Received: January 15, 2023; **Accepted:** August 02, 2023.
DOI. 10.30473/COAM.2023.66718.1226
Summer-Autumn (2023) Vol. 8, No. 2, (85-105)
Research Article

Open Access

Control and Optimization in Applied Mathematics - COAM

Emotion Recognition for Persian Speech Using Convolutional Neural Network and Support Vector Machine

Saeed Hashemi, **Saeed Ayat***

Department of Computer Engineering and Information Technology, Payame Noor University (PNU), Tehran, Iran.

Correspondence:
Saeid Ayat
E-mail:
dr.ayat@pnu.ac.ir

Abstract. The paper discusses the limitations of emotion recognition in Persian speech due to inefficient feature extraction and classification tools. To address this, we propose a new method for detecting hidden emotions in Persian speech with higher recognition accuracy. The method involves four steps: preprocessing, feature description, feature extraction, and classification. The input signal is normalized in the preprocessing step using single-channel vector conversion and signal resampling. Feature descriptions are performed using Mel-Frequency Cepstral Coefficients and Spectro-Temporal Modulation techniques, which produce separate feature matrices. These matrices are then merged and used for feature extraction through a Convolutional Neural Network. Finally, a Support Vector Machine with a linear kernel function is used for emotion classification. The proposed method is evaluated using the Sharif Emotional Speech dataset and achieves an average accuracy of 80.9% in classifying emotions in Persian speech.

Keywords. Emotion recognition in speech, Mel-Frequency cepstral coefficients, Convolutional neural network, Support vector machine.

MSC. 68XX.

How to Cite
Hashemi, S., Ayat, S. (2023). "Emotion recognition for Persian speech using convolutional neural network and support vector machine", Control and Optimization in Applied Mathematics, 8(2): 85-105.

1 Introduction

Speech is a highly effective tool for making fast communication and eliciting reactions among humans. It can also be utilized for communication between humans and machines. However, expressing various feelings and emotions during speech is exclusive to humans, and recognizing different speech emotions can be challenging for machines [17]. While recognizing emotions from speech characteristics may be relatively simple for humans, it can be challenging for emotionless machines. The characteristics of sound, accent, and the way of expressing emotions vary in different languages, affecting the accuracy of speech emotion recognition [18]. This is especially true for the Persian language, which has several sound characteristics, dialects, and accents. Despite the numerous studies conducted on Speech Emotion Recognition (SER) in different languages, only a small number of them have focused on the Persian language [16]. Furthermore, research on the recognition of emotions in Persian speech has faced several challenges, such as insufficient use of efficient and diverse methods for feature extraction and classification, highlighting a research gap in the field of emotion recognition in Persian speech.

Recognizing emotions in speech is a complex task that presents many challenges. One of these challenges is the complexity of emotional states, as some emotional states are the result of combining two or more basic emotional states [1]. Therefore, it is crucial to be able to distinguish basic emotional states in speech. Moreover, recognizing emotions in Persian speech is particularly important due to the lack of an accurate method for achieving this task, which is a result of using inefficient feature extraction and classification tools. In this paper, we present an efficient method for recognizing emotions in Persian speech that overcomes existing challenges and is expected to yield better results. Our proposed method combines signal processing and machine learning techniques. The combination of features used in our method for identifying emotions in speech is one of the innovative aspects of our research.

The remainder of this paper is organized as follows: In the second section, we review related works. The third section provides a detailed description of the proposed method. In the fourth section, we discuss our research findings. Finally, the fifth section presents our conclusions based on the research findings.

2 Related Works

The recognition of emotions in speech has a well-established research background, with most studies focusing on English speech. In this section, we discuss recent efforts in this area. Ke et al. [9] employed two models, Artificial Neural Network (ANN) and Support Vector Machine (SVM), to recognize emotions in Chinese speech using the CASIA Chinese Emotional Corpus database. The authors evaluated the effects of reducing feature dimensions, comparing the effect of feature reduction on these two models. Alghifari, Gunawan, and Kartiwi [2] employed the Deep Neural Network (DNN) classification method on a custom database to recognize speech emotions, considering only the Mel-frequency cepstral coefficients (MFCC) features for describing attributes of speech.

Kumbhar and Bhandari [10] used the MFCC features, along with a feature reduction mechanism, to analyze speech signals. Due to the high generality of MFCC features compared to other features, the authors claimed that their method is a language-independent approach. In this method, 39 coefficients

were extracted for each speech signal, and a Long Short-Term Memory (LSTM) was used to recognize emotions.

Ravanbakhsh et al. [15] used Short-Time Fourier Transform (STFT) and MFCC features to extract emotional-related features from speech and classified these features using an ANN. The Berlin database with 535 audio files in the German language was used in their experiments. This database involved seven basic emotions of happiness, sadness, disgust, anger, surprise, fear, and neutrality. Their research showed that the efficiency of emotion recognition depends on three factors: training algorithm, extracted features, and emotion type. According to their results, using certain training functions in ANN, the STFT features led to a better recognition rate for some emotion classes, while MFCC features could outperform STFT for other classes.

Fahad et al. [6] compared DNN and Hidden Markov Model (HMM) classification methods using two MFCC and Epoch-based features. They compared these cases using four emotional states of happiness, sadness, anger, and neutrality in the IEMOCAP database. Their results showed the superiority of DNN when fed by MFCC features. Horkus and Guerti [7] proposed a method for recognizing anger's emotional state using an SVM classifier with three types of features: MFCC, formant, and prosodic features (such as the lowest and highest pitch). They also examined cases where the combination of these three features was used. This method was tested on four datasets, including the *Sharif Emotional Speech* dataset (ShEmo). However, this paper is limited to the detection of neutral and angry emotional states, and the highest accuracy is achieved when the combination of features is used for emotion recognition.

Liu et al. [11] demonstrated that the classification method can be sensitive to a small number of phonetic labels clustered by the K-Means method, which may ignore other feature components. These specific phonetic components are taken from the MFCC feature and speech emotions are recognized by DNN and SVM classification methods using six datasets, including ShEmo. The dimensionality reduction approach aimed at reducing calculations, execution time, and cost, although it did not achieve high accuracy.

3 The Proposed Method

The proposed solution for recognizing emotions in Persian speech can be summarized in the following phases:

1. Preprocessing audio signals,
2. Describing features using MFCC and Spectro Temporal Modulation (STM),
3. Extracting features using Convolutional Neural Network (CNN),
4. Classifying extracted features using SVM.

The steps of the proposed method are illustrated in Figure 1.

According to this diagram, the proposed method commences with preprocessing the input signals. The primary objective of the preprocessing phase is to convert all input signals into a standardized intermediate form. To achieve this, the input signals are first converted into mono-channel vectors. Next, the frequency of all signal vectors is converted to 16 kHz. At the end of the preprocessing phase, the

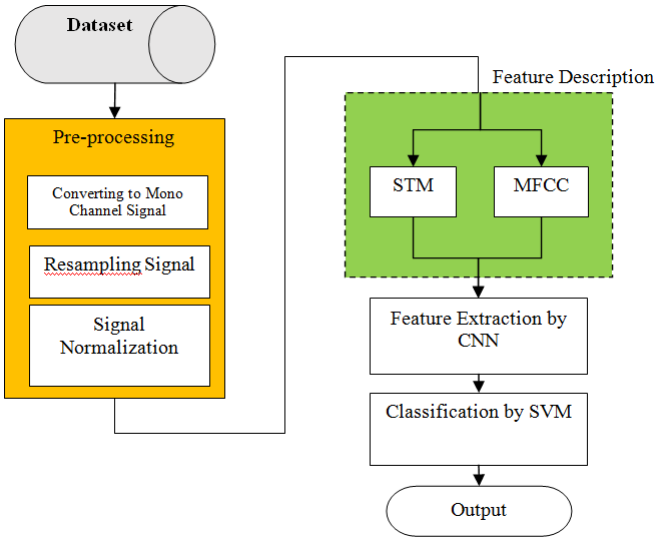


Figure 1: Diagram of the proposed method.

signal vector is normalized by transforming it into a vector with zero mean and unit variance. In the second phase of the proposed method, the feature description process is performed. Two sets of features are utilized to describe speech characteristics: MFCC and STM features. Each of these techniques processes the speech signal independently and produces a matrix of corresponding features. The resulting matrices are merged to perform feature extraction in the third phase of the proposed method. The feature extraction process is executed in the third phase of the proposed method using deep learning techniques. In this phase, the feature matrix obtained from the combination of MFCC and STM for each input speech signal is processed using a CNN, and the characteristics of the corresponding signal are described in the form of a set of features. In the last phase of the proposed method, the extracted features are classified by a set of SVM models that cooperate through the one-vs-all mechanism. In the following, we will expound on the details of each phase of the proposed method.

3.1 The Preprocessing Phase

The proposed method commences with preprocessing the input signals. This phase comprises three primary steps:

- Vectorizing stereo signals,
- Resampling for frequency conversion,
- Normalizing the signal.

The preprocessing phase aims to eliminate superfluous information from the input signal, standardize all signals into an intermediate form, and prepare them for further processing. To achieve this, the operations of audio signal normalization, signal resampling, and conversion of stereo signals to mono-channel vector form are carried out. At the beginning of the preprocessing phase, the multi-channel

nature of the input audio signal is checked. If it is multi-channel, it is converted to a mono-channel signal. Since audio signals are recorded under diverse conditions and using different devices, the frequency of some input signals may differ from others. Two signals with different frequencies describe the same temporal data with different data rates. The variation in the number of samples describing the time unit may cause inaccuracies in the detection process during the subsequent phases of the proposed method. Hence, in the second step of the preprocessing phase, the input signals are converted to a uniform frequency of 16 kHz. This operation aims to standardize the rate conditions for all input signals. At the end of the preprocessing phase, the signal vector is transformed into a vector with zero mean and unit variance, eliminating the possible specific conditions of a signal such as high or low volume as much as possible.

3.2 Feature Description by MFCC and STM

The second phase of the proposed method is devoted to describing the audio signal's characteristics. To obtain a comprehensive set of features that can describe the emotional characteristics of Persian speech from various perspectives, two different techniques have been employed:

- Mel-Frequency Cepstral Coefficients,
- Spectro-Temporal Modulation.

Each of these techniques processes the input signal independently and describes its characteristics in the form of a matrix. In the following sections, we will expound on the feature description process using these two techniques.

3.2.1 Feature Description by MFCC

The first set of features utilized to describe audio signal characteristics is MFCC. The MFCC technique is modeled based on the behavior of the human auditory system in analyzing an audio signal. One of the reasons for the high efficiency of this technique is its high resolution, which enables the recognition of even minor changes in a signal that can be well recognized through the Cepstral coefficients of Mel frequency. Another strength of this method is the use of discrete cosine transform (DCT), which efficiently summarizes the features while removing the details of the spectral structure. The calculation process for extracting MFCC involves the following steps:

a. Pre-emphasis filter: A high-pass filter is applied to the input signal to remove unwanted spectral effects such as sudden changes in the input signal caused by momentary intense noises and make the signal uniform.

b. Framing, windowing and overlapping: The signal is divided into smaller parts called frames and the characteristics of each frame are extracted. Typically, the size of each frame is between 10 and 50 milliseconds, and the frames overlap each other. The overlapping level between the frames is chosen variably (between 25% and 75% of the frame length). In the proposed method, frame length and overlapping level between frames were experimentally selected as 35 milliseconds

and 40%, respectively, resulting in the best results for Persian speech. The resulting frames are multiplied by a window to reduce the effect of signal discontinuity at the beginning and end of each frame and prevent interference between frames in the frequency domain.

c. Calculation of spectrum and filter bank in Mel scale: To enable easier and faster calculations, the Fourier transform is used to transfer the speech signal to the frequency domain. In this step, the spectrum is estimated using the fast Fourier transform. A non-linear transformation called the Mel scale is used on the speech spectrum to model the sensitivity of the human ear to different frequency domains. The Mel scale demonstrates that the human auditory system assigns more importance to the information related to the lower domain, and for this reason, the signal spectrum is passed through 40 filters with the bandwidth of the Mel scale. The number of filters was determined to achieve the highest accuracy in emotion recognition. These filters simulate the frequency resolution of the human auditory system and overlap in a triangular shape, with the beginning of each filter corresponding to the center frequency of its previous filter and its termination corresponding to the center frequency of the filter after it. The peak of each filter is located at its center frequency.

d. Applying logarithmic and discrete cosine transformation: To reduce the number of components in the feature vector, the logarithmic values □□obtained from the 40 filters are multiplied by DCT. The resulting number of target MFCC coefficients is equal to the number of components in the feature vector. The output of this transformation is called the Cepstrum coefficient. This set of coefficients is less correlated, and the fewer components it contains, the more important the information is. On the other hand, more components in this vector have less information and are, therefore, less important.

e. Calculation of the derivatives of Cepstral coefficients: Cepstral coefficients describe the set of characteristics of the speech signal and can be effective in increasing recognition accuracy. To increase the accuracy of the detection system, these coefficients can be derived by regarding time. Cepstral coefficients model static information in the speech signal and are sensitive to the speech states and the changes that occur in them. On the other hand, derivatives of Cepstral coefficients have dynamic information of transfer between different speech states. In this way, the combination of Cepstral coefficients and their derivatives can effectively increase the richness of speech descriptive features.

3.2.2 Feature Description by STM

The second set of features utilized to describe the characteristics of the speech is STM. This technique also employs an auditory system modeling strategy in the feature description process and comprises two basic processing steps:

- Modeling the human auditory system,
- Generation of temporal modulation based on the auditory spectrum.

In the first step, the human auditory system is modeled, during which the speech signal is converted into a neural pattern, known as the auditory spectrogram. This image is a time-frequency distribution along

the tonotopic axis or logarithmic frequency axis, obtained by applying three stages of transformation stages to the input signal. In the second stage, the content of time modulation is obtained through the audio spectrum and by applying wavelet transformation to each row of the auditory spectrogram. The calculation process for the auditory spectrogram involves three main steps, which we will describe in the following section. It should be noted that these two steps are applied to the preprocessed signal obtained in the first phase of the proposed method.

Modeling the Human Auditory System

The process of modeling the human auditory system consists of three main steps, based on the initial stages of sound processing by humans. In the first step, a constant Q transformation is applied to the input signal. This conversion is accomplished using a filter bank, in which all the filters have a constant ratio between the value of the central frequency and resolution. In the proposed method, 96 overlapped filters are utilized, with linearly and uniformly distributed central frequencies. To distribute these filters correctly, the logarithmic frequency axis is divided into the following four-octave intervals:

1. Octave 1: 100 to 200 Hz,
2. Octave 2: 200 to 400 Hz,
3. Octave 3: 400 to 800 Hz,
4. Octave 4: 800 to 1600 Hz.

The 96 filters are distributed on the logarithmic frequency axis to cover these four octaves. If we denote the logarithmic frequency of this filter bank as f, then the impulse response of each filter can be expressed as $h_{\text{cochlea}}(t, f)$. Given the impulse response caused by each filter and considering $s(t)$ as the input speech signal, the output of the cochlear filter can be described as the following equation [13]:

$$y_{\text{cochlea}}(t, f) = s(t) \star_t h_{\text{cochlea}}(t, f), \tag{1}$$

where $\star t$ denotes convolution in the time domain. In this way, the first stage of modeling the auditory system is completed by calculating the output of the cochlear filter. In the second step, the output obtained from the previous step (i.e., $y_{\text{cochlea}}(t, f)$) is transformed into an auditory neural pattern by a hair cell. Using this process, the cochlear output can be modeled as an intracellular pattern. This transformation can be implemented through the following steps: First, a derivative is taken the concerning time from the output obtained from each filter (as $\frac{\partial y_{\text{coachlea}}}{\partial t}(t, f)$) that this action acts as a high pass filter. Then, by applying a non-linear compression function such as $gh_c(\cdot)$ to the output obtained from the previous step, ion channels can be modeled. The compressor function $gh_c(\cdot)$ is defined as follows [13]:

$$gh_c(f) = \frac{1}{1 + e^{-\gamma * f}} - 0.5. \tag{2}$$

Finally, the output of hair cells in the auditory system can be modeled by utilizing a low-pass filter, $\mu h_{c(t)}(0)$. This filter allows frequencies higher than 4.5 kHz to pass through it. The three steps described in the second stage of auditory system modeling can be represented by the following equation [13]:

$$y_{\text{an}}(t, f) = gh_c\left(\frac{\partial y_{\text{cochlea}}}{\partial t}(t, f)\right) \star t\ \mu h_c(t), \tag{3}$$

where $y_{\text{an}}(t, f)$ represents the auditory neural pattern obtained through processing speech signals. Next, the discontinuities of the response along the logarithmic frequency for the existing auditory neural pattern are determined by applying the lateral inhibition network. This lateral inhibition network can be simulated using the first-order differential in terms of logarithmic frequency as follows [13]:

$$y\text{LIN}\,(t, f) = \max(\frac{\partial y_{an}}{\partial f}\,(t, f)\,, 0). \tag{4}$$

The final step in the process of modeling the human auditory system is to integrate the result of the above equation ($y\text{LIN}\,(t, f)$) over a short range, which can be described as the following equation [13]:

$$\mu_{\text{midbrain}}\,(t; \tau) = e^{-\frac{t}{\tau}}\,u(t). \tag{5}$$

Here, $u(t)$ represents the unit step function and τ specifies a short time constant in the range of 2 to 8 milliseconds. With these explanations, the auditory spectrogram $y\,(t, f)$ can be described as follows [13]:

$$y\,(t, f) = y\text{LIN}\,(t, f) \star t\;\mu_{\text{midb□rain}}\,(t; \tau)\,. \tag{6}$$

The process for modeling the human auditory system is illustrated in Figure 2. The output matrix resulting from the steps described above is an auditory spectrogram, an example of which is shown in the lower part of Figure 2.

Construction of Temporal Modulation

At the higher levels of the human central auditory system, particularly in the primary cortex of the auditory system, the analysis is performed on the auditory spectrum by estimating the signal content. To model the human auditory system's perception of temporal modulation, the proposed method uses the process of analyzing the dimensions of modulation to provide a more detailed view of the spectro-temporal characteristics of speech signals. The previous research has shown the best mechanism to model the human auditory system's perception of time modulation can be achieved by using the logarithmic frequency vector along with the constant Q discriminator [4]. In this way, the constant effect of Q can be efficiently modeled by applying continuous wavelet transformation to each row of the auditory spectrogram in the proposed method [5]. Instead of using the standard spectrogram, the auditory spectrogram is used as the input of the modulation dimension analysis step.

The modulation dimension analysis process consists of two main steps. First, a wavelet filter is applied to each temporal row of the auditory spectrogram $(y(t, f))$ assuming r coefficients [5]:

$$X^{\text{SP}}\,(r, t, f) = \frac{1}{r} y\,(t, f) \star t\;\Psi(-\frac{t}{r}). \tag{7}$$

By applying Equation (7), the output obtained from each cochlear channel can be filtered. To reduce complexity and increase computational efficiency, wavelet filters can be simulated by a filter bank consisting of a set of Gabor filters. Each of these filters can be adjusted for different values □□of spectro-temporal parameters (low to high rates). The modulation rates for the Gabor filter bank are $r = \{2, 4, 8, 16, 32, 64, 128, 256\}$ Hz.

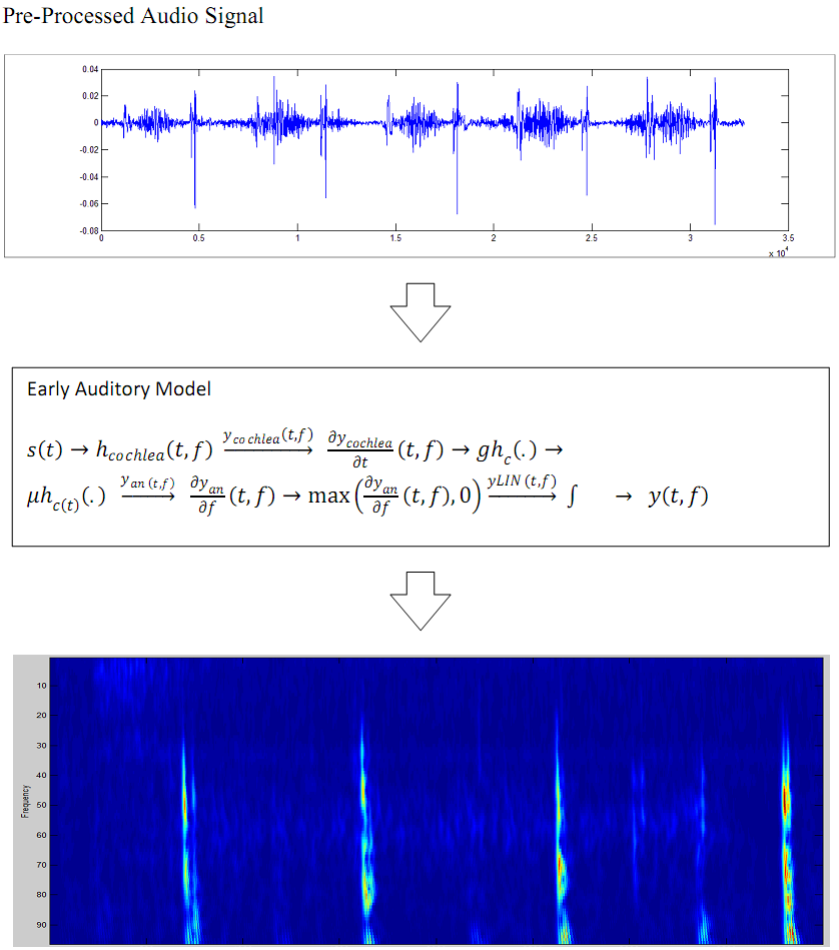


Figure 2: Auditory system modeling diagram.

It should be noted that the output obtained in this step utilizes rate-time-frequency criteria to analyze the input signal. In this way, the spectrum of the resulting audio signal can be depicted as a three-dimensional matrix in terms of rate, time, and frequency. The mentioned filters are applied to each line of this matrix.

The process of temporal modulation production is completed by integrating temporal from the three-dimensional matrix obtained from the previous step. This process can be implemented as an integration of each member of the $X^{\mathrm{SP}}(r,t,f)$ set. By doing this, a two-dimensional model is obtained in terms of rate and frequency [5].

$$X^{\mathrm{JF}}(r,f) = \int \left| X^{\mathrm{SP}}(r,t,f) \right|^2 dt. \tag{8}$$

The obtained two-dimensional model is called auditory temporal modulation. The process of creating temporal modulation based on the audio spectrum is given as a diagram in Figure 3.

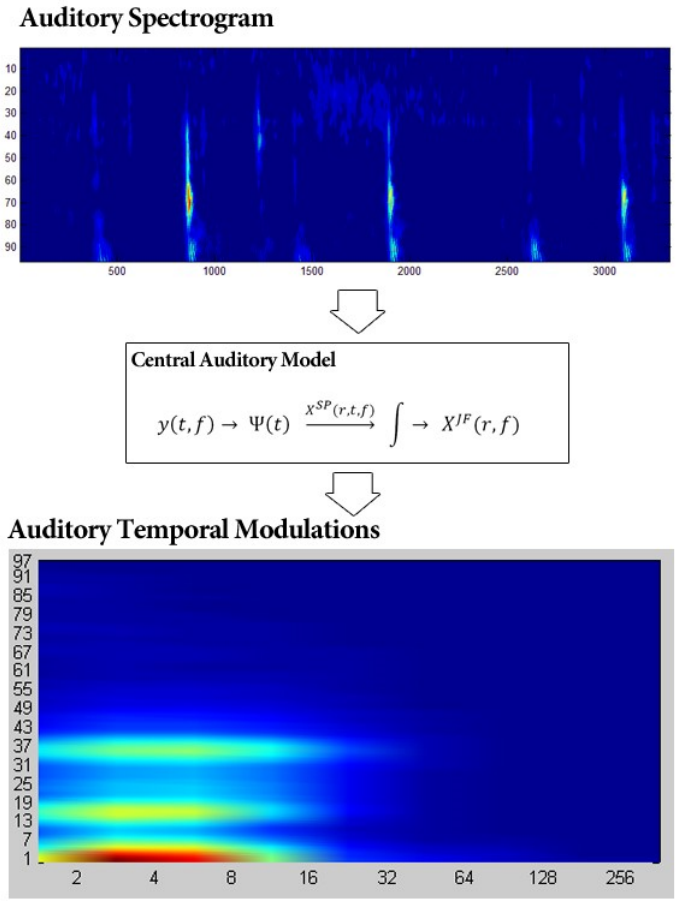


Figure 3: Diagram of the production steps of temporal modulation based on the auditory spectrum.

Figure 3 displays the auditory spectrogram in the upper part and the diagram of the extracted temporal modulation from this signal in the lower part.

The length of all feature vectors, such as the MFCC vector, the STM vector, their integrated feature vector for neural network input, and the CNN output feature vector, are presented in Table 1 after completing all these steps. It should be noted that the input vectors of the neural network have a merged structure with two fields.

3.3 Feature Extraction Using CNN

The proposed method involves a third phase, which entails the extraction of signal features via the use of matrices of MFCC and STM. A CNN is employed to extract the features of each signal. Specifically, the merged matrix resulting from the combination of the two matrices of MFCC and STM serves as the input of the CNN. The weight values □□obtained from the last fully connected layer in this neural network are then used as the final extracted features from the speech signal. It is worth noting that the proposed

Table 1: Length of feature vectors used

Feature vector	Length of the feature vector	Description
STM	96×10	—
MFCC	$13 \times x$, $202 < x < 1992$	Choose the minimum x to merge vectors
Input merged vectors (MFCC, STM)	96×10 merges to $13 \times \min(x)$	2126 samples
CNN neural network output feature vector	128	In the next section

method utilizes multiple layers to optimize the CNN, and if LSTM is employed for feature extraction instead of CNN, the LSTM model may face the issue of gradient vanishing. The structure of this CNN is illustrated in Figure 4.

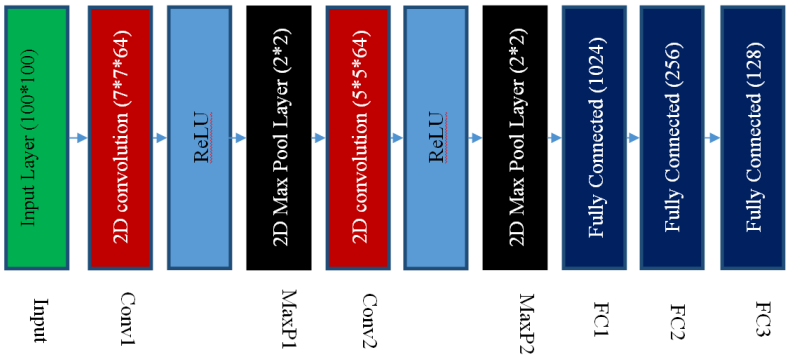


Figure 4: Convolutional neural network structure used in the proposed method for extracting features.

The proposed CNN for feature extraction consists of the following layers:

- One input layer: This layer is the first layer in the proposed CNN and receives the matrix resulting from the integration of the Mel frequency cepstral coefficients and spectral time modulation matrices.
- Two two-dimensional convolution layers: As the input matrices have a two-dimensional structure, the convolution layers used in the proposed deep neural network are also two-dimensional. The dimensions of the convolution filters in the first layer are 7×7 and in the second layer, they are 5×5. Both layers have 64 convolution filters. These values have been determined experimentally and by repeating the experiment for different values. It was found that using more than 64 filters in convolution layers can cause overfitting.
- Two Rectified Linear Unit (ReLU) layers: Each ReLU layer is placed directly after the two-dimensional convolution layers. These layers simulate the function of the activation function to transform the data obtained from the convolution layers. The ReLU function is used in the proposed CNN because of the type of inputs applied to this neural network. Negative outputs

from the convolution layers are converted to zero, while positive outputs are directly transferred to the next layer.

- Two 2D MaxPool layers: These layers are used to reduce the spatial size of the features extracted through convolution layers. The size of the window in these layers is 2×2, reducing the dimensions of the features obtained from the convolution layers by half.
- Three successive fully connected (FC) layers: The purpose of these layers is to transform the features obtained through the previous layers into a vector. The dimensions of the FC layers at the end of the proposed CNN are 1024, 256, and 128 neurons, respectively. The features are expanded and compressed in the second and third layers, respectively. Using this combination of three consecutive FC layers allows for a better abstraction of the features in the input samples.

As shown in Figure 4, the necessary layers for feature classification are not considered in the proposed CNN model. This neural network is only used for feature extraction. The output obtained from the last fully connected layer of the neural network (FC3), which is a numerical vector with a length of 128, is considered as the extracted features from the samples. After extracting the signal features, they are organized in the form of a vector to be used as the input of the learning model in the last step of the proposed method. Based on the set of extracted features, emotions can be recognized in speech.

3.4 Classification by Support Vector Machine

A support vector machine is an efficient supervised classifier for classifying two or more classes. It distinguishes between two classes by drawing a division boundary between them using data belonging to each class [3]. In the last step of the proposed method, An SVM is employed for classification, which can be represented as two very wide regions, a boundary, and a specific position relative to each other. Each wide part belongs to one of the target classes, and the smallest distance between the samples of each class with the border is considered the margin. The SVM is a classification algorithm that increases classification accuracy by maximizing the margin between the support planes of the samples for each class. The algorithm obtains the line that separates the classes by using two parallel lines and the opposite direction of movement so that each line reaches an example of a specific category on its side. Subsequently, a strip or border is formed between these two parallel lines. The wider this strip is, the more the algorithm has been able to maximize the margin, and the goal is to maximize the margin [14]. Since SVM is a binary classifier, each SVM model can classify samples into one of two target classes. To use the SVM model in the multi-class emotion recognition problem, the one-vs-all method has been employed.

4 Numerical Results

To implement the proposed method in this research, MATLAB 2016a software was utilized. The tests were conducted on a desktop computer system running Windows 10 64-bit operating system. This system is equipped with an Intel core i7 processor with a processing power of 3.2 GHz and 16 GB of

memory. To test the performance of the proposed method, the ShEMO database containing 3000 audio files was used [12]. All the speeches in this database are collected by the Sharif University of Technology from radio shows. This database contains a total of 3 hours and 25 minutes of speech data from 78 Persian-speaking speakers, covering the six basic emotions of anger, fear, happiness, sadness, surprise, and neutral state. The emotional state of each part of speech is labeled separately by 12 experts, and the final label of each part of speech is selected based on the majority of the voters. The agreement of the taggers is found to be 14% according to the Kappa criterion, indicating high agreement. Each of the mentioned categories contains samples with the voice of males and females. To classify each sample, 1 second of the signal is processed because signals with a length of less than 1 second might not provide sufficient and effective information to describe the feature with MFCC and STM methods. Therefore, samples with a length of less than one second are ignored, resulting in the removal of 874 samples. Consequently, the experiments carried out in this research are performed using 2126 audio signal samples with a length of more than one second.

In the research experiments, the cross-validation method was employed with 10-fold, 15-fold, and 20-fold divisions. Ultimately, better accuracy results are obtained with 10-fold cross-validation. The samples were divided sequentially, and after 10 repetitions, all the samples were tested. Figure 5, illustrates the correct recognition results for each fold by the proposed model. The results presented in Figure 5 display the percentage of correct detection for each of the 10 repetitions of the experiment. Additionally, Figure 6 shows the results of experiments with different folds and percentages of training and test samples, indicating that the number of repetitions and the percentage of training and test samples for the proposed method are well chosen. As shown in Figures 5 and 6, the proposed method can improve the accuracy of emotion detection compared to other methods. Figure 5 demonstrates that by using the lowest accuracy of correct emotion recognition is 76.6%, the highest accuracy is 84.51%, and the average accuracy is 80.9%. Based on these results, the average accuracy of the proposed method is higher than other learning models. These findings confirm that the techniques used in the proposed method can significantly increase the accuracy of emotion recognition in Persian speech. Furthermore, the proposed method, exhibits a more limited range of changes in the accuracy criterion during different repetitions, in addition to higher average accuracy.

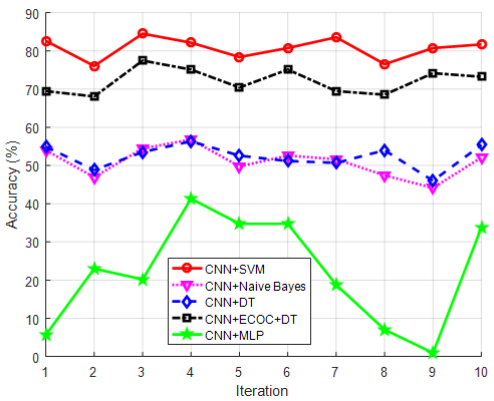
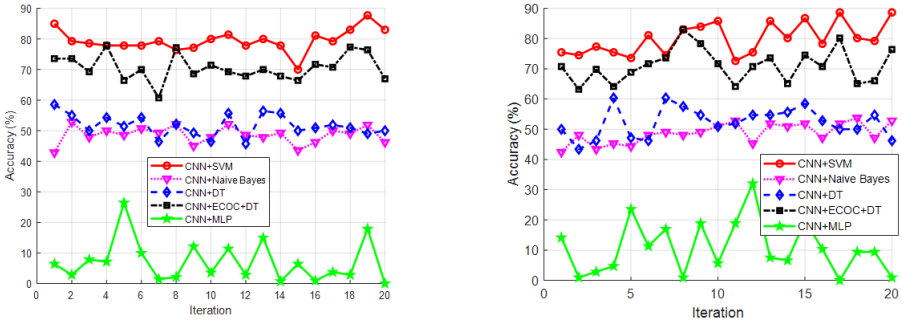


Figure 5: Correct emotion detection of the proposed method during each fold of the experiment.



(a) Accuracy of CNN+SVM = 78.42%. **(b)** Accuracy of CNN+SVM = 80.04%.

Figure 6: (a) The proposed method in 15-fold, and (b) the proposed method in 20-fold.

The results indicate that the proposed method yields a broader range of achievable accuracy values (the lowest and the highest accuracy values). These findings confirm that implementing the proposed model can effectively enhance the reliability of emotion recognition outputs in speech. Figure 7 displays a box plot of the accuracy changes of each classification algorithm, along with their respective accuracy change intervals. In this plot, the dashed range of each box represents the upper and lower limits of the algorithm's accuracy changes during different iterations, while the middle circle in each box represents the median accuracy. As illustrated in Figure 7, the proposed method, apart from having a higher average accuracy, yields more closely bound values □□for the limits of accuracy changes during different iterations. Figure 7 indicates that the proposed method can detect hidden emotions in Persian speech with an accuracy of at least 76.6%.

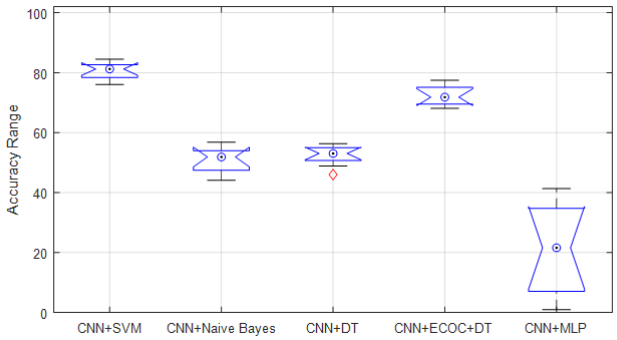


Figure 7: Box diagram of the accuracy of algorithms during 10 repetitions of experiments.

Figure 8 shows the confusion matrix resulting from emotion recognition in the proposed model.

In the confusion matrix presented in this figure, the sum of the values □□of the first column represents the number of test samples belonging to the "Anger" category. Similarly, the sum of values □□in the first row indicates the number of test samples classified by the SVM in the "Anger" category. The intersection of these two sets (first row and first column) displays the total number of correctly classified anger-based samples by the proposed algorithm, which amounts to 732 samples with 92.4% accuracy. The interpretation of the output of the proposed method for other target categories is carried out simi-

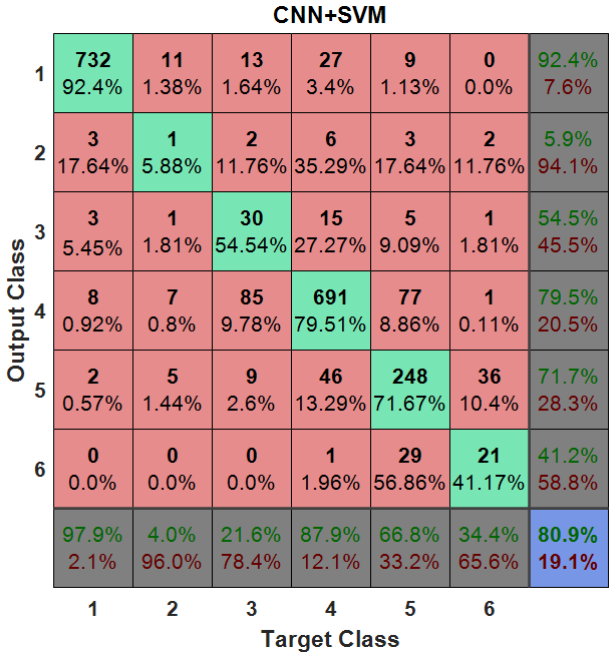


Figure 8: Confusion matrix resulting from emotion recognition by the proposed model.

larly. The proposed method exhibits good recognition in the "natural" and "sad" classes, with accuracies of 79.5% and 71.7%, respectively. Regarding the accuracy of class (6) (surprise), due to the similarity between the extracted features of this class and class (5) (sadness), almost half of the samples in class (6) have been misclassified as class (5) by the model. However, the small number of samples in class (6) also influences the recognition result of this class. Additionally, the low accuracy of class (2) is due to the small number of samples in the database (38 samples among 3000) and the removal of data less than 1 second during the model training process, resulting in a decrease in the number of samples in this class. This affects the accuracy of the average.

These findings demonstrate that, by using the proposed method, 80.9% of the database samples are classified correctly. Furthermore, the confusion matrix results related to other compared algorithms are displayed in Figure 9. Comparing the confusion matrices presented in Figures 8 and 9 reveals that the proposed method outperforms other algorithms in identifying hidden emotions in Persian speech.

The receiver operating characteristic (ROC) curve is a graphical representation created using the true positive rate (sensitivity) on the vertical axis and the false positive rate (specificity) on the horizontal axis at various thresholds (cut points). When the sensitivity increases, the false positive rate also increases. Therefore, the ROC curve allows for the assessment and comparison of the number of true positives and false positives at any point on the curve. The area under the curve (AUC) indicates the overall quality of classification. Tests with the same AUC exhibit overall classification performance, but not necessarily equal sensitivity and specificity. Figure 10 displays the ROC curve for the experiments. Figure 10 illustrates that the area under the ROC curve is greater when using the proposed method compared to other classification algorithms. This graph indicates that the proposed method reduces the

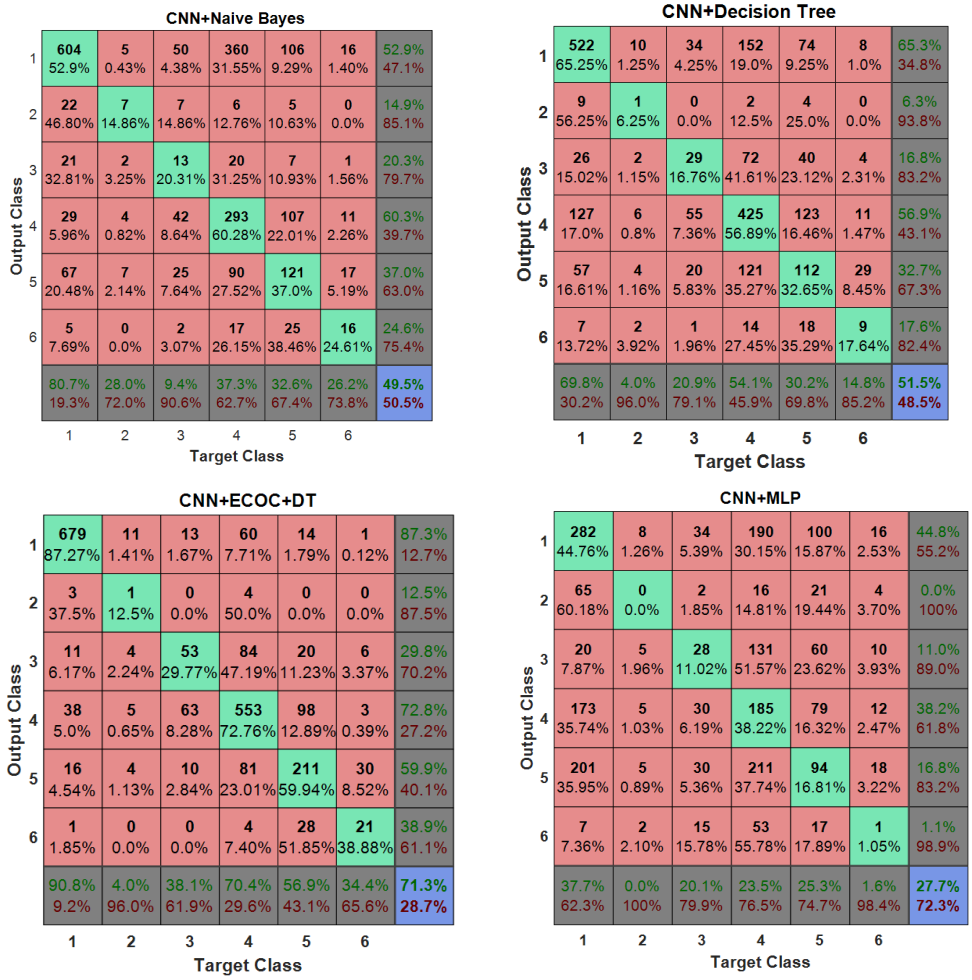


Figure 9: Confusion matrix resulting from (a) Naive Bayes (b) Decision tree (c) ECOC (d) MLP.

false positive (FP) rate and increases the true positive (TP) rate compared to other algorithms, rendering it more effective in accurately detecting emotions.

To assess the model's performance, conventional evaluation criteria should always be applied [8]. To evaluate unbalanced data, the Precision-Recall curve of different classes and the area under their (AUPRC) diagram are also presented in Figure 11. Additionally, the accuracy, recall, and F-Measure criteria can be used to evaluate the proposed method. These results are presented shown in Table 2.

As Table 2 reveals, the proposed method outperforms other algorithms in terms of both the percentage of correct detection and other performance criteria. These findings demonstrate that the proposed method can serve as an effective tool for detecting hidden emotions in Persian speech.

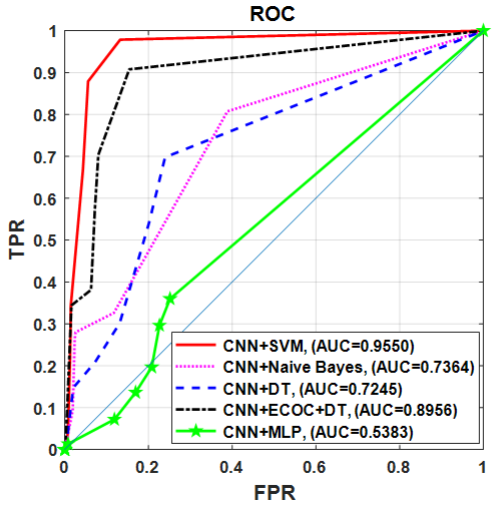


Figure 10: Comparison of the proposed method's ROC curve with other approaches.

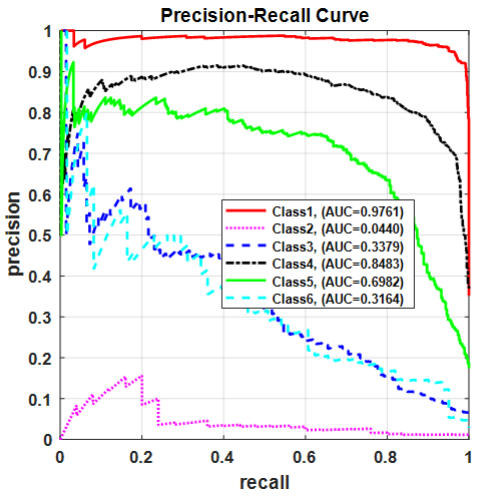


Figure 11: Precision-Recall curves of different classes and their area under the curve (AUPRC).

Table 2: Comparing the efficiency of the proposed method with other classification models

Method	Accuracy	F-measure	Recall	Precision
Proposed (CNN+SVM)	**80.8920**	**53.4894**	**52.1050**	**57.5369**
Naive Bayes	49.4836	33.7230	35.7038	35.0081
Decision Tree	51.5493	32.3001	32.2772	32.5762
ECOC (Error-correcting output coding)	71.2676	49.1532	49.0934	50.1909
MLP (Multi-layer perceptron)	17.4178	11.5732	17.9157	18.8200

5 Discussion

In this research, all available classes and samples from a natural Persian database were utilized, which made our work more challenging than when using synthetic databases, due to the presence of noise in this database that impairs recognition accuracy.

The study conducted by Horkus and Guerti [7] employed the same database as this research but only detected two emotional states "anger" and "neutral" out of six states. To compare the efficiency of the proposed method with that of Horkus and Guerti's [7] study, an experiment was conducted by ignoring other target classes and performing emotion recognition solely based on samples belonging to the two classes, "anger" and "neutral."

Horkus and Guerti's method [7] achieved 90.97% accuracy in recognizing the two emotional states "anger" and "neutral" without separating gender samples. Meanwhile, the proposed method recognized these emotional states with 93.23% accuracy. The results indicate that the proposed method can increase detection accuracy by 2.26% compared to Horkus and Guerti's method. Figure 12 displays the confusion matrix resulting from the detection of the two mentioned emotional states by the proposed method. According to the results of this matrix, for the anger class, 753 samples were correctly identified as true positive (TP) and 21 samples as false positive (FP). Similarly, for the neutral class, 692 samples were correctly identified as TP, and 84 samples were falsely identified as FP. Although information about the confusion matrix of Horkus and Guerti's method [7] is not provided, they stated in their paper that they performed the classification using only 144 samples (72 samples in the neutral category and 72 samples in the anger category). Thus, it is evident that the accuracy of their method would decrease when using all the samples in all six database classes.

Figure 13 compares the accuracy of the proposed method with other evaluation methods used in Horkus and Guerti's study [7] . These results demonstrate that using a combination of temporal spectral modulation features and Capstral coefficients Mel frequency in the proposed method increased the recognition accuracy of the two emotions, "anger" and "neutral", compared to all the methods employed in Horkus and Guerti's study [7] .

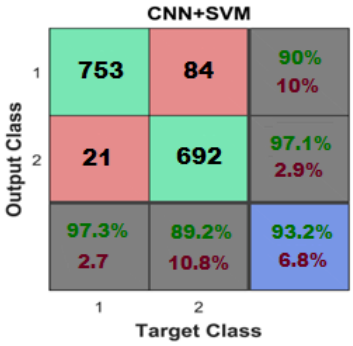


Figure 12: Confusion matrix of the proposed method to recognize two states of anger and neutral.

In another study by Liu et al. [11], the classification was performed using methods such as DNN and SVM, and the phonetic tag method was used to reduce cost and time. However, the average emotion recognition accuracy for the ShEmo database was 70.53%, which is lower accuracy compared to the

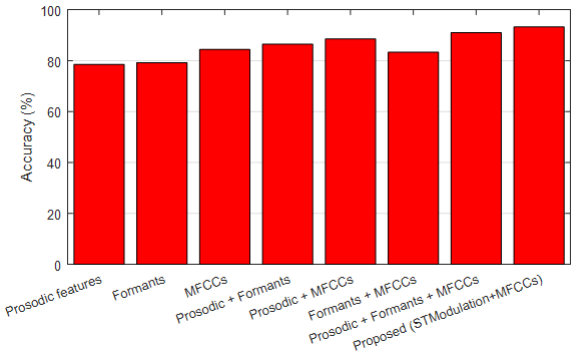


Figure 13: Comparing the accuracy of the proposed method against other approaches.

result of this research, which is 80.9%. As shown in Figure 14 in the confusion matrix of Liu et al.'s [11] study, signals less than 1 second were not removed, and all samples were used in the experiment, which may account for the lower accuracy. The highest recognition accuracies for "anger" and "neutral" emotions in Liu et al.'s study were 82.15% and 75.97%, respectively, which are still lower accuracies compared to the proposed method.

ShEMO

	A	F	H	N	S	U
A	870	1	59	84	20	25
F	5	1	6	5	13	8
H	39	0	80	44	24	14
N	85	0	52	781	70	40
S	28	1	42	84	265	29
U	24	1	16	31	34	119

UAR: 52.08%, WAR: 70.53%

Figure 14: Confusion matrix of research [11] on the ShEmo dataset.

6 Conclusion

This paper proposed a new solution for identifying emotions in Persian speech using machine learning techniques. The proposed method can recognize the six base emotions of "anger", "fear", "happiness", "sadness", "surprise" and "neutral". The method consists of four main phases: preprocessing, feature description, feature extraction, and classification. Two categories of features, MFCC and STM, were used in the proposed method to describe speech characteristics. Each technique independently processes the speech signal and produces a corresponding features matrix. In the third phase, feature extraction is performed using a CNN to merge these two matrices. To evaluate the proposed method's performance in recognizing emotions in speech, the accuracy, precision, recall, and F-Measure criteria were used. Based on the obtained results, the proposed method can recognize the six basic emotional states in Persian speech with an average accuracy of 80.9%, resulting in an increase in accuracy of at least 8.8% compared to other models. In future works, other feature extraction algorithms can be studied to describe speech

features more accurately. Additionally, the use of optimization algorithms to select an optimal subset of speech features could be the subject of future research.

Declarations

Availability of supporting data

All data generated or analyzed during this study are included in this published paper.

Funding

This study received no funds, grants, or other financial support.

Competing interests

The authors declare no competing interests are relevant to the content of this paper.

Authors' contributions

The main manuscript text is collectively written by all the authors.

References

[1] Alabsi, A., Gong, W., Hawbani, A. (2022). "Emotion recognition based on wireless, physiological and audiovisual signals: A comprehensive survey", In International Conference on Smart Computing and Cyber Security: Strategic Foresight, Security Challenges and Innovation, 121-138.

[2] Alghifari, M.F., Gunawan, T.S., Kartiwi, M. (2018). "Speech emotion recognition using deep feedforward neural network", Indonesian Journal of Electrical Engineering and Computer Science, 10 (2), 554-561.

[3] Badie, A., Moragheb, M.A., Noshad, A. (2021). "An efficient approach to mental sentiment classification with EEG-based signals using LSTM neural network", Control and Optimization in Applied Mathematics, 6 (1).

[4] Edraki, A., Chan, W.Y. G., Jensen, J., Fogerty, D. (2019). "Improvement and assessment of spectro-temporal modulation analysis for speech intelligibility estimation", In Interspeech 2019, 1378-1382.

[5] Edraki, A., Chan, W.Y., Jensen, J., Fogerty, D. (2022). "Spectro-temporal modulation glimpsing for speech intelligibility prediction", Hearing Research, 108620.

[6] Fahad, M., Deepak, A., Pradhan, G., Yadav, J. (2021). "DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features", Circuits, Systems, and Signal Processing, 40 (1), 466-489.

[7] Horkous, H., Guerti, M. (2021). "Recognition of anger and neutral emotions in speech with different languages", International Journal of Computing and Digital Systems, 10, 563-574.

[8] Hossin, M., Sulaiman, M.N. (2015). "A review on evaluation metrics for data classification evaluations", International Journal of Data Mining & Knowledge Management Process (IJDKP), 5, 3-9.

[9] Ke, X., Zhu, Y., Wen, L., Zhang, W. (2018). "Speech emotion recognition based on SVM and ANN", International Journal of Machine Learning and Computing, 8 (3), 198-202.

[10] Kumbhar, H.S., Bhandari, S.U. (2019). "Speech emotion recognition using MFCC features and LSTM network", In 2019, 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), 1-3.

[11] Liu, Z.T., Rehman, A., Wu, M., Cao, W.H., Hao, M. (2021). "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence", Information Sciences, 563, 309-325.

[12] Nezami, M.O., Jamshid Lou, P., Karami, M. (2019). "ShEMO: A Large-Scale Validated Database for Persian Speech Emotion Detection", Language Resources & Evaluation.

[13] Panagakis, Y., Kotropoulos, C., Arce, G.R. (2009). "Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification", IEEE Transactions on Audio, Speech, and Language Processing, 18 (3), 576-588.

[14] Pisner, D.A., Schnyer, D.M. (2020). "Support vector machine", In Machine Learning, 101-121, Academic Press.

[15] Ravanbakhsh, M., Setayeshi, S., Pedram, M.M., Mirzaei, A. (2020). "Evaluation of implicit emotion in the message through emotional speech processing based on Mel-frequency Cepstral coefficient and short-time Fourier transform features", Advances in Cognitive Science, 22 (2), 71-81.

[16] Siadat, S.R., Voronkov, I.M., Kharlamov, A.A. (2022). "Emotion recognition from Persian speech with 1D Convolution neural network", In 2022 Fourth International Conference Neurotechnologies and Neurointerfaces (CNN), 152-157.

[17] Tiwari, P., Darji, A.D. (2022). "A novel S-LDA features for automatic emotion recognition from speech using 1-D CNN", International Journal of Mathematical, Engineering and Management Sciences, 7 (1), 49.

[18] Yadav, S.P., Zaidi, S., Mishra, A., Yadav, V. (2022). "Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN)", Archives of Computational Methods in Engineering, 29 (3), 1753-1770.