exploration.

Author Contributions: Conceptualization, L.Y. and P.H.; Formal analysis, L.Y. and S.-C.C.; Methodology, L.Y., S.-C.C. and J.-S.P.; Software, L.Y. and P.H.; Writing—Original draft preparation, L.Y.; Writing—Review and editing, P.H., S.-C.C. and J.-S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Henan Provincial Philosophy and Social Science Planning Project (2022BJJ076), and the Henan Province Key Research and Development and Promotion Special Project (Soft Science Research) (222400410105).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. De Bruyne, L.; Karimi, A.; De Clercq, O.; Prati, A.; Hoste, V. Aspect-Based Emotion Analysis and Multimodal Coreference: A Case Study of Customer Comments on Adidas Instagram Posts. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 574–580.
2. Pastor, M.A.; Ribas, D.; Ortega, A.; Miguel, A.; Lleida, E. Cross-Corpus Training Strategy for Speech Emotion Recognition Using Self-Supervised Representations. *Appl. Sci.* **2023**, *13*, 9062. [CrossRef]
3. Fahad, M.S.; Ranjan, A.; Yadav, J.; Deepak, A. A survey of speech emotion recognition in natural environment. *Digit. Signal Process.* **2021**, *110*, 102951. [CrossRef]
4. Choi, Y.J.; Lee, Y.W.; Kim, B.G. Residual-based graph convolutional network for emotion recognition in conversation for smart Internet of Things. *Big Data* **2021**, *9*, 279–288. [CrossRef] [PubMed]
5. Koduru, A.; Valiveti, H.B.; Budati, A.K. Feature extraction algorithms to improve the speech emotion recognition rate. *Int. J. Speech Technol.* **2020**, *23*, 45–55. [CrossRef]

6. Jin, P.; Si, Z.; Wan, H.; Xiong, X. Emotion Classification Algorithm for Audiovisual Scenes Based on Low-Frequency Signals. *Appl. Sci.* **2023**, *13*, 7122. [CrossRef]
7. Mustaqeem; Kwon, S. A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors* **2019**, *20*, 183. [CrossRef] [PubMed]
8. Peng, Z.; He, W.; Li, Y.; Du, Y.; Dang, J. Multi-Level Attention-Based Categorical Emotion Recognition Using Modulation-Filtered Cochleagram. *Appl. Sci.* **2023**, *13*, 6749. [CrossRef]
9. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control.* **2020**, *59*, 101894. [CrossRef]
10. Wani, T.M.; Gunawan, T.S.; Qadri, S.A.A.; Kartiwi, M.; Ambikairajah, E. A comprehensive review of speech emotion recognition systems. *IEEE Access* **2021**, *9*, 47795–47814. [CrossRef]
11. Sajjad, M.; Kwon, S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access* **2020**, *8*, 79861–79875.
12. Abdullah, S.M.S.A.; Ameen, S.Y.A.; Sadeeq, M.A.; Zeebaree, S. Multimodal emotion recognition using deep learning. *J. Appl. Sci. Technol. Trends* **2021**, *2*, 52–58. [CrossRef]
13. Zehra, W.; Javed, A.R.; Jalil, Z.; Khan, H.U.; Gadekallu, T.R. Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex Intell. Syst.* **2021**, *7*, 1845–1854. [CrossRef]
14. Ancilin, J.; Milton, A. Improved speech emotion recognition with Mel frequency magnitude coefficient. *Appl. Acoust.* **2021**, *179*, 108046. [CrossRef]
15. Akçay, M.B.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *116*, 56–76. [CrossRef]
16. Semero, Y.K.; Zhang, J.; Zheng, D. PV power forecasting using an integrated GA-PSO-ANFIS approach and Gaussian process regression based feature selection strategy. *CSEE J. Power Energy Syst.* **2018**, *4*, 210–218. [CrossRef]
17. Elaziz, M.A.; Ahmadein, M.; Ataya, S.; Alsaleh, N.; Forestiero, A.; Elsheikh, A.H. A Quantum-Based Chameleon Swarm for Feature Selection. *Mathematics* **2022**, *10*, 3606. [CrossRef]
18. Aragón-Royón, F.; Jiménez-Vílchez, A.; Arauzo-Azofra, A.; Benítez, J.M. FSinR: An exhaustive package for feature selection. *arXiv* **2020**, arXiv:2002.10330.
19. Baioletti, M.; Milani, A.; Santucci, V. Variable neighborhood algebraic differential evolution: An application to the linear ordering problem with cumulative costs. *Inf. Sci.* **2020**, *507*, 37–52. [CrossRef]
20. Santos, S.P.; Gomez-Pulido, J.A.; Sanchez-Bajo, F. Deconvolution of X-ray Diffraction Profiles Using Genetic Algorithms and Differential Evolution. In Proceedings of the Advances in Computational Intelligence: 13th International Work-Conference on Artificial Neural Networks, IWANN 2015, Palma de Mallorca, Spain, 10–12 June 2015; pp. 503–514.
21. Yogesh, C.; Hariharan, M.; Ngadiran, R.; Adom, A.H.; Yaacob, S.; Polat, K. Hybrid BBO_PSO and higher order spectral features for emotion and stress recognition from natural speech. *Appl. Soft Comput.* **2017**, *56*, 217–232.
22. Shahin, I.; Alomari, O.A.; Nassif, A.B.; Afyouni, I.; Hashem, I.A.; Elnagar, A. An efficient feature selection method for arabic and english speech emotion recognition using Grey Wolf Optimizer. *Appl. Acoust.* **2023**, *205*, 109279. [CrossRef]
23. Yogesh, C.K.; Hariharan, M.; Ngadiran, R.; Adom, A.H.; Yaacob, S.; Berkai, C.; Polat, K. A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal. *Expert Syst. Appl.* **2017**, *69*, 149–158.
24. Butta, R.; Maddu, K.; Vangala, S. Cat swarm optimized ensemble technique for emotion recognition in speech signals. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e7319. [CrossRef]
25. Akinpelu, S.; Viriri, S. Robust Feature Selection-Based Speech Emotion Classification Using Deep Transfer Learning. *Appl. Sci.* **2022**, *12*, 8265. [CrossRef]
26. Kaur, B.; Rathi, S.; Agrawal, R. Enhanced depression detection from speech using Quantum Whale Optimization Algorithm for feature selection. *Comput. Biol. Med.* **2022**, *150*, 106122. [CrossRef]
27. Gideon, J.; Schatten, H.T.; McInnis, M.G.; Provost, E.M. Emotion recognition from natural phone conversations in individuals with and without recent suicidal ideation. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019.
28. Gharsellaoui, S.; Selouani, S.A.; Yakoub, M.S. Linear Discriminant Differential Evolution for Feature Selection in Emotional Speech Recognition. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019; pp. 3297–3301.
29. Chen, Y.L.; Wang, N.C.; Ciou, J.F.; Lin, R.Q. Combined Bidirectional Long Short-Term Memory with Mel-Frequency Cepstral Coefficients Using Autoencoder for Speaker Recognition. *Appl. Sci.* **2023**, *13*, 7008. [CrossRef]
30. Sun, L.; Qin, H.; Przystupa, K.; Cui, Y.; Kochan, O.; Skowron, M.; Su, J. A hybrid feature selection framework using improved sine cosine algorithm with metaheuristic techniques. *Energies* **2022**, *15*, 3485. [CrossRef]
31. Qin, A.K.; Suganthan, P.N. Self-adaptive differential evolution algorithm for numerical optimization. In Proceedings of the 2005 IEEE Congress on Evolutionary Computation, Edinburgh, UK, 2–5 September 2005; Volume 2, pp. 1785–1791.