


## Article

# Multi-Layer Hybrid Fuzzy Classification Based on SVM and Improved PSO for Speech Emotion Recognition

Shihan Huang <sup>1</sup>, Hua Dang <sup>1</sup>, Rongkun Jiang <sup>1,2,3</sup> , Yue Hao <sup>1</sup>, Chengbo Xue <sup>1,2</sup> and Wei Gu <sup>4,\*</sup>

<sup>1</sup> School of Integrated Circuits and Electronics, Beijing Institute of Technology (BIT), Beijing 100081, China; huangshihan@bit.edu.cn (S.H.); hua\_dang@163.com (H.D.); jiangrongkun@bit.edu.cn (R.J.); hy@bit.edu.cn (Y.H.); xue\_chengbo@163.com (C.X.)

<sup>2</sup> BIT Chongqing Innovation Center, Chongqing 401120, China

<sup>3</sup> BIT Chongqing Center for Microelectronics and Microsystems, Chongqing 401332, China

<sup>4</sup> School of Information and Electronics, Beijing Institute of Technology (BIT), Beijing 100081, China

\* Correspondence: bitguwei@163.com

**Abstract:** Speech Emotion Recognition (SER) plays a significant role in the field of Human–Computer Interaction (HCI) with a wide range of applications. However, there are still some issues in practical application. One of the issues is the difference between emotional expression amongst various individuals, and another is that some indistinguishable emotions may reduce the stability of the SER system. In this paper, we propose a multi-layer hybrid fuzzy support vector machine (MLHF-SVM) model, which includes three layers: feature extraction layer, pre-classification layer, and classification layer. The MLHF-SVM model solves the above-mentioned issues by fuzzy c-means (FCM) based on identification information of human and multi-layer SVM classifiers, respectively. In addition, to overcome the weakness that FCM tends to fall into local minima, an improved natural exponential inertia weight particle swarm optimization (IEPSO) algorithm is proposed and integrated with fuzzy c-means for optimization. Moreover, in the feature extraction layer, non-personalized features and personalized features are combined to improve accuracy. In order to verify the effectiveness of the proposed model, all emotions in three popular datasets are used for simulation. The results show that this model can effectively improve the success rate of classification and the maximum value of a single emotion recognition rate is 97.67% on the EmoDB dataset.

**Keywords:** speech emotion recognition; fuzzy c-means; particle swarm optimization; support vector machines



check for updates

**Citation:** Huang, S.; Dang, H.; Jiang, R.; Hao, Y.; Xue, C.; Gu, W. Multi-Layer Hybrid Fuzzy Classification Based on SVM and Improved PSO for Speech Emotion Recognition. *Electronics* **2021**, *10*, 2891. <https://doi.org/10.3390/electronics10232891>

Academic Editor: Ahmad Taher Azar

Received: 12 October 2021

Accepted: 20 November 2021

Published: 23 November 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

With the gradual enrichment of material life, people's attention has gradually shifted from the physical world to the spiritual world [1]. Research on human–computer interaction (HCI) for emotion recognition is increasingly becoming a hot topic. In the field of emotion recognition, it is necessary to develop machines that can understand human emotions better. In the field of HCI, the computer can recognize emotions through gesture, audio signals, body poses, facial expression, physiological signals, and neuroimaging methods, etc. [2,3]. Apart from expressing emotions and communicating, speech is the most natural and fastest method for speakers to convey emotions by intonation, volume, and speed compared to others. Speech emotion recognition (SER) is a method of emotion identification by extracting the emotional state of speakers from their speech signals [4–6]. In recent decades, SER, as the main form of emotional display, has focused on achieving a more natural interaction between people and machines and has become deeply involved in a wide bank of real-life applications, such as public safety [7], diagnosis of psychiatric diseases [8], adjustment of driving behavior from the state of drivers [9], web games, emergency call centers [10], and so on.

The wide application of SER depends on the rapid development of its technology. With the extensive expansion of deep learning research [11–16], SER technology has gradually

been applied beyond the use of basic methods such as hidden Markov model (HMM), the Gaussian mixture model (GMM), support vector machine (SVM), and k-nearest neighbor (KNN) [17–20]. Deep learning and its various variants also optimize the system from the aspects of feature extraction and classifier performance [21–25], which are the forms of SER model become increasingly diverse. In addition to analyzing only speech signals, the emotion recognition system based on deep learning also uses the multimodal recognition method to analyze emotions combined with attributes such as text, facial expression, gesture, and electrocardiography [26–28]. Meanwhile, researchers have also studied the SER system on contextual dependencies of speech signals by optimizing recurrent neural networks (RNN) and long short-term memory (LSTM) networks [29,30]. However, speech signal is still one of the most relevant attributes of emotion recognition, and SVM is still widely used in classification models because it provides quite good performance for SER system with less complexity. Therefore, the importance of the most common method cannot be ignored.

The aforementioned studies improve recognition accuracy by extracting more appropriate speech emotion features and training better models, but there are two easily overlooked reasons affecting the recognition accuracy of the SER system. One is that the low recognition accuracy of one or several emotions will result in a decline of overall accuracy when the number of emotions is large. It can be observed from existing studies that the recognition rate of neutral emotion is low, and it is difficult to distinguish fear and disgust [31,32]. The other is the impact of identification information on emotional understanding. People of different gender, nationality, and age express the same feelings in different ways.

In order to solve above problems, this paper proposes an SER model called multi-layer hybrid fuzzy support vector machine (MLHF-SVM) to improve the overall recognition rate. This model uses SVM-based integrated classifier to train specific emotions that are difficult to recognize. The pre-classification layer assisted on enhanced fuzzy c-means (FCM) is added to classify speech features according to the identification information, and the classified features will be sent to the corresponding classifier. Moreover, we propose an improved particle swarm optimization (PSO)-based FCM clustering method called improved natural exponential inertia weight PSO (IEPSO) to enhance the ability of pre-classification feature parameters and the overall accuracy of the system. Meanwhile, in the feature extraction layer, feature parameters incorporate non-personalized features to ameliorate the universality and adaptability of speech signals. We train and test the model on three commonly used speech emotion databases, and the ability of the model to distinguish emotions has been greatly promoted.

The main contributions of this paper are as follows:

- We propose an MLHF-SVM model, which adopts FCM based pre-classification layer and multi-layer SVM ensemble classifier to increase the recognition accuracy of some specific emotions and reduce the impact of identification information. This method effectively increases the overall emotion recognition accuracy of the SER system.
- We present an improved natural exponential inertia weight PSO (IEPSO) to alleviate the situation that FCM is easy to fall into local minima because of the initial value. This hybrid method (FCM-IEPSO) is used as the pre-classification layer to enhance the effect of pre-classification feature parameters.
- We select three databases for comprehensive simulations, compare the recognition ability of KNN and SVM as the base classifier, and contrast the performance of PSO aided on other inertial weights. The results show that this model efficiently promotes the overall performance of the system.

The remainder of this paper is organized as follows. Section 2 discusses the related studies and tries to reveal the differences and similarities between them. Section 3 briefly introduces the basic structure of the SER system, including feature extraction, feature pre-classification, and feature classification. Section 4 elaborates on the proposed model, other

innovative parts, and their application in this system. Section 5 provides the simulation results and explanation. Section 6 makes the final conclusion.

## 2. Related Work

The amount of research on speech emotion recognition has grown with the extensive applications of SER and accessibility of speech datasets. In order to highlight the research focus and fully compare and analyze the existing literature, we reviewed the following related literature. We particularly focus on the research on emotion recognition based on speech signal rather than multimodality.

In [33], mel-frequency cepstral coefficients (MFCC) and linear predictive cepstral coefficients (LPCC) features were extracted from input speech signals as speech feature parameters, and HMM and SVM were used as classifiers to distinguish emotions according to characteristic parameters. In addition to extracting spectral features such as MFCC and LPCC, reference [31] also extracted other acoustic features such as zero crossing rate (ZCR), energy, and fundamental frequency. The fused features contain more information on speech emotion, which makes the trained classifier more accurate. They utilized SVM and Linear Discriminant Analysis (LDA) assisted classifiers to classify speech signal emotions and tested the model in Berlin Database of Emotional Speech (EmoDB) and RML Emotion Database (RED). Considering that the use of all the features extracted will increase the total workload of SER system and that the influence of existing features on the final recognition rate is not clear, reference [34] refined the existing features by a feature reduction method after feature extraction, and the model increased the success rate of emotion recognition. Reference [35] adopted a bagged ensemble classifier comprising SVM in order to classify reduced features. The results showed that ensemble learning performs more superior than single estimators. Gradually, the design and implementation of deep learning have been more and more feasible. The feature parameter extraction and classification of signals in speech emotion recognition system have relied on deep learning.

In [21], a convolutional neural network (CNN) filter distributed in all spectrum ranges extracted more concentrated frequency domain features to accurately identify emotions, and the average accuracy was 66.1%. In order to improve classification accuracy, some papers optimize the architecture of CNN. Reference [25] modified the pooling strategy of CNN as a filter to learn depth frequency characteristics. The new lightweight effective SER model was trained on the extracted speech frequency features, and the recognition result was 77.01%. Issa et al. [36] extracted MFCC, chromagram, Tonnetz representation, and spectral contrast features from the speech signals as the input of CNN and used an incremental method to optimize SER model. One discovery is that the information of speech emotions is embedded in the long temporal context, and contextual information can assist in judging emotions [37]. Motivated by GoogleNet, Li et al. [24] proposed an attention pooling based model for SER, which utilized two groups of filters to obtain the features with context information in time domain and frequency domain and then fed the features to CNN for classification. In [23], the combination model of CNN and LSTM was used to consider the contextual information in the data. They stacked two layers of LSTM on the top of CNN and conducted end-to-end training. Similarly, Zhao et al. [30] designed a network composed of CNN and LSTM to learn long-term dependencies from the extracted features. A 1D CNN LSTM and a 2D CNN LSTM were constructed to learn local and global features from speech and a log-mel spectrogram, respectively. However, the problem that has not been considered in the above literature is that the recognition accuracy of single emotion or some confusing emotions is lower than that of other emotions, which affects the overall recognition rate.

The recognition accuracy of neural emotion is relatively low, and it is difficult to distinguish fear and disgust when the SER system recognizes speech emotions [32]. Badshah et al. [38] presented a deep CNN composed of three convolutional layers and three fully connected layers to train voice signals. The simulation results showed that the recognition accuracy of fear was still low. The bottleneck features extracted by DNN were trained

by SVM to reduce the confusion of emotion recognition in [39]. This method extracted different bottleneck features for training according to different emotions and effectively reduced the confusion between emotions and increased the accuracy of fear. Reference [40] created a hybrid feature vector composed of acoustic features and depth features extracted from depth network architecture. The hybrid feature vector described speech features more accurately and could improve the success rate of single emotion classification. In order to increase the speech emotion recognition accuracy of fear by extracting more accurate features, K. Zvarevashe and O. Olugbara [41] extracted prosodic features, spectral features, and others to form a hybrid acoustic feature vector. They verified the effectiveness of the features in two benchmark datasets and trained the features on the integrated classifier based on random forest. The review of numerous literature has revealed that there is little literature on training SER model for increasing the recognition accuracy of a single emotion, and most of them optimized the model from the perspective of diversification of characteristic parameters and development of a classification model.

Different from the perspective of the above paper, we add a pre-classification layer to classify the feature parameters according to the identification information and assist pre-classification by using the proposed FCM-IEPSO. Then, the integrated classification is used to train and classify emotions in parallel so as to improve the recognition rate of each emotion.

### 3. Materials and Methods

#### 3.1. Emotional Speech Database

Three databases were adopted to improve the universality and reliability of the study, and all speech samples in the databases were used.

The Berlin Database of Emotional Speech (EmoDB) [42] contains seven emotions, which are anger, boredom, disgust, fear, happiness, sadness, and neutralness. The speech samples were recorded in a studio by ten professional German actors (5 actors and 5 actresses) at a sampling frequency of 16 kHz. The Surrey Audio-Visual Expressed Emotion (SAVEE) Database [43] records the speech samples of four native English-speaking male graduate students and researchers from Surrey University, aged from 27 to 31. The database contains seven emotions: anger, disgust, fear, happiness, neutral, sadness, and surprise. The eNTERFACE'05 [44] is an audio-visual emotion dataset, which contains 42 subjects from 14 different nationalities. The distributions and types of databases are shown in Table 1.

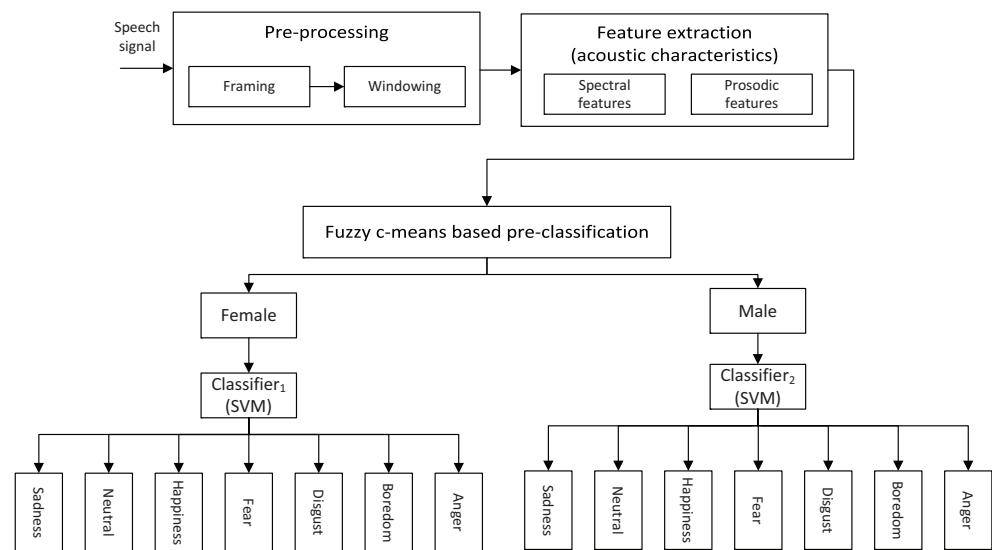
**Table 1.** Distribution and types of datasets used in this paper (Ang—anger; Bor—boredom; Dis—disgust; Fea—fear; Hap—happiness; Neu—neutral; Sad—sadness; Sur—surprise).

Database	Type	Distribution							
		Ang	Bor	Dis	Fea	Hap	Neu	Sad	Sur
EmoDB	Acted	127	81	46	69	71	79	62	-
SAVEE	Acted	60	-	60	60	60	120	60	60
eNTERFACE'05	Elicited	126	-	68	196	198	-	182	205

#### 3.2. Basic Methods

Generally, SER is composed of three steps: preprocessing, feature extraction, and classification. Considering the identification information, this system adds a pre-classification part before the classification step. This process uses FCM to cluster the database into multiple subclasses and then feeds each of them into the feature classification module, respectively. The flow diagram is shown in Figure 1.

In the figure, it is assumed that FCM divides speech signals into two categories according to gender, male and female, and the classifier divides emotions into seven categories based on EmoDB. The specific structure of the classifier is shown in Section 4.



**Figure 1.** Speech emotion recognition (SER) flow diagram including pre-classification layer.

3.2.1. Feature Extraction

Feature extraction is one of the important aspects in the SER system, and the features extracted should accurately reflect the emotional information of the speech. Two feature groups in this paper have been investigated. One is based on fundamental frequency (F0) value, zero-crossing rate (ZCR), and root-mean-square (RMS) signal frame energy. The other originates from MFCC. Meanwhile, in order to reduce excessive computation caused by feature redundancy and to improve the universality of feature parameters, seven statistical functions of each frame speech signal are calculated in this paper. The statistical functions obtained are presented in Table 2.

**Table 2.** Extracted emotional speech features and statistical functions.

Group	Features	Statistical Functions
Prosodic features	F0, ZCR, RMS	maximum value minimum value mean value standard deviation
Spectrum features	1–16 MFCC	skewness value kurtosis value median value

Group 1: F0, ZCR, and RMS energies are the most widely used prosodic features as features that can be perceived by human beings [45–47]. Compared with common features such as LPCC and format, taking features in Group 1 as the feature parameters performs better [48]. F0 reflects the rhythm and intonation of speech in which contour and mean values will change with different emotion. ZCR shows the situation that adjacent samples have different algebraic symbols. RMS energy is highly correlated with emotional state. The energy of high-level arousal emotions (anger, happiness, and surprise) is much higher than that of low-level arousal emotions (sadness and disgust). In addition to these prosodic features, this paper also extracts the spectral features based on MFCC.

Group 2 : MFCC is a spectrum feature that can simulate human auditory perception mechanism, and it extracts parameters through human ear perception of frequency signals [49]. The features based on MFCC have a higher success rate in the classification than those based on other spectral feature linear prediction coefficients (LPC). MFCC provides an envelope that is the real logarithm of the short-term energy spectrum provided. In this paper, sixteen MFCC coefficients were extracted for classification and recognition.

### 3.2.2. FCM Based Pre-Classification

Identification information has an important influence on the understanding of emotions. People with different identification information express their emotions in different ways (i.e., gender, age, and region). Moreover, preliminary classification of speech signal based on identification information is the purpose of pre-classification.

In training the pre-classification model, input data  $B$  all proceed through the feature extraction of speech signal, where  $B = \{b_1, b_2, \dots, b_N\}$ .  $b_j$  represents the  $j$ -th speech feature vector, which contains all its features. After clustering,  $N$  speech signals are clustered into  $C$  subsets. Let the set of subsets be  $S = \{s_1, s_2, \dots, s_C\}$ , and each subset has a cluster center  $s_i$ . Each subset contains several speech signals in  $B$  for which its characteristics are the most similar. This clustering process means that the voice signals with the same identification information are pre-classified into the same subset. In the process of testing, the trained pre-classification model will detect the clustering center closest to the input feature, and the corresponding subset of the center is the result of input speech signal pre-classification. Suppose  $b_2$  is a speech feature vector, the pre-classification result of  $b_2$  can be expressed as follows.

$$b_2 \in \arg \min_{s_i \in S} \|b_2 - s_i\|^2. \tag{1}$$

FCM is the most commonly used clustering method that utilizes the concept of weight to improve the defects of hard clustering [50]. The clustering process of FCM is classified according to the membership degree of each data. The flow chart of FCM pre-classification is shown in Figure 2.

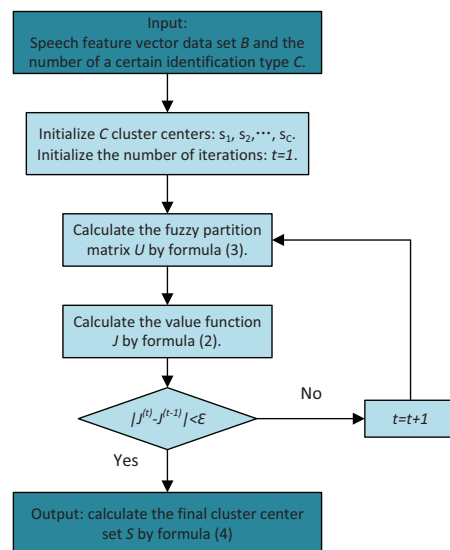


Figure 2. Pre-classification flow chart based on FCM.

The following steps briefly describe the pre-classification of FCM. Firstly,  $C$  cluster centers are randomly initialized according to the preset identification information types, and the membership degree  $u_{ij}$  from data point  $b_j$  to the cluster center  $s_i$  is calculated in order to obtain a fuzzy partition matrix  $U = [u_{ij}]$ . The value function of FCM is defined as follows:

$$J(U, S) = \sum_{i=1}^C \sum_{j=1}^N u_{ij}^m d_{ij}^2 = \sum_{i=1}^C \sum_{j=1}^N u_{ij}^m \|b_j - s_i\|^2$$

$$s.t. \quad \sum_{i=1}^C u_{ij} = 1 \quad \forall j \in 1, 2, \dots, N; \forall i \in 1, 2, \dots, C, \tag{2}$$

$$0 < \sum_{j=1}^N u_{ij} < N \quad \forall j \in 1, 2, \dots, N; \forall i \in 1, 2, \dots, C,$$



where  $d_{ij}$  is the Euclidean distance between  $b_j$  and  $s_i$ , and  $m \in [1, +\infty)$  is the fuzzification parameter that affects the membership degree and adjusts the clustering fuzziness degree. In the iterative process, the algorithm ends only when  $|J^{(t)} - J^{(t-1)}| < \varepsilon$ , where  $\varepsilon$  is the known sensitivity threshold, and  $t$  is the number of iterations. When the termination condition of the algorithm does not satisfy, the membership matrix  $u_{ij}$  and clustering center  $s_i$  are recalculated according to the following formulas, and the iteration continues until the  $J$  value is small enough.

$$u_{ij} = \begin{cases} \frac{d_{ij}}{\sum_{k=1}^C d_{kj}}, & t = 1, \\ \frac{1}{\left(\sum_{k=1}^C \frac{d_{ij}^m}{d_{kj}^m}\right)^{\frac{1}{m-1}}}, & 1 < t \leq MG, \end{cases} \quad (3)$$

$$s_i = \frac{\sum_{j=1}^N u_{ij}^m x_j}{\sum_{j=1}^N u_{ij}^m}. \quad (4)$$

After FCM clustering, each subclass is placed into its subclassifier for the following emotion classification.

The ability of FCM to process high-dimensional data flexibly is very suitable for pre-classification, that is, to classify characteristic parameters according to prior knowledge (identification information). Nevertheless, in the initialization step, FCM randomly selects initial values for iteration, which will affect the final classification results once a bad initial value is selected.

### 3.2.3. Classifier

As shown in Figure 1, the sub-feature sets after pre-classification are used to train corresponding sub-classifiers, respectively, and each sub-classifier trains all features and emotions at the same time. SVM and KNN are the classifiers used in this paper.

SVMs are supervised classifiers based on the idea of binary classification, which find the optimal hyperplane for the linearly separable dataset. The method of SVM is to find the hyperplane that separates the two kinds of data with the largest distance between the classification edges, and this hyperplane is the support vector plane. It can also be applied to multiple classification issues. If the dataset is linearly inseparable, the kernel function is introduced to map the input data to a higher dimensional space for further processing. The definition of the polynomial kernel function is shown below.

$$K(x_i, x_j) = (ax_i^T x_j + c)^d. \quad (5)$$

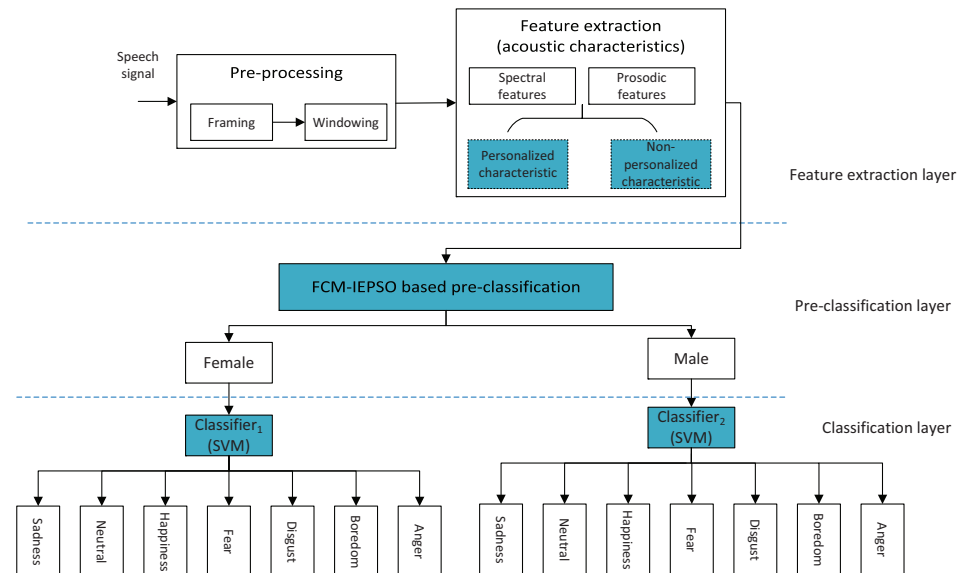
In the kernel function  $K(x_i, x_j)$ ,  $x_i$  and  $x_j$  are any two vectors in the training sample,  $a$  and  $c$  are hyperparameters, and  $d$  represents the degree of the polynomial. In speech emotion recognition, the recognition ability of this technology is relatively strong.

KNN is a basic instance-based supervised learning method, and the distance is calculated by Euclidean distance. Instead of looking at the nearest neighbor category, KNN takes into account the votes of  $k$  nearest neighbors. Specifically, this method selects  $k$  with the smallest distance between test data and training sample, that is, the closest  $k$  sequences, and endows the classification result category to the class with the highest frequency among the  $k$  sequences.

## 4. Proposed Multi-Layer Hybrid Fuzzy Support Vector Machine for SER

A multi-layer hybrid fuzzy support vector machine is presented for emotions that are difficult to classify, and it is optimized in the feature extraction layer and pre-classification

layer to improve its emotion recognition performance. Under the preset conditions in Figure 1, the rough flow of this model is shown in Figure 3. In this section, we will introduce in detail the improved feature extraction layer, pre-classification layer based on proposed FCM-IEPSO, and the classification layer based on multi-layer SVM.



**Figure 3.** Architecture of multi-layer hybrid fuzzy support vector machine for SER. (The improved operation is marked with blue box).

#### 4.1. Improvement of Feature Parameters

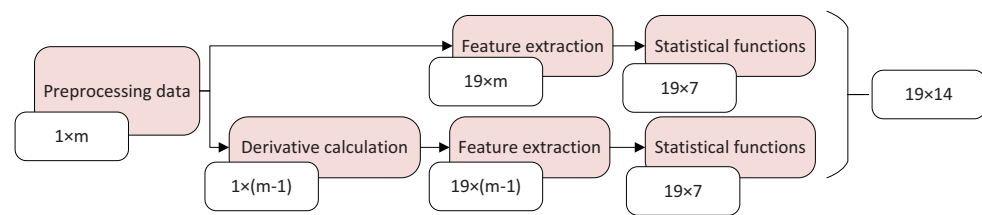
In speech emotion recognition systems, the processing of characteristic parameters usually involves reducing unnecessary parameters or training better parameters. However, the parameters obtained by using these methods are specific to the speech sample and are only generated according to the manner of individual emotion expression, which is not representative and universal. In this paper, derivative-based non-individualized features are added to increase the common ground and regularity of the same category of feature parameter vectors.

Assume that  $m$ -frame data are obtained after a certain speech sample is preprocessed, and the  $m$ -frame data will be extracted through the following steps:

- (1) Extract 19 feature parameters (16 MFCC, F0, ZCR, and RMS) for each frame of input data to obtain  $19 \times m$  parameters. Then, calculate the statistical data of each feature according to Table 2 to obtain  $19 \times 7$  personalized features;
- (2) Calculate the derivative of the input data, and its size is  $m - 1$ . Operate Step 1 to extract the characteristic parameters and corresponding statistical data and obtain  $19 \times 7$  non-personalized features;
- (3) A combination of non-personalized and personalized statistical features lead to  $19 \times 14$  features per sample.

The operation of this process is presented in Figure 4. It is obvious that the extracted feature is high dimensional, and it can describe the speech signal more accurately. The following hybrid method based on FCM and IEPSO has a good effect on processing high-dimensional features.





**Figure 4.** The feature extraction process combines personalized features and non-personalized features (red box represents extraction steps, and white box represents feature number).

#### 4.2. Hybrid Clustering Method Based on FCM and IEPSO

In order to solve the defect that FCM is easy to be affected by initial parameters, which reduces the accuracy of pre-classification, this subsection proposes a hybrid method called FCM-IEPSO. An improved natural exponential inertia weight PSO (IEPSO) is presented to ameliorate the above problem. This method is described in detail below.

##### 4.2.1. Improved Natural Exponential Inertia Weight PSO (IEPSO)

PSO is a population-based optimization algorithm that simulates bird foraging behavior [51]. It seeks the optimal solution through cooperation and information sharing among individuals in the group based on iteration. In each iteration of the algorithm, the optimal particle is searched by updating the particle velocity and position, and the fitness of the particle is determined by the fitness function. Updates of particle velocities require two best positions: personal best position  $pbest$  and global best position  $gbest$ , where  $pbest$  is the best position visited by the current particle and  $gbest$  is the best position visited by all particles. In addition, the inertia weight  $\omega$  and the acceleration constant  $c$  also affect particle velocity. Based on statistical theory analysis, inertia weight has a greater effect [52].

In general, larger inertial weights have better global search capabilities, while smaller inertial weights are more focused on local exploitation. In order to balance exploration capability and development capability, an improved natural exponential inertia weight adjustment strategy based on fitness difference before and after iteration is introduced into PSO, which is IEPSO, and its function is as follows:

$$\omega(t) = \begin{cases} \omega_{\min} + (\omega_{\max} - \omega_{\min}) \cdot e^{-z \cdot \frac{t}{MG}}, & t = 1, \\ \omega_{\min} + (\omega_{\max} - \omega_{\min}) \cdot e^{-z \cdot \frac{t}{MG} \cdot e^{\theta(t)}}, & 1 < t \leq MG, \end{cases} \quad (6)$$

$$\theta(t) = f(t) - f(t-1), \quad (7)$$

where  $\omega_{\min}$  and  $\omega_{\max}$  are the minimum and maximum of the inertial weight, and  $\omega(t)$  is expressed in the inertia weight value of the  $t$ th iteration.  $z$  is a parameter representation for finetuning the inertia weight, taking 20 from experience.

By using this improved inertial weight, the velocity and position of particle  $p$  are updated as follows.

$$V_p(t+1) = \omega V_p(t) + c_1 R_1 (pbest_p(t) - X_p(t)) + c_2 R_2 (gbest(t) - X_p(t)), \quad (8)$$

$$X_p(t+1) = X_p(t) + V_p(t+1). \quad (9)$$

In the above formula,  $c_1$  and  $c_2$  are acceleration factors;  $R_1$  and  $R_2$  are the random numbers in the range  $[0, 1]$  and are independent of each other.

##### 4.2.2. FCM-IEPSO Strategy

FCM always has the disadvantage of being seriously affected by the initial value, which will cause the result to fall into the local minimum value. In this paper, a hybrid fuzzy algorithm combining FCM and IEPSO is proposed, which uses the searching ability of particle swarm to find the appropriate initial value for FCM and avoids falling into the

local optimum. The results indicate that this method optimizes the fitness of FCM and reduces convergence time. The pseudo code of this process is described in Algorithm 1.

As shown, the algorithm consists of two stages, including particle search and clustering. In the first stage, the objective function of FCM updates the position of the particle as the fitness function:  $f_p = J_{\min}$ . Then, the appropriate particles are fed to the next step as initial values, and the optimal clustering centers are obtained by iteration. According to the membership degree, the initial dataset is grouped into class C.

---

#### Algorithm 1 FCM-IEPSO Algorithm

---

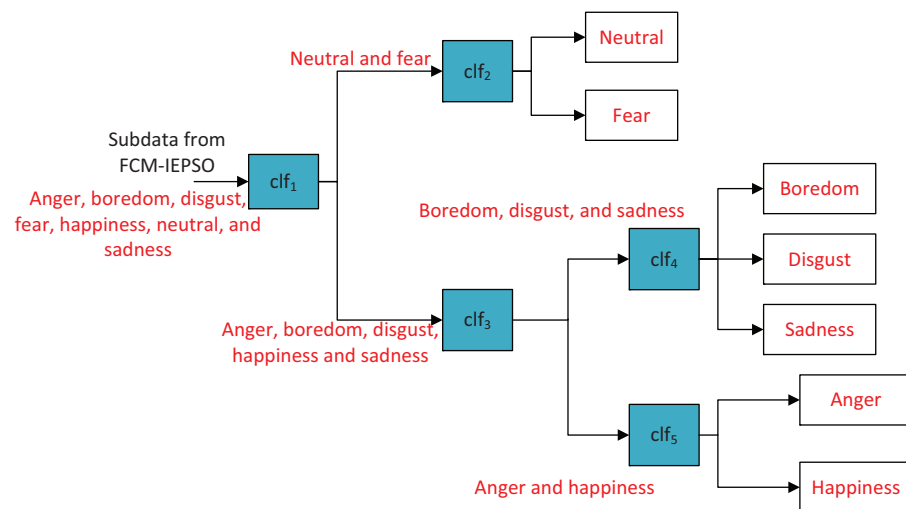
**Input:** The extracted feature parameter set  $B$  and the number of clusters  $C$

**Output:** clustering results  $\{S_1, \dots, S_C\}$  and minimum fitness value  $J_{\min}$

- 1: Initialize parameters:  $MG, c_1, c_2, \omega_{\min}, \omega_{\max}$  and number of particles  $P$  for IEPSO;  $m, \varepsilon$  and maximum number of iterations  $T_{\max}$  for FCM;
  - 2: Calculate the particle position range  $[X_{\min}, X_{\max}]$ ;
  - 3: Initialize each particle ( $p = 1, 2, 3, \dots, P$ ) position  $X_p$ , velocity  $V_p$ ,  $pbest$  and  $gbest$ ;
  - 4: **while** ( $t \leq MG$ ) **do**
  - 5: Calculate fitness function  $f$ , which is FCM objective function  $J$  by (2);
  - 6: Update  $\omega, V_p$  and  $X_p$  by (6)–(9);
  - 7: Update particle position boundary;
  - 8: Update  $pbest_p$  using fitness function;
  - 9: Store  $gbest$  after traversing all  $p$ ;
  - 10:  $t = t + 1$ ;
  - 11: **end while**
  - 12: Obtain particle  $p^*$  as the initialization one for the next iteration;
  - 13: Calculate the membership degree matrix  $u$  and initial cluster centers  $s(p^*)$  by (3) and (4);
  - 14: **while** ( $t \leq T_{\max}$ ) **do**
  - 15: Calculate the objective function  $J(t)$  by (2);
  - 16: **if**  $[(t > 1) \cap (J(t) - J(t - 1) < \varepsilon)]$  **then**
  - 17: break
  - 18: **else**
  - 19:  $t = t + 1$ ;
  - 20: **end if**
  - 21: Calculate  $u$  and  $s$ , respectively;
  - 22: **end while**
  - 23: The dataset  $B$  is divided into  $C$  parts by the final cluster centers and membership matrix;
  - 24: Output clustering results  $\{S_1, \dots, S_C\}$  and  $J_{\min}$ .
- 

#### 4.3. Multi-Layer Hybrid Fuzzy SVM

For the situation where certain emotions are difficult to classify, this paper presents a multi-layer SVM algorithm to increase the recognition accuracy of poorly behaved emotions. The algorithm uses prior information to divide the classification difficulty of each emotion  $E_l (l = 1, 2, \dots, L)$  and trains classifiers containing different categories. Each classifier  $clf_m (m = 1, 2, \dots, M)$  distinguishes at least two classes. It should be noted that the class that is the most difficult to separate needs to be trained first. The priority of classifier training is determined by the difficulty of emotion classification, and its priority also determines the order of test data entering the sub-classifiers. In order to reflect the actual structure of multi-layer classifiers, EmoDB database containing seven emotions is taken as an example. Neutralness and fear, as emotions that are difficult to distinguish, are trained with a separate classifier, while other emotions are trained with another classifier. Since anger and happiness are easiest to distinguish, the classifier trained on them is the lowest level. In order to display the emotion classification of all subclassifiers, the specific structure of multi-layer classifiers based on EmoDB is shown in Figure 5.



**Figure 5.** The structure of multi-layer support vector machine for hard separated emotions. (The categories of emotions are marked in red and the subclassifiers are in blue box).

In this system, SVM is selected as the base classifier of multi-layer classifier model. Combined with above improved methods, the entire process of multi-layer hybrid fuzzy SVM (MLHF-SVM) is shown as follows:

- (1) The training data from the EmoDB database are preprocessed, and then  $19 \times 14$  characteristic parameters from each sample are extracted by using the combined method of non-personalization and personalization;
- (2) Characteristic parameters are clustered into  $C$  subclasses using FCM-IEPSO, and the number of particles is defined as particle number  $\times$  the dimension of feature parameter;
- (3) Train each subdata using multi-layer SVM, and each sub-classifier trains different data;
- (4) Verify the trained model with test data, and integrate the results of the sub-classifiers. The final classification results are obtained by a 5-fold cross-validation method.

## 5. Simulations and Results

### 5.1. Data and Environment Setting

In order to validate the proposed MLHF-SVM model, we have carried out extensive simulations and specific analyses. All the speech samples in these three datasets have been used for simulations. For the purpose of realizing the model independent of speaker and environment, personalized features and non-personalized features of speech were extracted respectively. Sixteen MFCC, F0, ZCR, and RMS features were extracted as feature parameters of each speech, and seven statistical features were calculated with a total quantity of  $19 \times 14$ . Since the gender labels of the three databases were known, we chose gender as a prior information for pre-classification. According to prior knowledge, people of different genders have different methods of expressing emotions. Thus, in the pre-classification stage, the data generated by the feature extraction layer were clustered into two subsets by FCM-IEPSO according to gender, that is,  $C = 2$ . Each subclass was divided into data for training and testing. According to the benchmark database, neutral and fear emotions are often confusing and have low recognition rates. As a result, in the training stage, the classifier of neutral and fear emotions sets and the other five emotions sets were trained first. Then, neutral and fear emotions were trained separately, and the remaining emotions were trained simultaneously. SVM and KNN were used as base classifiers to observe the influence of base classifiers on the results.

In order to verify the validity of the proposed model,  $k$ -fold cross-validation was used for training and testing all data. Considering the amount of data, the cyclic variable  $k$  is set to 5, and the final result is the average value of five cycles. All simulations are implemented

in MATLAB R2018a and run on a Windows 10 operative system with 64-bit support using an Intel Core i5 CPU at 2.30 GHz and 7.85 GB of RAM.

### 5.2. Impact of FCM-IEPSO Based Pre-Classification Layer

In order to analyze the effect of the proposed pre-classification layer, in this subsection, we evaluate the impact of the proposed IEPSO on FCM and the impact of the new pre-classification layer on the overall system recognition accuracy. We test the combination of four types of PSO with FCM: the original PSO, the proposed IEPSO, and two inertia weight variants of PSO. The first variation of the inertia weight, linearly decreasing inertia weight PSO (LPSO) [53], adopts the strategy of monotonically decreasing inertia weight with the number of iterations. The second is natural exponential inertia weight PSO (EPSO) [54], which combines the idea of diminishing inertia weight with a natural exponent to improve convergence speed. The inertia weight is the only variable. When setting parameters, all other parameters except inertia weight are set to the same. Parameter settings are shown in Table 3.

**Table 3.** Parameter settings of PSO and its variants.

Parameter	Value
Popsiz	100
Acceleration factors ( $c_1, c_2$ )	1.49445
The maximum of inertial weight ( $w_{max}$ )	0.9
The minimum of inertial weight ( $w_{min}$ )	0.4
Sensitivity threshold ( $\epsilon$ )	$10^{-5}$
Fuzzification parameter ( $m$ )	2

Tables 4 and 5 compare the  $J$  value [55] calculated from formula (2) and operation time of different clustering algorithms, respectively, on three databases. We compare the system using FCM-PSO, FCM-LPSO, FCM-EPSO, and the proposed FCM-IEPSO method as pre-classification. The smaller the  $J$  value, the better the clustering effect, and the higher the fitness of the algorithm. In order to intuitively observe the results and ensure the randomness of calculation, the average value of the results of 50 runs is calculated, and the best and worst values are selected as auxiliary values. The best values calculated by the different algorithms are highlighted in bold.

**Table 4.**  $J$  value of pre-classification algorithm on three databases. (The best results are highlighted in bold).

Algorithm	EmoDB			SAVEE			eNTERFACE'05		
	Ave	Best	Worst	Ave	Best	Worst	Ave	Best	Worst
FCM-PSO	4656.6	4160.3	4789.9	3546.9	3278.3	3741.6	6231.2	5956.6	6435.2
FCM-LPSO	4236.8	3979.5	4336.9	3215.4	3035.7	3541.5	5796.6	5564.1	6156.3
FCM-EPSO	4196.7	4319.4	<b>4264.5</b>	3054.6	2978.5	3255.6	5642.6	5314.0	<b>5956.3</b>
FCM-IEPSO	<b>4127.1</b>	<b>3959.1</b>	4319.4	<b>2960.9</b>	<b>2890.6</b>	<b>3221.5</b>	<b>5456.8</b>	<b>5134.2</b>	6054.1

**Table 5.** Running time of pre-classification algorithm on three databases.

Algorithm	EmoDB			SAVEE			eNTERFACE'05		
	Ave(s)	Best(s)	Worst(s)	Ave(s)	Best(s)	Worst(s)	Ave(s)	Best(s)	Worst(s)
FCM-PSO	46.68	43.63	47.65	35.64	33.45	37.75	88.61	82.34	90.53
FCM-LPSO	57.17	53.70	58.20	50.03	43.12	53.68	92.68	87.35	94.02
FCM-EPSO	49.88	46.94	52.67	37.98	36.19	42.89	90.32	86.94	92.48
FCM-IEPSO	48.99	47.33	51.12	36.35	33.93	40.68	90.15	85.16	91.44

It can be observed from Table 4 that the  $J$  value and running time are affected by different databases and pre-classification algorithms. However, on average, the FCM-IEPSO method in this paper obtains the smallest  $J$  value, and its running time is relatively short in three different datasets. In terms of operation time, the FCM-IEPSO algorithm performs better on SAVEE, which takes only 0.71 seconds more than the original PSO.

In order to verify the impact of pre-classification layer based on different algorithms on SER system, we evaluate accuracy, F-score, average recall, average precision, and average specificity based on three datasets, as shown in Tables 6–8. In addition to the algorithms covered in Tables 4 and 5, we also compared the system without a pre-classification layer to analyze the impact of this layer. For more comparable results, Tables 6–8 also demonstrate the performance of the pre-classification model from [56].

**Table 6.** Classification test for different pre-classification algorithm on EmoDB.

Algorithm	Index				
	Accuracy (%)	F-Score	Ave Recall	Ave Precision	Ave Specificity
NONE	77.60	0.77	0.77	0.77	0.95
FCM [56]	84.30	0.84	0.84	0.85	0.87
FCM-LPSO	83.27	0.81	0.81	0.84	0.87
FCM-EPSO	86.36	0.85	0.85	0.87	0.89
FCM-IEPSO	90.00	0.90	0.90	0.90	0.91

**Table 7.** Classification test for different pre-classification algorithm on SAVEE.

Algorithm	Index				
	Accuracy (%)	F-Score	Ave Recall	Ave Precision	Ave Specificity
NONE	65.40	0.66	0.66	0.65	0.67
FCM [56]	73.85	0.74	0.73	0.74	0.75
FCM-LPSO	74.54	0.74	0.74	0.75	0.77
FCM-EPSO	78.39	0.78	0.79	0.78	0.80
FCM-IEPSO	80.66	0.81	0.82	0.81	0.84

**Table 8.** Classification test for different pre-classification algorithm on eNTERFACE'05.

Algorithm	Index				
	Accuracy (%)	F-Score	Ave Recall	Ave Precision	Ave Specificity
NONE	58.65	0.59	0.59	0.60	0.62
FCM [56]	64.30	0.64	0.65	0.65	0.67
FCM-LPSO	63.65	0.64	0.64	0.65	0.67
FCM-EPSO	70.96	0.70	0.70	0.70	0.75
FCM-IEPSO	72.47	0.72	0.73	0.72	0.78

Obviously, the accuracy of the system without pre-classification is much lower than that of the system with pre-classification. Compared with the model from [56], our model increases 5.70%, 6.81%, and 8.17%, respectively, on EmoDB, SAVEE, and eNTERFACE'05. The results of FCM derived from LPSO are slightly worse on EmoDB and eNTERFACE'05 and slightly better on SAVEE than that of [56]. Although FCM-LPSO has little impact on system performance, proposed FCM-IEPSO still performs well.

### 5.3. Impact of the Proposed Multi-Layer SVM

The purpose of this subsection is to observe the impact of the proposed multi-layer SVM on the recognition accuracy of individual emotions and the overall recognition rate of the system.

The pre-classification method in this subsection uses the proposed FCM-IEPSO. Figures 6–8 show the confusion matrix of each emotion using multi-layer SVM and SVM [57]

as classifier, respectively. Among the seven emotions on EmoDB, six emotion accuracies increase, and one emotion accuracy (sadness) decreases in Figure 6. The success rates of anger, boredom, disgust, fear, happiness, and neutralness increased by 0.07, 0.01, 0.06, 0.09, 0.01, and 0.11. The recognition accuracies of all seven emotions on SAVEE increased in Figure 7. The values 0.03, 0.04, 0.06, 0.01, 0.04, 0.03, and 0.05 are increased emotion recognition accuracies of anger, disgust, fear, happiness, neutral, sadness, and surprise. Among the six emotions on eNTERFACE'05, the recognition accuracies of five emotions increased, and happiness decreases by 1% in Figure 8. The success rates of anger, disgust, fear, sadness, and surprise increased by 0.07, 0.14, 0.07, 0.05, and 0.04. Moreover, the use of multi-layer SVM greatly reduces the recognition confusion of fear and neutral and improve their success rate, respectively. In EmoDB, the success rate of fear increased from 0.77 to 0.86 and that of disgust with multi-layer SVM reached 0.98.



Figure 6. Confusion matrix with support vector machine (SVM) as classifier on EmoDB. (a) SVM [57], (b) multi-layer SVM.

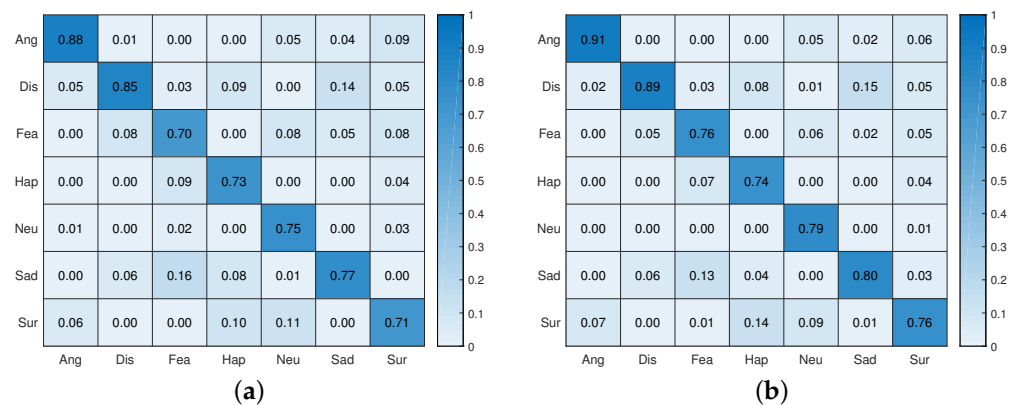
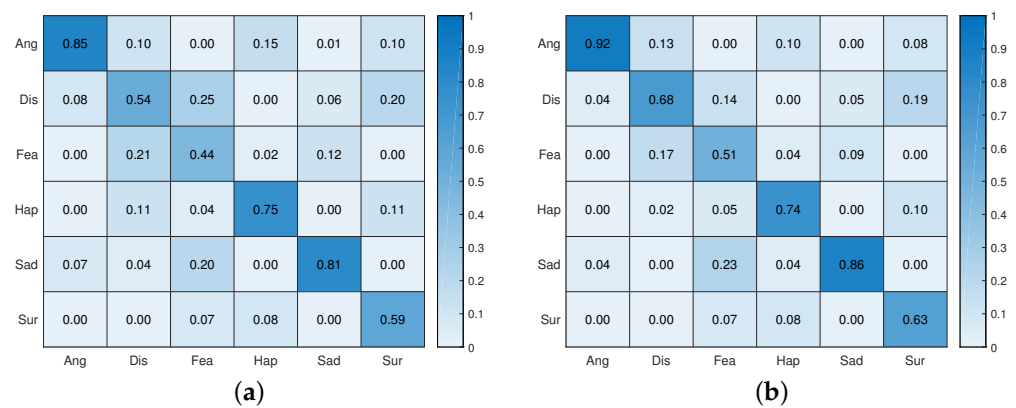


Figure 7. Confusion matrix with support vector machine (SVM) as classifier on SAVEE. (a) SVM [57], (b) multi-layer SVM.





**Figure 8.** Confusion matrix with support vector machine (SVM) as classifier on eINTERFACE'05. (a) SVM [57], (b) multi-layer SVM.

Table 9 compares the impact of using SVM classifier [57] and our proposed multi-layer SVM classifier on overall performance. It can be observed that the accuracies based on EmoDB, SAVEE, and eINTERFACE'05 improved by 4.67%, 3.73%, and 6.04%. Moreover, the classification model has obtained good average specificity on three databases, which are 0.91, 0.84, and 0.78, respectively. Therefore, multi-layer SVM as a classifier improves the recognition accuracy of some indistinguishable emotions so as to make the overall system perform better.

**Table 9.** Classification success rate based on different SVM classifiers.

Classifier	Accuracy (%)	F-Score	Ave Recall	Ave Precision	Ave Specificity
EmoDB					
SVM [57]	85.33	0.85	0.86	0.85	0.87
multi-layer SVM	90.00	0.90	0.90	0.90	0.91
SAVEE					
SVM [57]	76.93	0.77	0.78	0.77	0.81
multi-layer SVM	80.66	0.81	0.82	0.81	0.84
eINTERFACE'05					
SVM [57]	66.43	0.66	0.67	0.66	0.74
multi-layer SVM	72.47	0.72	0.73	0.72	0.78

#### 5.4. Impact of Base Classifier

In the classification layer, we chose SVM as a base classifier. Considering the influence of classifier on SER system, we switch the base classifier of classification layer to KNN for comparison; that is, the classifier becomes multi-layer KNN. Meanwhile, we also compared the impact of using KNN [58] as a classifier and multi-layer KNN on system performance.

The pre-classification method in this subsection uses the proposed FCM-IEPSO. The confusion matrix results of the multi-layer KNN are shown in Figures 9–11, and the overall success rates are shown in Table 10 on three data sets. It can be observed from the confusion matrix that on EmoDB, the success rate of five emotions increased and two emotions decreased; the recognition accuracy of all emotions on SAVEE and eINTERFACE'05 both showed an upward trend.

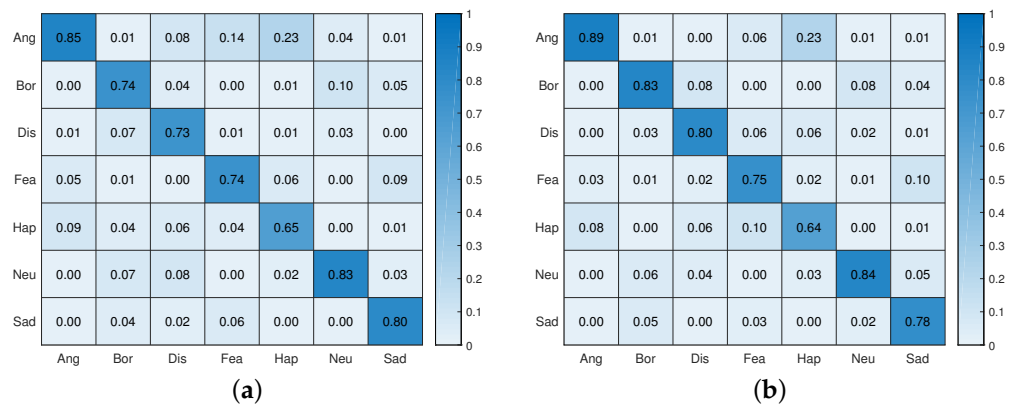


Figure 9. Confusion matrix with K-nearest neighbor (KNN) as classifier on EmoDB. (a) KNN [58], (b) multi-layer KNN.

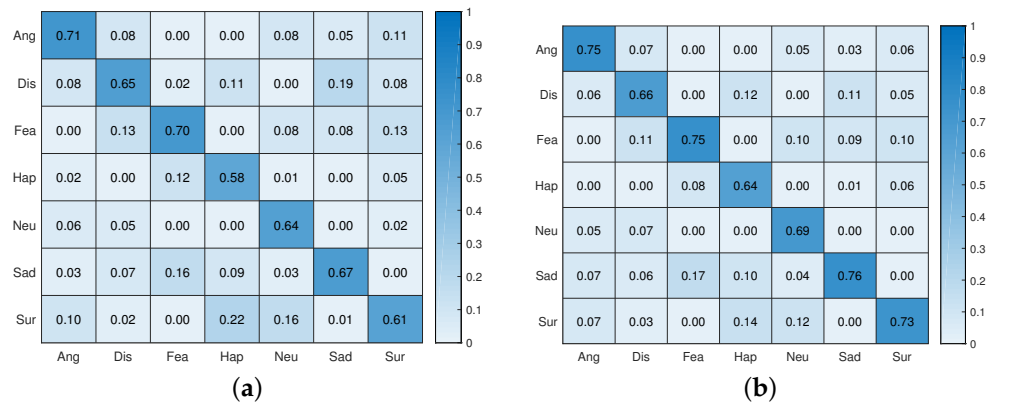


Figure 10. Confusion matrix with K-nearest neighbor (KNN) as classifier on SAVEE. (a) KNN [58], (b) multi-layer KNN.



Figure 11. Confusion matrix with K-nearest neighbor (KNN) as classifier on eINTERFACE'05. (a) KNN [58], (b) multi-layer KNN.

In Table 10, the success rate using the multi-layer KNN classifier increases by 2.94%, 6.09%, and 6.02% on three databases, respectively, compared with a single KNN classifier. Although the accuracy of multi-layer KNN has greatly improved compared that of KNN [58], the overall accuracy of multi-layer KNN is still lower than multi-layer SVM. Therefore, using a multi-layer can enhance the system's performance, and SVM performs better than KNN as the base classifier.

**Table 10.** Classification success rate based on different KNN classifiers.

Classifier	Accuracy (%)	F-Score	Ave Recall	Ave Precision	Ave Specificity
EmoDB					
KNN [58]	76.33	0.76	0.77	0.76	0.81
multi-layer KNN	79.27	0.77	0.78	0.77	0.83
SAVEE					
KNN [58]	65.11	0.65	0.66	0.65	0.73
multi-layer KNN	71.20	0.71	0.72	0.71	0.77
eNTERFACE'05					
KNN [58]	59.40	0.61	0.64	0.59	0.69
multi-layer KNN	65.42	0.61	0.64	0.60	0.70

### 5.5. Impact of Model Performance

In order to evaluate the performance of the proposed MLHF-SVM model, we compare the accuracy of our model with the accuracies of others from [56,57]. Table 11 shows the processes, methods, and recognition accuracies of these models based on EmoDB, SAVEE, and eNTERFACE'05.

**Table 11.** Performance and method comparison with other literature.

Approach	Feature Extraction	Pre-Classification	Classifier	Database	Accuracy (%)
[57]	Spectral features with FCM	-	SVM	EmoDB	80.62
				SAVEE	74.20
				eNTERFACE'05	63.38
[56]	Prosodic features Spectral features	FCM	Random forest	EmoDB	77.97
				SAVEE	72.85
				eNTERFACE'05	64.39
Proposed method	Prosodic features Spectral features	FCM-IEPSO	SVM	EmoDB	85.33
				SAVEE	76.93
				eNTERFACE'05	66.43
Proposed method	Prosodic features Spectral features	FCM-IEPSO	multi-layer SVM	EmoDB	90.00
				SAVEE	80.66
				eNTERFACE'05	72.47

It can be observed from Table 11 that the accuracies of model from [57] reach 80.62%, 74.20%, and 63.38%, and the accuracies of model from [56] reach 77.97%, 72.85%, and 64.39% on three databases. Ref. [57] selected MFCC features using FCM clustering and used an original SVM classifier to recognize speech emotion. They did not consider the impact of certain emotions and identification characteristics on overall performance. Compared with [57], the results of our model with original SVM improved by 4.71%, 2.73%, and 3.05% on EmoDB, SAVEE, and eNTERFACE'05 due to the pre-classification of our FCM-IEPSO algorithm. After adopting multi-layer SVM as the classifier, our model performs better.

Ref. [56] extracted spectral features and prosodic features, preprocessed features with original FCM, and identified emotions with random forest. They did not consider that FCM can easily fall into local minima. The results of our MLHF-SVM model are 12.03%, 7.81%, and 8.08% higher than the model of [56]. Therefore, the FCM-IEPSO algorithm we presented performs better in pre-classification. Compared with these state-of-art technologies, our model has more comprehensive consideration, higher recognition accuracy, and better performance.

## 6. Conclusions

An MLHF-SVM model based on clustering and classification was proposed for speech emotion classification. In MLHF-SVM, in order to alleviate the classification error caused by identified information, the proposed FCM-IEPSO divides the feature datasets into

corresponding subclasses according to prior information, and multi-SVM is presented to distinguish the emotional information of each subset. Meanwhile, in the process of feature extraction, the personalized features and non-personalized features of the speech sample are extracted simultaneously to render the speech signal more representative. The validity of the model is simulated by comparing different clustering algorithms and two classifier models on the EmoDB database. It is observed that the proposed IEPSO performs better than PSO, LPSO, and EPSO in alleviating the initial value and local minimum of fuzzy clustering. When IEPSO and FCM jointly serve for the pre-classification step, FCM-IEPSO performs better than other methods in the speed term, indicating that MLHF-SVM will not induce additional efficiency burdens. When the multi-classifier method is used for classification, the success rate is improved regardless of classifier used. By using the method we suggested, the maximum accuracy rate was achieved at 90.00% with SVM classifiers.

Our future studies will focus on more intelligent algorithms and methods of training classifiers in order to improve the performance of emotion recognition. In recent years, adversarial training, as one of the fields of machine learning, is a new training method for emotion recognition. Furthermore, we will examine techniques for labeling phonetic emotions more accurately in a more natural context.

**Author Contributions:** Methodology and writing—original draft preparation and software, S.H.; writing—review and editing, S.H. and R.J.; data curation and validation, S.H., R.J., Y.H., and C.X.; conceptualization and investigation, Y.H. and C.X.; supervision and investigation, H.D. and W.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

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

## References

1. Li, Y.; Jiang, Y.; Tian, D.; Hu, L.; Lu, H.; Yuan, Z. AI-enabled emotion communication. *IEEE Netw.* **2019**, *33*, 15–21. [[CrossRef](#)]
2. Wioleta, S. Using physiological signals for emotion recognition. In Proceedings of the 2013 6th International Conference on Human System Interactions (HSI), Sopot, Poland, 6–8 June 2013; pp. 556–561.
3. Domínguez-Jiménez, J.A.; Campo-Landines, K.C.; Martínez-Santos, J.C.; Delahoz, E.J.; Contreras-Ortiz, S.H. A machine learning model for emotion recognition from physiological signals. *Biomed. Signal Process. Control* **2020**, *55*, 101646. [[CrossRef](#)]
4. Wu, C.-H.; Liang, W.-B. Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels. *IEEE Trans. Affect. Comput.* **2011**, *2*, 10–21.
5. Wang, K.; An, N.; Li, B.N.; Zhang, Y.; Li, L. Speech Emotion Recognition Using Fourier Parameters. *IEEE Trans. Affect. Comput.* **2015**, *6*, 69–75. [[CrossRef](#)]
6. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Trans. Multimed.* **2018**, *20*, 1576–1590. [[CrossRef](#)]
7. Ye, L.; Liu, T.; Han, T.; Ferdinando, H.; Seppänen, T.; Alasaarela, E. Campus Violence Detection Based on Artificial Intelligent Interpretation of Surveillance Video Sequences. *Remote Sens.* **2021**, *13*, 628. [[CrossRef](#)]
8. Shu, L.; Xie, J.; Yang, M.; Li, Z.; Li, Z.; Liao, D.; Xu, X.; Yang, X. A review of emotion recognition using physiological signals. *Sensors* **2018**, *18*, 2074. [[CrossRef](#)]
9. Bosch, E.; Oehl, M.; Jeon, M.; Alvarez, I.; Healey, J.; Ju, W.; Jallais, C. Emotional GaRage: A workshop on in-car emotion recognition and regulation. In Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Toronto, ON, Canada, 23–25 September 2018; pp. 44–49.
10. Bojanić, M.; Delić, V.; Karpov, A. Call redistribution for a call center based on speech emotion recognition. *Appl. Sci.* **2020**, *10*, 4653. [[CrossRef](#)]
11. Schulz, M.-A.; Yeo, B.T.T.; Vogelstein, J.T.; Mourao-Miranada, J.; Kather, J.N.; Kording, K.; Richards, B.; Bzdok, D. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nat. Commun.* **2020**, *11*, 1–15. [[CrossRef](#)] [[PubMed](#)]
12. Jiang, R.; Fei, Z.; Cao, S.; Xue, C.; Zeng, M.; Tang, Q.; Ren, S. Deep Learning-Aided Signal Detection for Two-Stage Index Modulated Universal Filtered Multi-Carrier Systems. *IEEE Trans. Cogn. Commun. Netw.* **2021**, *1*. [[CrossRef](#)]
13. Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; Bennamoun, M. Deep Learning for 3D Point Clouds: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4338–4364. [[CrossRef](#)] [[PubMed](#)]
14. Jiang, R.; Wang, X.; Cao, S.; Zhao, J.; Li, X. Deep Neural Networks for Channel Estimation in Underwater Acoustic OFDM Systems. *IEEE Access* **2019**, *7*, 23579–23594. [[CrossRef](#)]

15. Tian, C.; Fei, L.; Zheng, W.; Xu, Y.; Zuo, W.; Lin, C.-W. Deep learning on image denoising: An overview. *Neural Netw.* **2020**, *131*, 251–275. [[CrossRef](#)]
16. Zhao, J.; Jiang, R.; Wang, X.; Gao, H. Robust CFAR Detection for Multiple Targets in K-Distributed Sea Clutter Based on Machine Learning. *Symmetry* **2019**, *11*, 1482. [[CrossRef](#)]
17. Anila, R.; Revathy, A. Emotion recognition using continuous density HMM. In Proceedings of the 2015 International Conference on Communications and Signal Processing (ICCCSP), Melmaruvathur, India, 10–11 October 2015; pp. 919–923.
18. Trabelsi, L.; Amami, R.; Ellouze, N. Automatic emotion recognition using generative and discriminative classifiers in the GMM mean space. In Proceedings of the 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, Tunisia, 21–23 March 2016; pp. 767–770.
19. Dahake, P.P.; Shaw, K.; Malathi, P. Speaker dependent speech emotion recognition using MFCC and Support Vector Machine. In Proceedings of the 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, India, 9–10 September 2016; pp. 1080–1084.
20. Lanjewar, R.B.; Mathurkar, S.; Patel, N. Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques. *Procedia Comput. Sci.* **2015**, *49*, 50–57. [[CrossRef](#)]
21. Bertero, D.; Fung, P. A first look into a convolutional neural network for speech emotion detection. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5115–5119.
22. Darekar, R.V.; Dhande, A.P. Emotion recognition from Marathi speech database using adaptive artificial neural network. *Biol. Inspired Cogn. Archit.* **2018**, *23*, 35–42. [[CrossRef](#)]
23. Tzirakis, P.; Zhang, J.; Schuller, B.W. End-to-end speech emotion recognition using deep neural networks. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5089–5093.
24. Li, P.; Song, Y.; McLoughlin, I.V.; Guo, W.; Dai, L.-R. An attention pooling based representation learning method for speech emotion recognition. *Int. Speech Commun. Assoc.* **2018**. [[CrossRef](#)]
25. Anvarjon, T.; Kwon, S. Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. *Sensors* **2020**, *20*, 5212. [[CrossRef](#)]
26. Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; Manocha, D. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 1359–1367.
27. Wagner, J.; Andre, E.; Lingensfelder, F.; Kim, J. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Trans. Affect. Comput.* **2011**, *2*, 206–218. [[CrossRef](#)]
28. Wu, X.; Zheng, W.-L.; Lu, B.-L. Investigating EEG-based functional connectivity patterns for multimodal emotion recognition. *arXiv* **2020**, arXiv:2004.01973.
29. Yu, Y.; Kim, Y.-J. Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database. *Electronics* **2020**, *9*, 713. [[CrossRef](#)]
30. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323.
31. Semwal, N.; Kumar, A.; Narayanan, S. Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models. In Proceedings of the 2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), New Delhi, India, 23–24 February 2017; pp. 1–6.
32. Khan, A.; Roy, U.K. Emotion recognition using prosodie and spectral features of speech and Nave Bayes Classifier. In Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 22–24 March 2017; pp. 1017–1021.
33. Chenchah, F.; Lachiri, Z. Acoustic emotion recognition using linear and nonlinear cepstral coefficients. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **2015**, *6*, 1–4. [[CrossRef](#)]
34. Özseven, T. A novel feature selection method for speech emotion recognition. *Appl. Acoust.* **2019**, *146*, 320–326. [[CrossRef](#)]
35. Bhavan, A.; Chauhan, P.; Shah, R.R. Bagged support vector machines for emotion recognition from speech. *Knowl.-Based Syst.* **2019**, *184*, 104886. [[CrossRef](#)]
36. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control* **2020**, *59*, 101894. [[CrossRef](#)]
37. Latif, S.; Rana, R.; Qadir, J.; Epps, J. Variational autoencoders for learning latent representations of speech emotion: A preliminary study. *arXiv* **2017**, arXiv:1712.08708.
38. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech emotion recognition from spectrograms with deep convolutional neural network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Korea, 13–15 February 2017; pp. 1–5.
39. Sun, L.; Zou, B.; Fu, S.; Chen, J.; Wang, F. Speech emotion recognition based on DNN-decision tree SVM model. *Speech Commun.* **2019**, *115*, 29–37. [[CrossRef](#)]
40. Er, M.B. A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features. *IEEE Access* **2020**, *8*, 221640–221653. [[CrossRef](#)]

41. Zvarevashe, K.; Olugbara, O. Ensemble learning of hybrid acoustic features for speech emotion recognition. *Algorithms* **2020**, *13*, 70. [[CrossRef](#)]
42. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the 9th European Conference on Speech Communication and Technology, Lisboa, Portugal, 4–8 September 2005; pp. 1517–1520.
43. Jackson, P.; Haq, S. *Surrey Audio-Visual Expressed Emotion (Savoe) Database*; University of Surrey: Guildford, UK, 2014.
44. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The eINTERFACE'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.
45. Zeng, Z.; Pantic, M.; Roisman, G.I.; Huang, T.S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 39–58. [[CrossRef](#)]
46. Busso, C.; Lee, S.; Narayanan, S. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *17*, 582–596. [[CrossRef](#)]
47. Philippou-Hübner, D.; Vlasenko, B.; Böck, R.; Wendemuth, A. The performance of the speaking rate parameter in emotion recognition from speech. In Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops, Melbourne, VIC, Australia, 9–13 July 2012; pp. 296–301.
48. Wang, K.X.; An, N.; Li, L. Emotional speech recognition using a novel feature set. *J. Comput. Inf. Syst.* **2013**, *9*, 1–8.
49. Davis, S.; Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **1980**, *28*, 357–366. [[CrossRef](#)]
50. Dunn, J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *J. Cybern.* **1973**, *3*, 32–57. [[CrossRef](#)]
51. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the ICNN'95-International Conference on Neural Networks, Perth, WA, Australia, 27 November–1 December 1995; pp. 1942–1948.
52. Peng, Y.; Peng, X.-Y.; Zhao-Qing, L. Statistic analysis on parameter efficiency of particle swarm optimization. *Acta Electron. Sin.* **2004**, *32*, 209–213.
53. Eberhart, R.C.; Shi, Y. Comparing inertia weights and constriction factors in particle swarm optimization. In Proceedings of the 2000 Congress on Evolutionary Computation, CEC00 (Cat. No. 00TH8512), La Jolla, CA, USA, 5–9 June 2000; pp. 84–88.
54. Chen, G.; Huang, X.; Jia, J.; Min, Z. Natural exponential inertia weight strategy in particle swarm optimization. In Proceedings of the 2006 6th World Congress on Intelligent Control and Automation, Dalian, China, 21–23 June 2006; pp. 3672–3675.
55. Izakian, H.; Abraham, A. Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. *Expert Syst. Appl.* **2011**, *38*, 1835–1838. [[CrossRef](#)]
56. Chen, L.; Su, W.; Feng, Y.; Wu, M.; She, J.; Hirota, K. Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Inf. Sci.* **2020**, *509*, 150–163. [[CrossRef](#)]
57. Demircan, S.; Kahramanli, H. Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech. *Neural Comput. Appl.* **2018**, *29*, 59–66. [[CrossRef](#)]
58. Liogienė, T.; Tamulevičius, G. Multi-stage recognition of speech emotion using sequential forward feature selection. *Sci. J. Riga Tech. Univ. Electr. Control Commun. Eng.* **2016**, *10*, 35–41. [[CrossRef](#)]