applied sciences

MDPI

Article

English Speech Emotion Classification Based on Multi-Objective Differential Evolution

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
[2] School of Computer and Software, Nanyang Institute of Technology, Nanyang 473004, China
[3] College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
[4] Department of Information Management, Chaoyang University of Technology, Taichung 413310, Taiwan
* Correspondence: jengshyangpan@gmail.com

Abstract: Speech signals involve speakers' emotional states and language information, which is very important for human–computer interaction that recognizes speakers' emotions. Feature selection is a common method for improving recognition accuracy. In this paper, we propose a multi-objective optimization method based on differential evolution (MODE-NSF) that maximizes recognition accuracy and minimizes the number of selected features (NSF). First, the Mel-frequency cepstral coefficient (MFCC) features and pitch features are extracted from speech signals. Then, the proposed algorithm implements feature selection where the NSF guides the initialization, crossover, and mutation of the algorithm. We used four English speech emotion datasets, and K-nearest neighbor (KNN) and random forest (RF) classifiers to validate the performance of the proposed algorithm. The results illustrate that MODE-NSF is superior to other multi-objective algorithms in terms of the hypervolume (HV), inverted generational distance (IGD), Pareto optimal solutions, and running time. MODE-NSF achieved an accuracy of 49% using eNTERFACE05, 53% using the Ryerson audio-visual database of emotional speech and song (RAVDESS), 76% using Surrey audio-visual expressed emotion (SAVEE) database, and 98% using the Toronto emotional speech set (TESS). MODE-NSF obtained good recognition results, which provides a basis for the establishment of emotional models.

Keywords: speech signals; feature selection; multi-objective; differential evolution

Citation: Yue, L.; Hu, P.; Chu, S.-C.; Pan, J.-S. English Speech Emotion Classification Based on Multi-Objective Differential Evolution. *Appl. Sci.* **2023**, *13*, 12262. https://doi.org/10.3390/app132212262

Academic Editor: Javier Hernando

Received: 13 October 2023
Revised: 3 November 2023
Accepted: 9 November 2023
Published: 13 November 2023

1. Introduction

Speech signals are information rich and extend the content of written messages via the speakers' identity, their emotional state, and their intonation patterns [1]. They are easier to capture compared to other physiological signals [2].

Speech recognition technology grants machines the ability to express emotions and enables them to recognize human emotions. A lot of research has been conducted on speech emotion recognition (SER), and its applications are increasingly popular in the field of human–computer interaction, distance education, and emotional therapy. However, significant work is still required to make the applications more natural. In fact, the factors that affect a person's emotions are complex and diverse. Individuals experience various psychological changes in different emotional states. These changes lead them to associate emotional fluctuations with speech, and provide key emotional information for SER. Speech features are extracted to describe this information.

Emotion recognition relies on extracting meaningful features from speech signals. Currently, these features mainly include intonation features, spectrum features, voice quality features, and other acoustic features. Many features are used in speech recognition, and excellent results can be achieved through training with various machine learning methods. However, speech features introduce redundancy, and negatively impact recognition results.

Feature selection achieves dimensionality reduction by removing irrelevant and redundant features [3,4]. This is widely used in SER to reduce processing time and enhance recognition efficiency. Differential evolution (DE) mimics the natural concept of survival of the fittest, and gradually converges towards an optimal or near-optimal solution [5–7]. DE is known for its computational efficiency in optimizing feature subsets. It explores search space effectively, and it shows significant advantages in single- and multi-objective feature selection [8–11].

While multi-objective DE is fast, it has a few drawbacks. For example, it has trouble with unstable convergence and may lose population diversity. In this study, we investigated multi-objective DE to recognize speech emotion through feature selection, and the main contributions of this paper are summarized as follows:

1. We propose a model for speech emotion recognition;
2. We propose a feature extraction approach from speech signals;
3. We propose a multi-objective feature selection algorithm based on DE in which the number of selected features (NSF) guides the initialization, crossover, and mutation of DE;
4. We validated the performance of the proposed algorithm on K-nearest neighbor (KNN) and random forest (RF) classifiers with four English speech emotion datasets.

The structure of this paper is organized as follows. Section 2 introduces the related works of SER. Section 3 describes the proposed algorithm. Section 4 represents the experimental results with discussions, and Section 5 provides the conclusions.

2. Related Works

Existing research in speech emotion recognition is classified into single- and multi-objective optimization according to different goals.

Sun et al. proposed a SER method based on decision tree (DT), support vector machine (SVM), and Fisher feature selection [12]. The Fisher criterion is employed to filter out feature parameters with a high discrimination ability. The DT and SVM framework is first established by calculating the confusion of emotions, and then features with high discrimination are selected for each SVM in the DT according to the Fisher's criterion. Finally, SER is realized based on the model. Partila et al. discussed the impact of classification methods and feature selection on the accuracy of SER, and found the best combination of methods and feature sets for stress detection in human speech [13]. Selecting appropriate parameters for a classifier is an important part of reducing computational complexity, especially for systems intended for real-time applications. The classification accuracy of an artificial neural network, KNN, and Gaussian mixture model is measured considering the selection of foreground, spectral, and speech quality features. Traditional feature selection methods often rely on supervised learning, where emotion labels are used to guide the selection of relevant features. However, these methods may not be efficient when labeled data are scarce or expensive to obtain. To address this challenge, Bandela et al. proposed a novel approach that leverages unsupervised feature selection algorithms to identify the most informative and discriminative features from speech data without using emotion labels [14]. They explore various unsupervised feature selection algorithms, such as principal component analysis (PCA), independent component analysis (ICA), and clustering-based methods. Akinpelu and Viriri integrated robust feature selection and deep transfer learning to improve the performance and robustness of speech emotion classification [15]. The robust feature selection chooses features that are less affected by noise and irrelevant variations in data. The deep transfer learning employs knowledge learned from a pre-trained neural network model on a large dataset. Transfer learning allows the model to benefit from knowledge gained from one domain (e.g., a large general speech dataset) and apply it to a related but different domain (e.g., speech emotion classification). Li et al. addressed the problem of recognizing emotions from speech signals [16]. Speech features are extracted from signals that may carry emotional cues, and these features include acoustic features (pitch, intensity, and spectrum), prosodic features (speaking rate and pitch contour), and linguistic features (lexical content and sentiment-related words). The research involves

a comprehensive analysis of the extracted speech features to identify their relevance and importance for emotion recognition.

In addition to recognition accuracy, there are several studies on emotion recognition from computational efficiency, classifier optimization, and unity. Brester et al. proposed a novel approach that combines heuristic feature selection methods with a multi-objective optimization framework [17], aiming to maximize classification accuracy and minimize computational complexity. They optimize computational efficiency by working in parallel and incorporating a technique for exchanging subsets of data. Furthermore, the approach uses a beneficial pre-processing step when combined with an ensemble of classifiers. It not only streamlines the feature selection process but also enhances the overall classification performance. Daneshfar and Kabudian combined discriminative dimension reduction and a modified quantum-behaved particle swarm optimization (QPSO) algorithm to implement SER and optimize the Gaussian mixture model (GMM) classifier's parameters [18]. The dimension reduction method preserves emotion-specific features and improves the discriminative power of the extracted features. The modified QPSO algorithm enhances the optimization process for feature selection in SER. Li et al. presented a novel approach for enhancing emotion recognition through multiple data sources [19]. The proposed model utilizes a sophisticated multi-objective optimization algorithm to create the multi-modal system, which effectively combines voice and facial information with the goal of simultaneously improving recognition accuracy and consistency. Yildirim et al. introduced a modified feature selection method that employs metaheuristic algorithms to identify the most important features for SER [20]. Various metaheuristic algorithms, such as NSGA-II and cuckoo search, are applied to optimize the feature selection process. These algorithms efficiently explore feature space and converge to optimal feature subsets.

Drawing from previous research, multi-objective speech emotion recognition mainly includes classification accuracy and the number of selected features. Although the research achieved good results, it usually considers recognition accuracy when searching for the optimal solutions. It neglects the fact that the number of features is also the main factor affecting multi-objective algorithms, resulting in a loss of population diversity. Compared to other evolutionary algorithms, the popularity of DE as a competitive optimization algorithm is due to its efficiency, simplicity, robustness, and global search ability. We utilize it as a feature selection technique for multi-objective speech emotion recognition.

3. Materials and Method

The whole process consists of several main steps: firstly, preparing the data (pre-process); secondly, extracting important information (feature extraction); thirdly, selecting the most relevant features (feature selection); and finally, using classifiers for making predictions, as shown in Figure 1.

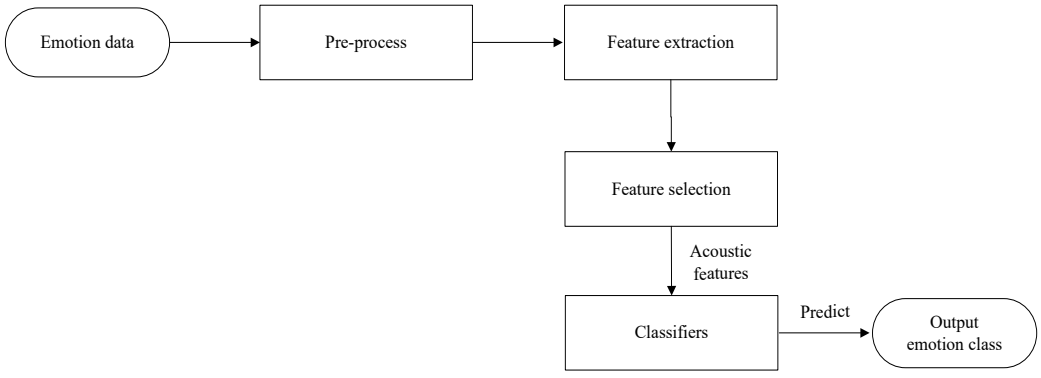


Figure 1. The flowchart of the proposed system.

3.1. Emotion Data

Four different English speech emotion datasets were used in this study. eNTERFACE05 is known for capturing spontaneous emotional expressions in unscripted scenarios, while the Ryerson audio-visual database of emotional speech and song (RAVDESS) focuses on emotional speech and song data. The Surrey audio-visual expressed emotion (SAVEE) dataset collects posed emotional expressions, and the Toronto emotional speech set (TESS) specializes in scripted emotional speech. Each database has unique characteristics that suit different research needs in the domain of emotion analysis and recognition. Through these datasets, the pros and cons of the algorithms can be more comprehensively evaluated.

1. eNTERFACE05: eNTERFACE05 is a well-known European research project and dataset that focuses on the development of technologies for human–computer interaction, particularly in the fields of facial and emotional expression recognition [21]. The eNTERFACE05 dataset contains emotional expressions such as happiness, sadness, anger, fear, disgust, and surprise;
2. Ryerson audio-visual database of emotional speech and song: RAVDESS is a valuable resource for studying emotions in speech and music because it includes both acted and natural emotional expressions performed by professional actors [22]. RAVDESS encompasses a wide range of emotions such as calm, happiness, sadness, anger, fear, surprise, and disgust;
3. Surrey audio-visual expressed emotion: SAVEE is a dataset containing audio and video recordings that display emotional expressions by English native speakers [23]. This dataset covers various emotions including happiness, anger, disgust, sadness, fear, and neutral;
4. Toronto emotional speech set: TESS consists of professionally acted and recorded speech segments spoken by North American English speakers [24]. TESS includes expressions of anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral.

3.2. Pre-Process

Pre-emphasis, framing, and windowing are important pre-process steps often applied to raw speech audio data. These steps ensure that speech data are cleaned, transformed, and organized in a way that allows machine learning algorithms to effectively learn and recognize emotional patterns.

Pre-emphasis is a filtering technique that is applied to raw speech signals before further analysis. It accentuates high frequencies in signals and improves the signal-to-noise ratio. Speech signals tend to have more energy in low frequencies, and pre-emphasis can balance this by boosting high-frequency components.

Framing involves dividing continuous speech signals into shorter overlapping segments (frames). The reason for this is that speech characteristics, such as pitch and spectral content, can change rapidly in short time intervals. By analyzing these frames individually, we capture variations more accurately. Each frame typically contains about 20–30 ms of speech data.

After framing, a windowing function is applied to each frame. Windowing reduces sudden changes at the edges of frames and prevents artifacts during the subsequent analysis, such as the Fourier transform. Common windowing functions include the Hamming, Hanning, and Blackman windows. These functions smoothly taper signals within frames, and decrease the effects of spectral leakage.

3.3. Feature Extraction

In this study, we extracted Mel-frequency cepstral coefficient (MFCC) features and pitch features from raw audios. A total of 141 values were extracted, and Table 1 describes their details.

1. Pitch features: Pitch features are important elements extracted from speech signals that provide information about the fundamental frequency (F0) and tonal characteristics of the human voice. They are extracted using autocorrelation, cepstral analysis, and

wavelet transform. By analyzing these features, SER systems can better understand and interpret the emotional nuances and linguistic characteristics of spoken language;

2. MFCC features: MFCC features capture the essential characteristics of speech signals, and ignore redundant or less important information. They mimic how human auditory systems process sounds by converting their frequency spectrum into a representation that's easier for computers to understand.

Table 1. The details of the features.

Features	Details
Pitch	spurt length the max, min, median, mean and variance of each pitch the max derivative, min derivative, median derivative, mean derivative and variance derivative of each pitch
MFCC	the max, min, median, mean and variance of each coefficient the max derivative, min derivative, median derivative, mean derivative and variance derivative of each coefficient

3.4. Improved Multi-Objective Differential Evolution for Feature Selection

The proposed MODE-NSF, shown in Figure 2, includes four new schemes. Firstly, the initialization is defined by the NSF. Then, this NSF is adopted to adjust crossover. Third, the NSF-based mutation strategy is introduced to balance exploration and exploitation. Finally, MODE-NSF combines new and old solutions, and deletes duplicate individuals. In our speech emotion recognition model, MODE-NSF implements the feature selection operation shown in Figure 1.

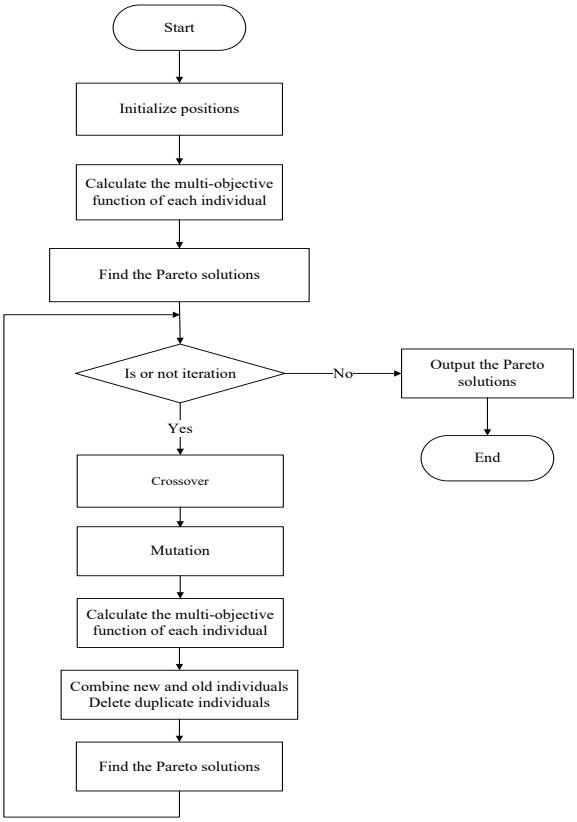


Figure 2. The flowchart of the proposed MODE-NSF.

3.4.1. Multi-Objective Feature Selection

Multi-objective algorithms are designed to find solutions for multiple objectives in decision space. Because these objectives are in conflict, a significant challenge is comparing two potential solutions. If solution S1 outperforms solution S2 in all objectives, we say that S1 dominates S2. Non-dominated solutions, like S1, are not outperformed by any other solution, and they are called Pareto solutions. The set of all non-dominated solutions forms a Pareto set, and their corresponding objective values make up the Pareto front.

A binary string represents a solution of multi-objective feature selection, where 0/1 represents the unselected/selected feature. Feature selection mainly involves two objectives: maximizing classification accuracy and minimizing the number of selected features [25], as shown in Equation (1).

$$\begin{aligned} min\ F(x) &= [f_1(x), f_2(x)] \\ subject\ to\ f_1 &= number(x) \\ f_2 &= 1 - accuracy(x) \end{aligned} \tag{1}$$

where f_1 is the number of feature sets (x), and f_2 denotes the classification accuracy of x.

In multi-objective feature selection, the number of selected features (f_1) defines an optimization objective, so we utilize it to implement a search.

3.4.2. Initialization

Random initialization is not employed to expand feature search space. MODE-NSF randomly generates the NSF and determines the initial positions using it. MODE-NSF defines a counter to store the usage frequency of the NSF, and Algorithm 1 provides the details of the novel initialization.

Algorithm 1: The initialization based on the NSF

```
% Input: dim represents the number of features
%        nPop means the population size
%        flags is used to count the usage frequency of the NSF
% Output: the positions (pop) of individuals
for i = 1:nPop do
    dd = randperm(dim,1) ;
    id = randperm(dim, dd);
    pop(i).Position(id) = 1;
    Execte Equation (1) ;
    flags(pop(i).Cost(1)) = flags(pop(i).Cost(1)) + 1;
end
```

3.4.3. Crossover

Crossover operators are crucial because they enhance exploration ability. This process, similar to reproduction in nature, ensures the survival of species. In DE, crossover generates new offspring, and it explores search space more effectively. During crossover, two or more individuals from the population act as parents to create children. These children inherit genetic information from their parents, which improves the algorithm's chances of finding the global optimal solution.

MODE-NSF utilizes Equations (2) and (3) to implement crossover. In addition, r_2 and r_3 are two random individuals, and r_1 comes from Pareto solutions with the minimum NSF difference of i. Moreover, λ and pCR are called the scale factor and crossover probability, respectively. The NSF difference between r_1 and i is the smallest. It prompts i to quickly approach r_1 and increases the convergence of the algorithm. However, if r_1 is a local

optimum, it may cause the algorithm to fall into a local trap, which can be solved by the mutation operation.

$$m_i(t+1) = Position_{r_1}(t) + \lambda * (Position_{r_2}(t) - Position_{r_3}(t)) \tag{2}$$

$$Position_i^j(t+1) = \begin{cases} m_i^j(t+1) & if(rand(j) \leq pCR) \quad or \quad j = randi(i) \\ Position_i^j(t) & if(rand(j) > pCR) \quad or \quad j! = randi(i) \end{cases} \tag{3}$$

3.4.4. Mutation

Mutation is the process of introducing a small random change into a solution. This change promotes diversity and exploration within solutions. In essence, mutation prevents the algorithm from getting stuck in local optima and can lead to the discovery of better solutions. Algorithm 2 presents the procedure of the proposed mutation, and *mt* means a random integer value between the maximum NSF of Pareto solutions and the minimum NSF of Pareto solutions. Lines 2–6 of the algorithm represent reducing the NSF, and lines 7–12 represent increasing the NSF.

Algorithm 2: The mutation based on NSF

```
1  % Input: mutation i
2  % Output: i
3  m_p = find(pop(i).Position==1) ;
4  if length(m_p) > mt then
5      m_a = randperm(length(m_p), length(m_p) - mt);
6      m_b = m_p(m_a);
7      pop(i).Position(m_b) = 0;
8  end
9  if length(m_p) < mt then
10     m_p0 = find(pop(i).Position==0);
11     m_a = randperm(length(m_p0), mt - length(m_p));
12     m_b = m_p0(m_a);
13     pop(i).Position(m_b) = 1;
14 end
```

3.5. Classifiers

1. K-nearest neighbor: KNN is a simple classification algorithm that is easy to comprehend and implement [26]. It is non-parametric, and it does not make strong assumptions about the distribution of data. This can be useful when dealing with diverse and complex emotional speech data;
2. Random forest: Random forest is an ensemble method that consists of multiple decision trees [27]. The ensemble approach tends to produce results that are robust and accurate, which reduces the risk of overfitting and improves generalization. RF is a suitable option for speech data that contain various types of noise in real-world scenarios.

   These classifiers offer a balance between simplicity and effectiveness, which is important in the context of speech emotion recognition. Both KNN and RF have demonstrated their utility in emotion recognition, making them reasonable choices for this study;
3. K-fold cross validation: K-fold cross validation prevents overfitting and provides a more precise depiction of a model's true performance [28]. By dividing a dataset into multiple subsets, it continuously trains and evaluates a model on different combinations of these subsets.

In this study, we used 5-Nearest Neighbor and RF (with 20 decision trees) classifiers to create models, and then assessed the performance of these models using 10-fold cross-validation.

4. Experimental Results and Analysis

4.1. Approaches Used for Comparisons

The proposed MODE-NSF's superiority was verified by comparing its classification performance with MOGA [17], MODE [29], and NSGA-II [20]. Table 2 provides more details concerning the algorithms.

The algorithms had a maximum number of iterations of 100 with 20 runs, and the population size was 20. Wilcoxon rank sum and Frideman test were employed to determine whether there were any significant differences in the experimental results. The significant level was chosen to be 0.05, which means that if p-value <= 0.05, an algorithm was significantly superior to the compared algorithms at a 95% confidence.

Table 2. The parameter settings of the algorithms.

Algorithm	Main Parameters
MOGA	pC = 1; mu = 0.02;
MODE	pCR = 0.2;
NSGA-II	tournament; mu = 20; mum = 20;
MODE-NSF	pCR = 0.2; beta_min = 0.2; beta_max = 0.8;

4.2. Experimental Analysis

4.2.1. Simulation Results on the KNN Classifier

1. Hypervolume (HV)

Table 3 provides the HV of the algorithms where *AVG* and *STD* denote the average and variance of the HV, respectively. MODE-NSF outperformed MOGA, MODE, and NSGA-II using eNTERFACE05, RAVDESS, and SAVEE, while NSGA-II obtained the best HV value using TESS. The algorithms achieved low HV values using eNTERFACE05 and RAVDESS, but high values using TESS. The multi-objective algorithms performed poorly in the Pareto optimal solutions for the eNTERFACE05 and RAVDESS datasets, while the optimal solutions for TESS were close to the ideal value. The Wilcoxon rank sum revealed that MOGA, MODE, NSGA-II, and MODE-NSF performed well on 0, 0, 3, and 4 datasets. MODE-NSF and NSGA-II produced similar experimental data for RAVDESS, SAVEE, and TESS. According to the Friedman test, their average ranks were 3, 4, 1.75, and 1.25, respectively, proving that MODE-NSF performed the best, followed by NSGA-II, MOGA, and MODE. The NSF improves the multi-objective solution ability of DE.

Table 3. The HV values of the algorithms.

Dataset	MOGA		MODE		NSGA-II		MODE-NSF	
	AVG	STD	AVG	STD	AVG	STD	AVG	STD
eNTERFACE05	0.0242	0.0364	0.0129	0.0278	0.0798	0.0578	**0.1476**	0.0842
RAVDESS	0.0408	0.0471	0.0195	0.0221	0.1343	0.0320	**0.1356**	0.0542
SAVEE	0.0915	0.0619	0.0538	0.0462	0.2140	0.0390	**0.2334**	0.0428
TESS	0.1640	0.1118	0.1078	0.0925	**0.4545**	0.1244	0.4254	0.1251
>/=/<	0/0/4		0/0/4		1/2/1		3/1/0	
Rank	3		4		1.75		1.25	
p-Value	1.12×10^{-2}							

2. Inverted generational distance (IGD)

Table 4 shows the IGD of the algorithms, along with their Wilcoxon rank-sum and Friedman test results. MODE-NSF exhibited the best performance using eNTERFACE05,

RAVDESS, SAVEE, and TESS. MODE-NSF's values were lower than those of the other algorithms, but its solutions were almost the same as the Pareto solutions obtained by them. The Wilicion rank sum indicates that MOGA, MODE, and NSGA-II did not share similar statistical data with MODE-NSF on the four datasets. The average ranks obtained by the Frideman test were 3, 4, 2, and 1, and the p-value was 1.12×10^{-2}. Experimental data and non-parameter validation show that MODE-NSF outperformed MOGA, MODE, and NSGA-II in terms of IGD.

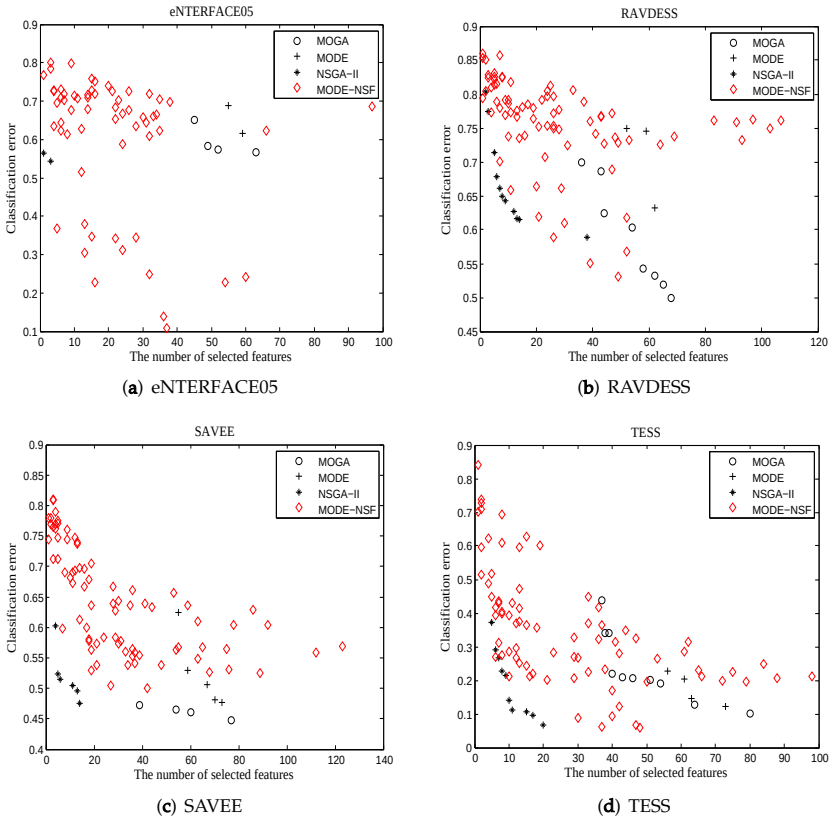
Table 4. The IGD values of the algorithms.

Dataset	MOGA		MODE		NSGA-II		MODE-NSF	
	AVG	STD	AVG	STD	AVG	STD	AVG	STD
eNTERFACE05	0.2820	0.1393	0.3245	0.0923	0.1441	0.1308	**0.0564**	0.0616
RAVDESS	0.2437	0.0677	0.2940	0.0446	0.0995	0.0554	**0.0623**	0.0336
SAVEE	0.2167	0.0642	0.2546	0.0433	0.1220	0.0543	**0.0783**	0.0507
TESS	0.2314	0.0754	0.2933	0.0579	0.1541	0.0733	**0.1070**	0.0635
>/=/<	0/0/4		0/0/4		0/0/4		4/-/-	
Rank	3		4		2		1	
p-Value	2.56×10^{-2}							

The hypervolume, inverted generational distance, and non-parametric statistical analysis verified that MODE-NSF has a remarkable distribution and convergence, and it balances exploration and exploitation.

3. Pareto solutions

Figure 3 shows the final Pareto optimal solutions acquired by the algorithms in which Pareto solutions were the solutions that were non-dominated and obtained from all algorithms after 20 runs.



(**a**) eNTERFACE05 (**b**) RAVDESS

(**c**) SAVEE (**d**) TESS

Figure 3. Pareto optimal solutions acquired by the algorithms.

For eNTERFACE05, MODE-NSF obtained more solutions than MOGA, MODE, and NSGA-II. Furthermore, it achieved the highest recognition accuracy of 49%, which was superior to the other algorithms. The solutions of NSGA-II were located in a low dimension, and the solutions of MOGA and MODE were mainly in a medium dimension. For RAVDESS, the solutions of MODE-NSF were distributed over the entire feature space. The solutions of NSGA-II remained in a low dimension, and the solutions of MOGA and MODE were located in a middle dimension. Although MODE did not achieve as many optimal solutions as MODE-NSF, it outperformed MODE-NSF in obtaining an optimal recognition accuracy. For SAVEE, MODE-NSF presented with diverse characteristics. The solutions of NSGA-II were in a low dimension, and the solutions of MOGA and MODE were in a medium dimension. The recognition accuracy of the Pareto optimal solutions obtained by NSGA-II, MOGA, and MODE was better than that obtained by MODE-NSF. For TESS, MODE-NSF excelled in terms of diversity and accuracy. NSGA-II achieved an accuracy of 90% using a small number of features, and MOGA acquired a better accuracy in the middle dimension than MODE. From the Pareto optimal solutions, it was found that MODE-NS utilized the NSF-guided mutation to enhance the solution's diversity and balances exploration and exploitation.

Table 5 presents the running time of the algorithms. MODE had the shortest computation time using eNTERFACE05, and MODE-NSF exhibited the quickest execution time using RAVDESS, SAVEE, and TESS. The proposed algorithm exhibited a fast execution and low time complexity. It is worth noting that the algorithms required less time to execute using eNTERFACE05 and SAVEE compared to TESS. This is because TESS contains a larger number of samples, which impacted the execution of the algorithms.

Table 5. The average running time of the algorithms (second).

Dataset	MOGA	MODE	NSGA-II	MODE-NSF
eNTERFACE05	2148.5126	**187.7856**	1860.2133	268.8662
RAVDESS	5858.3368	407.2977	2797.5543	**295.3447**
SAVEE	2226.02	180.2649	1802.0954	**178.759**
TESS	16,305.3481	954.1857	4681.1689	**543.2745**

4.2.2. Simulation Results on the RF Classifier

1. Hypervolume

Table 6 shows the HV values of the multi-objective algorithms. The data obtained by the algorithms using eNTERFACE05, RAVDESS, SAVEE, and TESS were better than those obtained using the KNN classifier. MODE-NSF outperformed the other algorithms on four datasets, especially TESS, where it achieved a value of 0.8101, a result close to the theoretical optimal value. The Friedman test showed that their average ranks were 2.75, 3.75, 2.5, and 1. The Wilcoxon rank sum revealed that they performed well on 2, 1, 3, and 4 datasets. MODE-NSF and MOGA produced consistent statistical data using eNTERFACE05 and RAVDESS. Additionally, MODE-NSF and MODE exhibited a similar performance using RAVDESS, while MODE-NSF and NSGA-II achieved similar experimental results using eNTERFACE05, RAVDESS, and SAVEE.

Table 6. The HV values of the algorithms.

Dataset	MOGA		MODE		NSGA-II		MODE-NSF	
	AVG	STD	AVG	STD	AVG	STD	AVG	STD
eNTERFACE05	0.1123	0.0232	0.0629	0.0292	0.1364	0.0514	**0.1661**	0.0446
RAVDESS	0.2794	0.0523	0.2563	0.0450	0.3169	0.0891	**0.3429**	0.0850
SAVEE	0.2049	0.0764	0.1686	0.0896	0.2935	0.0834	**0.348**	0.0619
TESS	0.6167	0.0278	0.5895	0.0374	0.5152	0.1988	**0.8101**	0.0507
>/=/<	0/2/2		0/1/3		0/3/1		4/-/-	
Rank	2.75		3.75		2.5		1	
p-Value	0.0256							

2. Inverted generational distance

Table 7 provides the IGD of the algorithms and their corresponding Wilcoxon rank-sum and Friedman test results. The IGD value of MODE-NSF was significantly smaller than that of the other algorithms, which means that the multi-objective solutions obtained by it were close to the Pareto front composed of all algorithms. The performance of MODE-NSF was better than the others. The Wilcoxon rank sum indicates that MOGA produced similar statistics to MODE-NSF using eNTERFACE05; in addition, MOGA, MODE, NSGA-II, and MODE-NSF performed well on 1, 0, 0, and 4 datasets. Their average ranks were 2.25, 3.5, 3.25, and 1, and the *p*-value was less than 0.05.

Table 7. The IGD values of the algorithms.

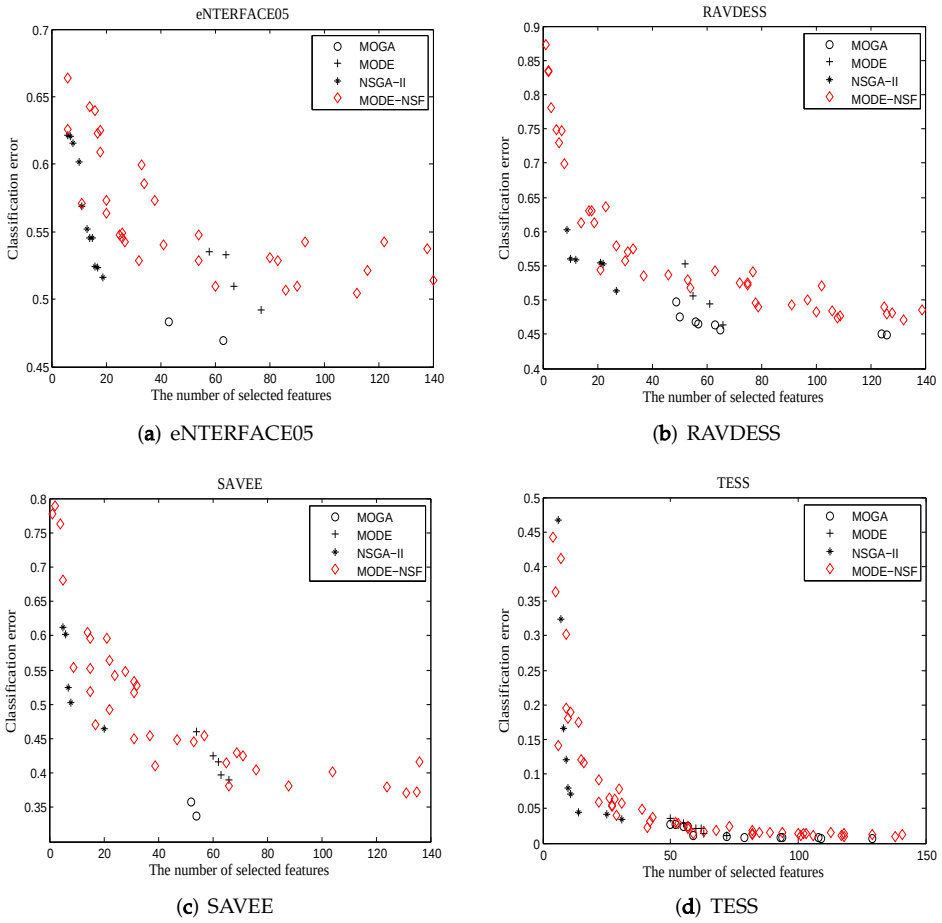
Dataset	MOGA		MODE		NSGA-II		MODE-NSF	
	AVG	STD	AVG	STD	AVG	STD	AVG	STD
eNTERFACE05	0.1489	0.0518	0.2545	0.0605	0.1816	0.0704	**0.0881**	0.0148
RAVDESS	0.1491	0.0172	0.2496	0.0400	0.2786	0.0246	**0.0631**	0.0118
SAVEE	0.2323	0.0257	0.2613	0.0326	0.1992	0.0415	**0.0485**	0.0114
TESS	0.1428	0.0268	0.2216	0.0208	0.3046	0.0195	**0.0407**	0.0140
>/=/<	0/1/3		0/0/4		0/0/4		4/-/-	
Rank	2.25		3.5		3.25		1	
p-Value	0.0256							

Tables 6 and 7 confirm that the proposed MODE-NSF demonstrates exceptional distribution and convergence, and it is suitable for multi-objective feature selection.

3. Pareto solutions

Figure 4 illustrates the Pareto optimal solutions acquired by the algorithms.

Using eNTERFACE05, MODE-NSF obtained a large number of solutions, while the solutions of NSGA-II were concentrated in a low dimension. MOGA acquired four feasible solutions in a middle dimension. MODE only obtained two optimal solutions, but their classification accuracy was superior to the other algorithms. Using RAVDESS, the solutions of MODE-NSF were distributed throughout the entire feature space, while the solutions of NSGA-II were primarily in a low dimension. The solutions of MOGA and MODE were located in a middle dimension. Using SAVEE, MODE-NSF had a better diversity than the other algorithms. The solutions of NSGA-II were distributed in the region [0, 20], while the solutions of MOGA were in the region [50, 70]. Although MODE obtained fewer solutions compared to the other algorithms, it had the highest classification accuracy. Using TESS, MODE-NSF exhibited excellent abilities in diversity and classification accuracy, and MODE and MOGA also used more features to achieve a low amount of recognition errors. NSGA-II outperformed the other algorithms in low dimensional solutions. From the Pareto optimal solutions of MODE-NSF on the four datasets, it can be seen that the NSF-guided mutation is able to search for solutions in more emotional space and improve the population's diversity.



(**a**) eNTERFACE05 (**b**) RAVDESS

(**c**) SAVEE (**d**) TESS

Figure 4. Pareto optimal solutions acquired by the algorithms.

Table 8 presents the running time of the algorithms. The maximum time complexity of RF was larger than that of KNN, resulting in a longer run time for the algorithms compared to KNN. MODE-NSF had a superior operational efficiency. The algorithms ran longer for TESS than they did for eNTERFACE05 and SAVEE.

Table 8. The average running time of the algorithms (second).

Dataset	MOGA	MODE	NSGA-II	MODE-NSF
eNTERFACE05	18,608.8552	4885.9528	11,366.0398	**4008.0458**
RAVDESS	61,563.0978	16,457.0307	33,405.7332	**11,584.3484**
SAVEE	18,906.203	5643.6755	12,084.5362	**3997.2735**
TESS	72,422.1074	19,536.5901	41,187.0129	**16,545.7497**

4.3. Discussion

The running time of MODE-NSF using the KNN classifier in the four datasets was 268.8662, 295.3447, 178.7590, and 543.2745 respectively, while the running time on the RF classifier was 4008.0458, 11,584.3484, 3997.2735, and 16,545.7497. It is also reported in [30,31] that RF has a large time complexity, while KNN has a small workload. Ref. [31] achieved a recognition accuracy of 41% for eNTERFACE05, while MONDE-NSF acquired an accuracy of 49%. In [32,33], they obtained an accuracy of 75% using MFCCs on RAVDESS, while MODE-NSF only exhibited an accuracy of 53%. Both the results reported in [34] and those of MODE-NSF exhibited an accuracy of 76% for SAVEE. Ref. [15] achieved a 97% accurate classification for TESS, compared to 98% for MODE-NSF. The algorithms acquired a high recognition accuracy using TESS and a low value for eNTERFACE05. MODE-NSF showed

excellent performance for eNTERFACE05, SAVEE, and TESS, which illustrates that the NSF-guided method is suitable for speech emotion recognition.

5. Conclusions

Humans intentionally or unintentionally engage in emotional recognition when they interact with others. Speech signals can be extracted and used to classify emotions, and significant progress has been made in the field of emotion recognition. However, there is still a need for research in multi-objective emotion recognition. For this reason, pre-processing and feature selection are important for SER. In this paper, we propose a speech emotion recognition model based on DE as a feature selection method, using KNN and RF for emotion classification. First, feature extraction is applied to speech data, and then MFCCs and pitch features undergo DE to acquire the most relevant emotion features and discard redundant features. An accurate and robust SER is achieved through the reconstruction of input data with meaningful acoustic features. The NSF-guided multi-objective DE algorithm is responsible for efficiently exploring the emotional feature space and identifying the features for emotion classification. In English speech emotion datasets, the proposed MODE-NSF achieved a higher recognition accuracy with fewer features compared to the other multi-objective algorithms. MODE-NSF demonstrated a great execution efficiency because the number of features is the main factor affecting the running time of feature selection algorithms.

In the future, the proposed MODE-NSF algorithm can be applied in other popular research applications, especially in customer service, voice assistants, and English education. Furthermore, this algorithm can employ more acoustic features such as LPC, LSF, and DWT.

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, L.Y.; writing—review & 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: The data presented in this study are available in the article.

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

References

1. Hasija, T.; Kadyan, V.; Guleria, K.; Alharbi, A.; Alyami, H.; Goyal, N. Prosodic feature-based discriminatively trained low resource speech recognition system. *Sustainability* **2022**, *14*, 614. [CrossRef]
2. Arslan, R.S.; Barışçı, N. Development of output correction methodology for long short term memory-based speech recognition. *Sustainability* **2019**, *11*, 4250. [CrossRef]
3. Zhao, Z.D.; Zhao, M.S.; Lu, H.L.; Wang, S.H.; Lu, Y.Y. Digital Mapping of Soil pH Based on Machine Learning Combined with Feature Selection Methods in East China. *Sustainability* **2023**, *15*, 12874. [CrossRef]
4. Song, D.; Huang, D.; Li, L.; He, H.L. Biomedical Named Entity Recognition Based on Feature Selection and Word Representations. *J. Inf. Hiding Multim. Signal Process.* **2016**, *7*, 729–740.
5. Yuan, S.; Ji, Y.; Chen, Y.; Liu, X.; Zhang, W. An Improved Differential Evolution for Parameter Identification of Photovoltaic Models. *Sustainability* **2023**, *15*, 13916. [CrossRef]
6. Feleke, S.; Pydi, B.; Satish, R.; Kotb, H.; Alenezi, M.; Shouran, M. Frequency stability enhancement using differential-evolution- and genetic-algorithm-optimized intelligent controllers in multiple virtual synchronous machine systems. *Sustainability* **2023**, *15*, 13892. [CrossRef]
7. Pan, Z.B.; Yang, L.; Xu, Z.X.; Wang, D.Y. A NEC-based parallel differential evolution algorithm with MKL/CUDA. *J. Netw. Intell.* **2022**, *7*, 114–128.
8. Li, T.; Dong, H.; Sun, J. Binary differential evolution based on individual entropy for feature subset optimization. *IEEE Access* **2019**, *7*, 24109–24121. [CrossRef]

9. Zhang, Y.; Gong, D.W.; Gao, X.Z.; Tian, T.; Sun, X.Y. Binary differential evolution with self-learning for multi-objective feature selection. *Inf. Sci.* **2020**, *507*, 67–85. [CrossRef]
10. Hancer, E. Fuzzy kernel feature selection with multi-objective differential evolution algorithm. *Connect. Sci.* **2019**, *31*, 323–341. [CrossRef]
11. Wang, P.; Xue, B.; Liang, J.; Zhang, M. Feature selection using diversity-based multi-objective binary differential evolution. *Inf. Sci.* **2023**, *626*, 586–606. [CrossRef]
12. Sun, L.; Fu, S.; Wang, F. Decision tree SVM model with Fisher feature selection for speech emotion recognition. *EURASIP J. Audio Speech Music Process.* **2019**, *2019*, 1–14. [CrossRef]
13. Partila, P.; Voznak, M.; Tovarek, J. Pattern recognition methods and features selection for speech emotion recognition system. *Sci. World J.* **2015**, *2015*, 573068. [CrossRef]
14. Bandela, S.R.; Kumar, T.K. Speech emotion recognition using unsupervised feature selection algorithms. *Radioengineering* **2020**, *29*, 353–364. [CrossRef]
15. Akinpelu, S.; Viriri, S. Robust Feature Selection-Based Speech Emotion Classification Using Deep Transfer Learning. *Appl. Sci.* **2022**, *12*, 8265. [CrossRef]
16. Li, D.; Zhou, Y.; Wang, Z.; Gao, D. Exploiting the potentialities of features for speech emotion recognition. *Inf. Sci.* **2021**, *548*, 328–343. [CrossRef]
17. Brester, C.; Semenkin, E.; Sidorov, M. Multi-objective heuristic feature selection for speech-based multilingual emotion recognition. *J. Artif. Intell. Soft Comput. Res.* **2016**, *6*, 243–253. [CrossRef]
18. Daneshfar, F.; Kabudian, S.J. Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. *Multimed. Tools Appl.* **2020**, *79*, 1261–1289. [CrossRef]
19. Li, M.; Qiu, X.; Peng, S.; Tang, L.; Li, Q.; Yang, W.; Ma, Y. Multimodal emotion recognition model based on a deep neural network with multiobjective optimization. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 6971100. [CrossRef]
20. Yildirim, S.; Kaya, Y.; Kılıç, F. A modified feature selection method based on metaheuristic algorithms for speech emotion recognition. *Appl. Acoust.* **2021**, *173*, 107721. [CrossRef]
21. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The eNTERFACE'05 audio-visual emotion database. In Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 3–7 April 2006; p. 8.
22. Livingstone, S.R.; Russo, F.A. 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* **2018**, *13*, e0196391. [CrossRef] [PubMed]
23. Vryzas, N.; Kotsakis, R.; Liatsou, A.; Dimoulas, C.A.; Kalliris, G. Speech emotion recognition for performance interaction. *J. Audio Eng. Soc.* **2018**, *66*, 457–467. [CrossRef]
24. Dupuis, K.; Pichora-Fuller, M.K. Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set. *Can. Acoust.* **2011**, *39*, 182–183.
25. Xue, Y.; Tang, Y.; Xu, X.; Liang, J.; Neri, F. Multi-objective feature selection with missing data in classification. *IEEE Trans. Emerg. Top. Comput. Intell.* **2021**, *6*, 355–364. [CrossRef]
26. Bansal, M.; Goyal, A.; Choudhary, A. A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decis. Anal. J.* **2022**, *3*, 100071. [CrossRef]
27. Zhou, J.; Huang, S.; Qiu, Y. Optimization of random forest through the use of MVO, GWO and MFO in evaluating the stability of underground entry-type excavations. *Tunn. Undergr. Space Technol.* **2022**, *124*, 104494. [CrossRef]
28. Rabinowicz, A.; Rosset, S. Cross-validation for correlated data. *J. Am. Stat. Assoc.* **2022**, *117*, 718–731. [CrossRef]
29. Ali, I.M.; Essam, D.; Kasmarik, K. Novel binary differential evolution algorithm for knapsack problems. *Inf. Sci.* **2021**, *542*, 177–194. [CrossRef]
30. Das, A.; Guha, S.; Singh, P.K.; Ahmadian, A.; Senu, N.; Sarkar, R. A hybrid meta-heuristic feature selection method for identification of Indian spoken languages from audio signals. *IEEE Access* **2020**, *8*, 181432–181449. [CrossRef]
31. Özseven, T. A novel feature selection method for speech emotion recognition. *Appl. Acoust.* **2019**, *146*, 320–326. [CrossRef]
32. Shahin, I.; Hindawi, N.; Nassif, A.B.; Alhudhaif, A.; Polat, K. Novel dual-channel long short-term memory compressed capsule networks for emotion recognition. *Expert Syst. Appl.* **2022**, *188*, 116080. [CrossRef]
33. Bhavan, A.; Chauhan, P.; Shah, R.R. Bagged support vector machines for emotion recognition from speech. *Knowl.-Based Syst.* **2019**, *184*, 104886. [CrossRef]
34. Liu, Z.T.; Xie, Q.; Wu, M.; Cao, W.H.; Mei, Y.; Mao, J.W. Speech emotion recognition based on an improved brain emotion learning model. *Neurocomputing* **2018**, *309*, 145–156. [CrossRef]