Hindawi
Journal of Electrical and Computer Engineering
Volume 2023, Article ID 9870204, 1 page
https://doi.org/10.1155/2023/9870204

Retraction

Retracted: Application of an Improved LSTM Model to Emotion Recognition

Journal of Electrical and Computer Engineering

Received 19 December 2023; Accepted 19 December 2023; Published 20 December 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

In addition, our investigation has also shown that one or more of the following human-subject reporting requirements has not been met in this article: ethical approval by an Institutional Review Board (IRB) committee or equivalent, patient/participant consent to participate, and/or agreement to publish patient/participant details (where relevant).

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Li, "Application of an Improved LSTM Model to Emotion Recognition," *Journal of Electrical and Computer Engineering*, vol. 2022, Article ID 3271074, 11 pages, 2022.

Hindawi
Journal of Electrical and Computer Engineering
Volume 2022, Article ID 3271074, 11 pages
https://doi.org/10.1155/2022/3271074

Research Article

Application of an Improved LSTM Model to Emotion Recognition

Yuan Li

School of Computer Science and Information Engineering, Anyang Institute of Technology, Anyang 454000, Henan, China

Correspondence should be addressed to Yuan Li; liyuan@ayit.edu.cn

Received 2 April 2022; Revised 6 May 2022; Accepted 10 May 2022; Published 26 May 2022

Academic Editor: Wei Liu

The rise of artificial intelligence technology has promoted the development of human-computer interaction and other fields. In human-computer interaction, in order to enable the machine to accurately perceive and understand the user's emotion in real time, thereby improving the service quality of the machine, user emotion recognition has been widely studied. In real life, because voice output not only is convenient, but also contains rich emotional information, human-computer interaction is mainly carried out in the form of voice. Speech carries a wealth of linguistic, paralinguistic, and nonlinguistic information that is essential for human-computer interaction. Understanding language information alone will not allow a computer to fully comprehend the speaker's purpose. For computers to behave like humans, speech recognition systems must be able to process nonverbal information, such as the emotional state of the speaker. As a result, developing machine understanding of human emotions requires speech-based emotion recognition. This paper proposes an improved long short-term memory network (ILSTM) for emotion recognition. Because the initial LSTM only analyzes the preceding moment's input, it will miss out on a lot of information for the full context scene. In this way, all the features in the speech segment can be extracted. In order to be able to select the feature that can express emotion the most among the many features, this paper also introduces the attention mechanism. Experiments are carried out on public datasets, and the experimental results show that the ILSTM used in this paper is very effective in classifying speech emotion data and the classification accuracy can reach more than 0.6. This fully shows that this research can be applied to actual products and has certain feasibility and reference value.

1. Introduction

Human-computer interaction is getting more humanized and sophisticated as artificial intelligence and deep learning technology advance. Professor Picard in the United States was the first to develop the concept of affective computing in the 1990s. He claimed that the goal of affective computing is to create a harmonious human-computer ecosystem by giving computers the ability to identify, interpret, express, and adapt to human emotions, as well as to provide computers higher and more complete intelligence. Professor Picard not only introduced affective computing, but also thoroughly examined its definition, the relationship between expression and cultural background, and so on. These factors are seen to be crucial criteria for people to connect smoothly [1, 2]. Now, voice recognition technology is primarily employed to convert speech signals to text information. This research has yielded surprising outcomes. However, just expressing information by text or speech might easily neglect the emotional content of its connotation, and it is impossible to fully understand the user's aim of speaking. When communicating in real life, people catch each other's emotional conditions through tone changes and intonation frustrations in addition to sharing written information. The absence of emotional semantics is implausible. In the CASIA Chinese emotional corpus, for example, "even if the wind blows, go out" can deduce the six emotions of anger, happiness, neutrality, sadness, fear, and surprise. Relying on speech recognition technology alone can lead to poor communication and discomfort during human-computer interaction. Relying on speech recognition technology alone can lead to poor communication and discomfort during human-computer interaction. The differ5ence between speech recognition research and speech emotion recognition research is that the latter pays more attention to the content of emotions in speech signals, because the emotions of

speech signals are closely related to people's emotions. For AI to be more humane and better serve the masses, machines must have the ability to understand human emotions.

In recent years, research on voice emotion identification has had a wide range of applications in the medical, educational, service, and car industries [3, 4]. In the medical profession, for example, speech emotion recognition can detect whether the speaker has symptoms such as depression or autism, allowing for timely psychological counseling and treatment of the patient. The chance of contracting the disease can thus be increased. The speech emotion identification system for the lonely elderly can identify the inner emotions of the elderly in real time and avoid the occurrence of mental disorders in the elderly [5]. It is extremely important in the field of education, particularly online education. Teachers cannot determine students' emotions in real time since they cannot observe their students' movements and expressions in real time during online teaching. Furthermore, as a result of emotional sadness, students will perform poorly in class, hurting their marks. Teachers can improve class quality by monitoring students' learning emotions in real time and modifying teaching methods and content as needed [6]. In the service industry, such as telecommunications, users' perceptions of intelligent machine customer service can be changed by recognizing emotional changes in clients in real time and giving humanized services that are more in accordance with customer wants. Furthermore, the speech emotion recognition system can be used to monitor customer service attitudes and improve customer satisfaction [7]. In the automobile manufacturing industry, emotionally unstable and irritable drivers are more likely to cause traffic accidents due to issues such as rush hour, time constraints, or fatigue driving. The voice emotion recognition system monitors the driver's emotions and sends corresponding reminders to make driving safer [8]. Speech emotion recognition technology has greatly facilitated medical, educational, service, automotive, and other industries. It is apparent that voice emotion recognition research is directly tied to human life. Voice emotion recognition will provide new advancements in the field of human-computer interaction with the continual growth of artificial intelligence and in-depth study of speech emotion recognition. As a result, the study of speech emotion recognition has enormous theoretical and research significance.

Reference [9] discovered in 1972 that emotional conditions have a significant impact on the pitch contour and average power of human speech. Reference [10] investigated the relationship between acoustic features and speech emotion later in the 1980s. Reference [11] discovered that the lowest value of the fundamental frequency of speech increased with cognitive and emotional stress and that the position of formant and pronunciation accuracy were related to emotional changes in female subjects, leading to the use of speech statistical features to identify speech emotion associations. In 1996, Dellaert et al. [12] proposed a pitch contour-based prosodic feature extraction method and applied it to the task of speech emotion classification. The experimental results show that this method performs well in terms of semantic emotion recognition. In 1999, Moriyama and Ozawa [13] created a system for recognizing and synthesizing emotional content in speech using simple linear operations on speech feature information associated with emotion, which had the first preliminary commercial application. Great strides have been made in the establishment of corpus, the extraction of speech emotion features, and the emotion recognition models as research in the field of emotion recognition has continued to deepen. In terms of establishing a corpus, the Technical University of Berlin recorded the German database EMO-DB in 2005, which is widely used in emotion research [14]. In 2010, [15] proposed a dimensional SEMAINE database for human-computer interaction and used the annotation tool FEELTRACE to annotate it on five emotional dimensions. The China Institute of Automation then established the China Natural Type Multimodal Database (CHEAVD) [16]. The creation of a large and diverse database has laid a solid foundation for future research on speech emotion recognition. In order to extract emotional features, common prosodic features such as time length [17], fundamental frequency [18], and energy [19] are used, as well as Linear Prediction Cepstral Coefficients [20], Mel Frequency Cepstral Coefficients [21], Log Frequency Power Coefficients [22], and other spectral features. The EMO-DB voice database is used in [23], and the spectrogram of the dataset is extracted as the input dataset, which is then fed into a convolutional neural network to automatically learn high-level emotional components, with a final recognition rate of more than 70%. There are now primarily machine learning-based emotion recognition models [24, 25] and deep learning-based emotion recognition models [26–28].

This paper considers that in real production and life, the use of voice-based human-computer interaction is the most common, and this method will also be the future development trend. Therefore, this paper mainly focuses on emotion recognition research on speech data. LSTM in deep learning model has unique advantages in speech recognition. Since the original LSTM only considers the input of the previous moment, it will lose a lot of information for the entire context scene. As a result, the ILSTM model is proposed in this study, which improves on the classic LSTM model. Because the model believes that the input at the current instant is related not just to the previous moment, but also to all past moments, it extracts all of the features from the speech segment. The ILSTM model additionally includes an attention mechanism in order to select the aspects that can best communicate emotion among multiple features. The suggested ILSTM model's effectiveness and superiority are demonstrated by experimental results on public datasets.

2. Knowledge Related to Speech Emotion Recognition

2.1. Emotional Categories. There are many kinds of human emotions, and researchers divide them into discrete and continuous emotion classification descriptions according to different basis. Among them, the discrete emotion

Table 1: Discrete emotion partition details.

Author	Sentiment classification type
Plutchik	Angry, happy, sad, disgusted, fearful, approving, anticipating, surprised
Ekman, Friesen, and Ellsworth	Angry, happy, sad, disgusted, fearful, surprised
Frijda	Happy, concerned, surprised, expectant, sad
Gray	Angry, happy, fearful, anxious
Arnold	Angry, disgusted, brave, frustrated, anticipating, disappointed, fearful, hopeful, loving, sad
James	Angry, scared, sad, loving
Mower	Joy, pain
McDougall	Angry, happy, disgusted, afraid, conquered, guilty, tender, surprised
Panksepp	Anger, expectation, fear, panic
Izard	Anger, contempt, disgust, grief, fear, guilt, concern, shame, surprise
Tomkins	Angry, contemptuous, disgusted, sad, scared, happy, concerned, ashamed, surprised
Waston	Angry, scared, loving
Weiner and Graham	Happy, sad
Oatley and Jolnson-Laird	Angry, happy, sad, disgusted, anxious

Table 2: Details of continuous emotion division.

Category	Details
Two-dimensional	Arousal is used to describe the intensity of emotions, such as anger and joy. Valence space is used to describe the degree of positive and negative emotions. It is used to distinguish between angry and happy emotions.
Three-dimensional	Pleasure is primarily used to assess whether an emotion is in a positive or negative state. Arousal is mainly used to describe the degree of emotional strength. Dominance is used to describe a situation in which an individual is in domination or being dominated.

classification is shown in Table 1. In the realm of emotion recognition, the six basic emotions proposed by Ekman et al. in the table, namely, anger, disgust, fear, happiness, sorrow, and surprise, are the most extensively utilized.

In contrast to discrete emotion classification, some scholars believe that emotion is continuous and gradually changing in space. Any emotion state can be mapped to a point in space, and the discrete emotion description model cannot fully cover the emotion in real life. Continuous emotional description uses continuous coordinate points in space to describe emotional states. The size of the coordinate value represents the intensity of emotion in each dimension. The spatial distance of coordinate points in dimensional space indicates the similarity and difference between emotions. Therefore, the purpose of emotion classification is to find the correspondence between coordinate points and emotional states in the dimensional space. The emotion categories are divided into four quadrants in the two-dimensional Cartesian coordinate system. The closer the coordinate system to the origin, the less intense the emotion, and vice versa. The continuous sentiment classification is shown in Table 2.

2.2. Speech Emotion Recognition Dataset. At present, in the field of speech emotion recognition research, there are many kinds of corpora available for research, such as EMO-DB German database, DES Danish database [29], CASIA (the Institute of Automation of the Chinese Academy of Sciences) database [30], and IEMOCAP English database [23]. However, due to the influence of different geographical locations, pronunciation habits, and direct differences between cultures and languages, different corpora have certain particularities. There are no particularly hard boundaries between sentiment labels in different databases. The definitions of tags are not uniform, so there is no general speech emotion database for all researchers to refer to. Table 3 mainly lists common speech databases from the language, size, type, emotional label, etc. of the corpus.

2.3. Speech-Based Emotion Recognition Process. Speech emotion recognition is mainly divided into the following links: the establishment of emotion database, speech signal preprocessing, feature extraction, model training, and model testing. The identification process is shown in Figure 1. The corpus is the data source for model training and testing, where the test samples can use data from the corpus or real-life voices. Preprocessing refers to converting the collected speech signal into a digital signal that can be recognized by the computer through analog and digital processing technology; applying hardware or software technology; and performing operations such as preemphasis, framing, windowing, and denoising. Feature extraction refers to extracting the acoustic features that can represent emotion through feature extraction tools such as openSMILE, openEAR open source tools, or principal component analysis and other feature extraction algorithms from the preprocessed data. The extracted features are required to be able to better represent the inherent characteristics of the original speech. Model training refers to the process of building a speech emotion recognition model. The training of general models is done using machine learning or deep learning algorithms. Model testing refers to calling the

Table 3: Common speech databases.

Database	Language type	Size (emotion ∗ number of people ∗ text)	Data-induced type	Emotion
EMO-DB	German	700 (7 ∗ 10 ∗ 10)	Imitative	Anger, joy, sadness, fear, disgust, boredom, neutrality
CASIA	Mandarin	1200 (6 ∗ 4 ∗ 50)	Imitative	Angry, happy, sad, surprised, fearful, neutral
LDC	English	1050 (15 ∗ 7 ∗ 10)	Imitative	Neutrality, panic, anxiety, longing, sadness, joy, hobby, boredom, shame, pride, contempt, etc.
DES	Danish	5 ∗ 4 ∗ (2 words, 9 sentences, and 2 articles)	Imitative	Angry, happy, sad, surprise, neutral
IEMOCAP	English	9000 (10 speakers, 7 emotions)	Imitative	Happy, angry, sad, neutral, surprised, disgusted, excited
KISMIT	English	1002 (3 women, 5 emotions)	Natural	Agree, attentive, forbidden, comfortable, neutral
SAVEES	English	480 (7 ∗ 4 ∗ 15)	Imitative	Angry, happy, sad, fearful, disgusted, surprised, neutral
MPEG-4	English	2440 (35 speakers, 7 emotion categories)	Imitative	Angry, happy, sad, fearful, disgusted, surprised, neutral
eNTERFACE	English	1287 (34 males, 8 females, 6 types of emotions)	Guided	Anger, joy, sadness, fear, disgust, surprise
Belfast	German	(20 males and 20 females, 5 types of emotions)	Imitative	Angry, happy, sad, fearful, neutral
FAU-AIBO	German	48401 (51 children, 11 emotion categories)	Natural	Angry, happy, bored, stressed, helpless, sarcastic, delighted, reprimanded, rested, surprised, sensitive
ACCorpus_SR	Mandarin	50000 (25 males and 25 females, 5 types of emotions)	Imitative	Angry, happy, sad, fearful, neutral
Pereiras	English	80 (5 ∗ 2 ∗ 8)	Imitative	High anger, low anger, happiness, sadness, neutrality
CLDC	Mandarin	1200 (4 speakers, 6 types of emotions)	Imitative	Angry, happy, sad, surprised, fearful, neutral
KES	Korean	5400 (10 speakers, 4 types of emotions)	Imitative	Angry, happy, sad, neutral
CHEAVD	Mandarin	7030 (238 speakers, 8 emotions)	Natural	Angry, happy, sad, surprised, worried, anxious, disgusted, neutral
MSP-IMPROV	English	7798 (6 males and 6 females, 4 types of emotions)	Imitative	Angry, happy, sad, neutral
VAM	German	947 (47 speakers)	Natural	Activation-valence-control dimension
SEMAINE	English	95 speech segments (20 speakers)	Natural	Activation-valence-arousal-expectation-intensity dimension

training model, inputting the test set into the trained model, using the classification result to calculate the evaluation index, and then judging the performance of the model according to the evaluation index.

3. ILSTM Model

The LSTM network introducing the attention mechanism can rely on this mechanism to learn the weight of each step and express it as a weighted combination. This multitask learning can better learn features in sentences. The LSTM network structure that introduces the attention mechanism is shown in Figure 2.

The structure is divided into stage 1 and stage 2. Stage 2 is sentiment classification. Stage 1 shares all tasks and handles the input and feature representation of the classification, and its top is a weighted pooling layer, which is calculated as (1) and (2). In stage 1, there is a fully connected layer consisting of 256 ReLU nodes and a bidirectional LSTM layer consisting of 128 nodes, followed by a weighted pooling layer. In stage 2, each task has a hidden layer that contains 256 ReLU neurons and a Softmax layer.

$$Weighted\,pooling = \sum_{T=t1}^{tn} A_T \times h_T, \tag{1}$$

$$A_T = \frac{\exp(W \cdot h_T)}{\sum_{T=t1}^{tn} \exp(W \cdot h_T)}, \tag{2}$$

where h_T is the output of the LSTM at T, A_T is the scalar of the corresponding weight at T, and the calculation process is as in (2). W is the learning parameter, and $\exp(W \cdot h_T)$ is the energy at T. If the energy of the frame at time T is high, its weight will increase, and the attention will be higher. Otherwise, the attention will be lower.

In traditional LSTM, the mechanism of data transmission is mainly that the data from the bottom layer and the previous moment is continuously output to the next layer. As shown in (3), the gate mechanism controls the flow of information through point multiplication, and the memory cell updates information. f_t and i_t are the forget gate and input gate outputs at t, respectively, and C_t is the new candidate unit value calculated as (4). *tanh* represents the activation function, W_c represents the learned weight

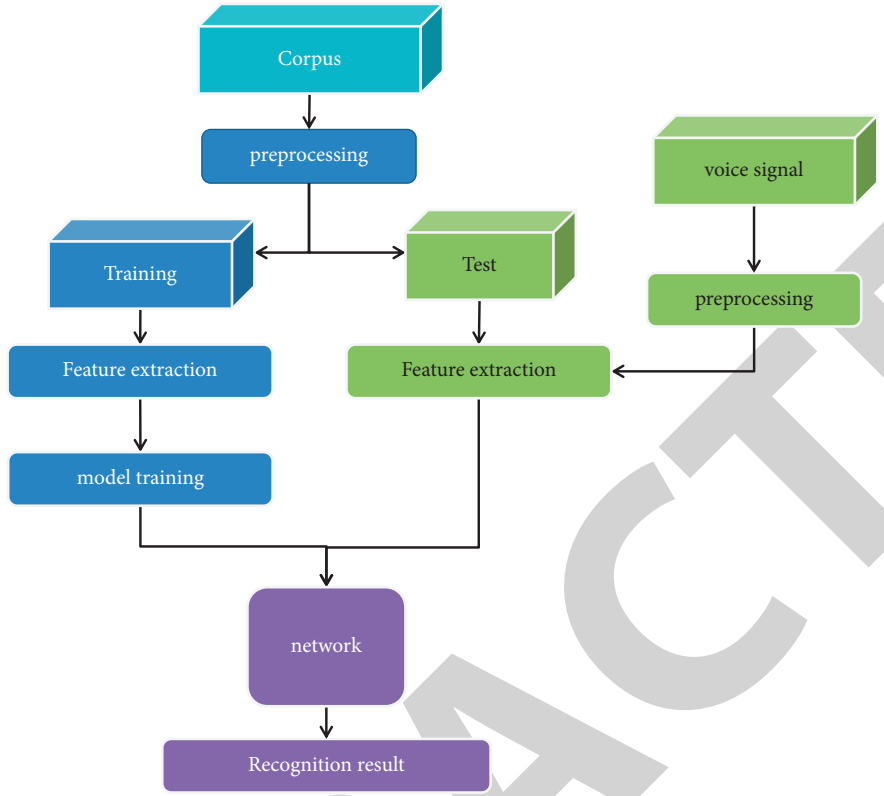

Figure 1: Speech emotion recognition process.

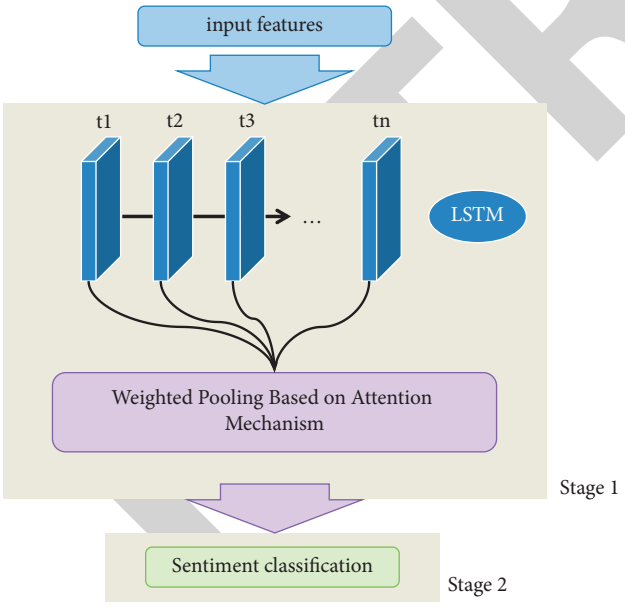

Figure 2: LSTM network structure based on attention mechanism.

set, and b_c represents the bias; $[h_{t-1}, x_t]$ represents the concatenation of the previous time step (h value) and the bottom layer (x value), and the h value at t is calculated. For example, in (5), O_t is the output gate, which computes C_t based on h_{t-1} and C_{t-1}.

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t, \tag{3}$$

$$\widetilde{C}_t = \tanh\left(W_C \cdot [h_{t-1}, x_t]\right) + b_c, \tag{4}$$

$$h_t = o_t \odot \tanh\left(C_t\right). \tag{5}$$

Equations (3) and (4) should be changed to (6) and (7), where C is the weighted sum of selected states and T is the set of selected time steps. Equation (9) computes scalar representing the weight corresponding to the time step. Equation (10) is used to calculate the implicit value at time t, which is the same as (5), but this time the unit value is C'. h' is calculated through (11) and (12). W is the learned shared parameter in (9) and (12), and C' and h' contain all of the states and implicit values in the set T.

The improved LSTM has a more flexible time-dependent modeling ability, similar to the human learning function, which can recall historical information and improve learning efficiency. In this paper, the attention mechanism is introduced into the above LSTM network to obtain the ILSTM network. The ILSTM structure is shown in Figure 3. The difference between Figures 3 and 2 is that the LSTM network in Figure 2 is replaced with the LSTM network structure shown in Figure 4. Its calculation process is as follows:

$$C_t = f_t \odot C'_{t-1} + i_t \odot \widetilde{C}_t, \tag{6}$$

$$\widetilde{C}_t = \tanh\left(W_c \cdot \left[h'_{t-1}, x_t\right]\right) + b_c, \tag{7}$$

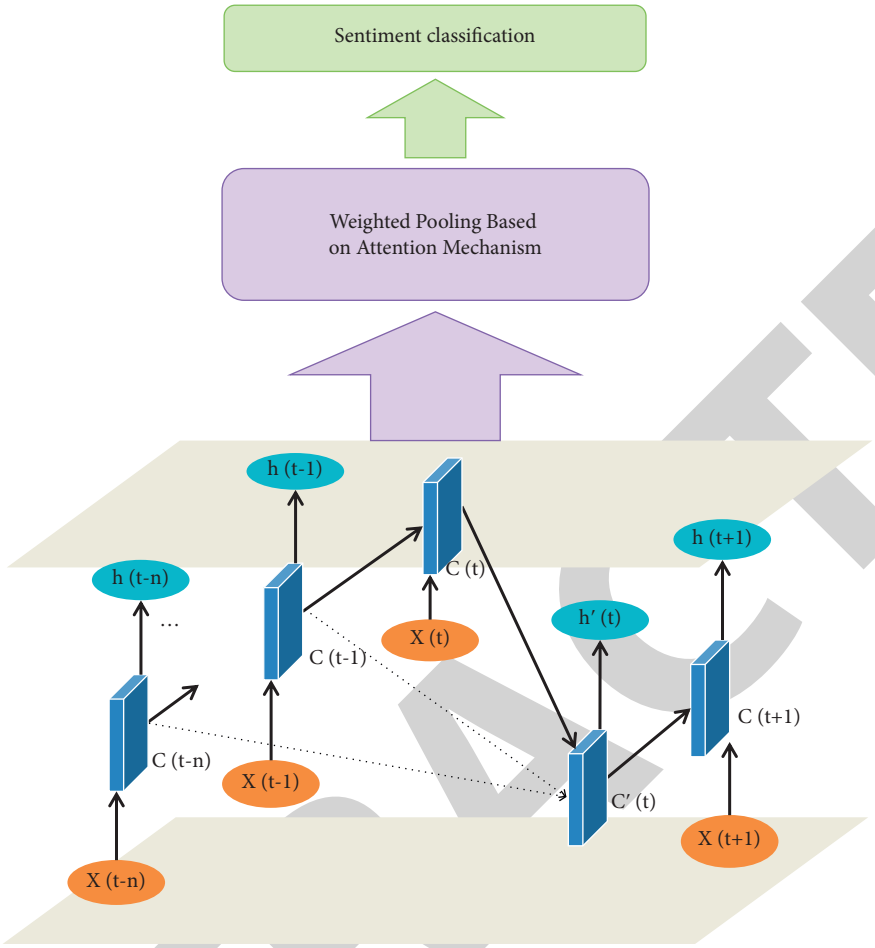

Figure 3: ILSTM network structure.

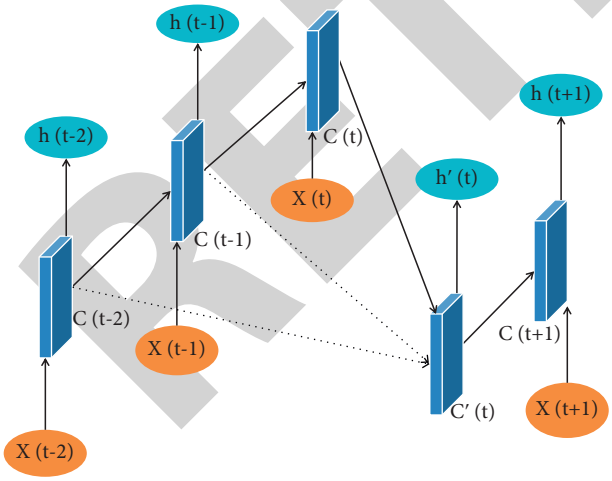

Figure 4: Improved LSTM structure diagram.

$$C' = \sum_{T} W_{C_T} \times C_T, \tag{8}$$

$$W_{C_T} = \frac{\exp\left(W \cdot C_T\right)}{\sum_{T} \exp\left(W \cdot C_T\right)}, \tag{9}$$

$$h_t = o_t \odot \tanh\left(C_t'\right), \tag{10}$$

$$h' = \sum_{T} W_{hT} \times h_T, \tag{11}$$

$$W_{h_T} = \frac{\exp\left(W \cdot h_T\right)}{\sum_{T} W \cdot h_T}. \tag{12}$$

4. Experimental Analysis

4.1. Experimental Dataset. In order to verify the recognition rate of the model in this paper for speech data in different languages, this paper selects the English dataset Belfast, the German dataset EMO-DB, and the Chinese dataset CASIA. The detailed introduction of each dataset is shown in Table 4.

4.2. Experimental Parameter Settings. This paper mainly uses dropout technology to prevent overfitting during training. LSTM layers all use dropout. It mainly detects the units by ignoring half of the features in each training batch. By reducing the interaction of the feature detection unit, the activation value of some neurons stops working with a certain probability. This makes the model more

Table 4: Dataset introduction.

Dataset	Details
Belfast	Collection unit: Queen University; volunteers: 40, 20 males and 20 females; emotion categories: 5, anger, sadness, joy, fear, neutrality
EMO-DB	Collection unit: Technical University of Berlin; volunteers: 10, 5 males and 5 females; emotion categories: 7, neutral, angry, fearful, happy, sad, disgusted, bored
CASIA	Collection unit: Institute of Automation, Chinese Academy of Sciences; volunteers: 4, 2 males and 2 females; emotional categories: 5, happy, sad, angry, frightened, neutral

Table 5: Model parameter settings.

Parameter	Parameter value
Learning rate	0.1, 0.05, 0.01, 0.005, 0.01
Batchsize	16, 32, 64, 128
Dropout	0–0.03, 0.04–0.5, 0.06–0.08, 0.09–0.11, 0.12–0.15, 0.15–0.2
Iterations	50, 100, 150, 200, 300, 500, 1000
K_folds	5, 10, 15
Optimizer	Adam, RMSprop, SGD

generalizable and does not depend on some local features. The parameters that need to be determined in the ILSTM model in this paper include batch size (Batchsize), iteration period (Iterations), training termination condition (Patience), and cross-validation times (K_folds). The values of these parameters are shown in Table 5.

The obtained model performance varies greatly depending on the parameter settings. The accuracy rate is the evaluation index used in this paper to determine the parameters of the optimal model. The accuracy rate refers to the positive sentiment data identified as positive plus the negative sentiment data identified as negative divided by the total number of samples. As our most commonly used indicator, the accuracy rate cannot reasonably reflect the classification ability of the model when the sample is unbalanced. For example, the test dataset has 90% positive samples and 10% negative samples. Assuming that the classification results of the model are all positive samples, the accuracy rate is 90%. However, the model has no ability to identify negative samples. At this time, the model's classification ability cannot be reflected by its high accuracy rate. The following is the formula for calculating accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}. \tag{13}$$

Precision indicates the number of actual positive samples in the samples classified as positive. This indicator mainly reflects the accuracy of the model. Its calculation formula is as follows:

$$\text{Pr}ecision = \frac{TP}{TP + FP}. \tag{14}$$

Recall is for data samples. In the data sample, the probability that the positive sample is correctly classified. Similar to how many questions a candidate answers on a test paper. It reflects the comprehensiveness of a model; that is, the model can find all the correctly answered questions. The calculation formula of recall is as follows:

$$\text{Re}call = \frac{TP}{TP + FN}. \tag{15}$$

Precision and recall are a pair of contradictory measures. Generally speaking, when precision is high, the recall value tends to be low. When the precision value is low, the recall value tends to be high. When the classification confidence is high, the precision is high; when the classification confidence is low, the recall is high. In order to be able to comprehensively consider these two indicators, F1 is proposed. The core idea of F1 is that while improving precision and recall as much as possible, we also want the difference between the two to be as small as possible. The formula for calculating F1 is as follows:

$$F1 = 2 \times \frac{\text{Pr}ecision * \text{Re}call}{\text{Pr}ecision + \text{Re}call}. \tag{16}$$

4.3. Analysis of Experimental Results

4.3.1. Parameter Determination Experiment. Experiments on the EMO-DB dataset were carried out in order to determine the optimal parameters of the model. Figure 5 depicts how the model's emotion recognition accuracy varies with parameter values. The effect of changing the learning rate on the recognition rate is depicted in Figure 5(a). The figure shows that as the learning rate increases, the accuracy of emotion recognition decreases gradually. The emotion recognition rate is highest when the learning rate is 0.001. Figure 5(b) shows the effect of the change of dropout value on the recognition rate. It can be seen from the figure that when dropout is 0.1, the recognition rate is the highest. Figure 5(c) shows the effect of the value of Batchsize on the recognition rate. It can be seen from the figure that when the Batchsize is 32, the recognition rate is the highest. Figure 5(d) shows the effect of Iterations on the recognition rate. It can be seen from the figure that when Iterations takes 300, the recognition rate approaches the highest value. As the number of times increases, the recognition rate does not increase significantly. Considering the factors of the recognition rate and the shortest possible time, Iterations is selected as 300.

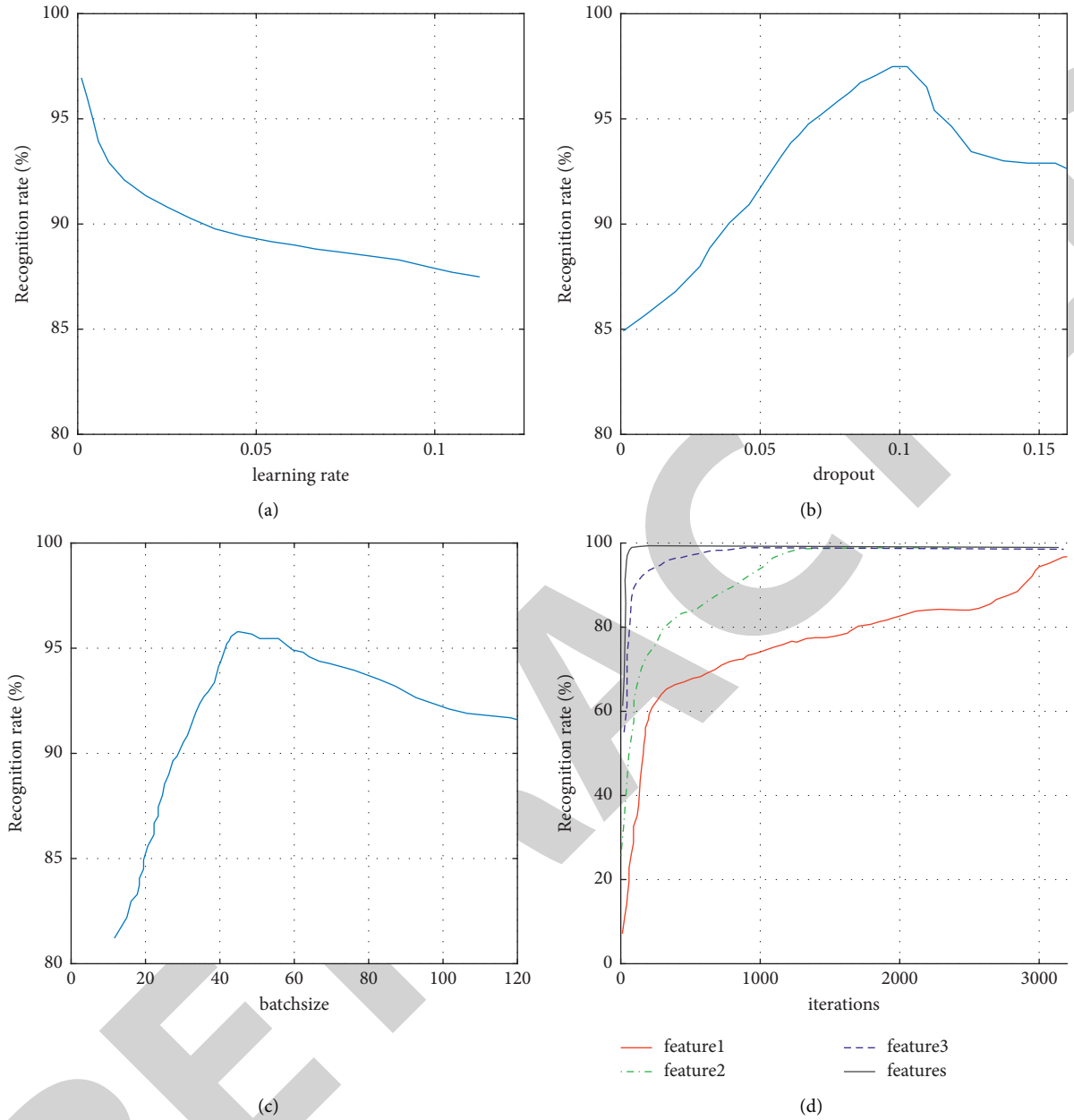

Figure 5: The recognition rate of different parameters.

Figure 6 shows the accuracy of the model under different cross-validation times and optimizers. Figure 6(a) shows that when K_folds is 10, the accuracy is the highest. Figure 6(b) shows that when the optimizer is Adam, the obtained accuracy is the best.

4.3.2. Model Classification Performance Experiment. In order to analyze the classification performance of the model in this paper on emotional data, the selected comparison models mainly include CNN [31], LSTM [32], BiLSTM [33], CNN-LSTM [34], and DCNN-LSTM [35]. The experimental steps are as follows: run the model 10 times, and take the average. The recognition accuracy, precision, recall, and F1 data of each model on the three datasets are shown in Tables 6–8, and 9, respectively.

From the experimental data shown in Table 6, the following experimental conclusions can be drawn:

(1) For the Belfast dataset, except for CNN, the classification accuracy of the models is above 0.6. Several other models are evolutionary models based on the LSTM model. This demonstrates that the LSTM model is better suited to the Belfast dataset. The ILSTM model used in this paper has the best classification effect among the evolutionary models of multiple LSTMs. This demonstrates that the ILSTM model in this paper successfully extracts all of the features in the speech segment by taking into account

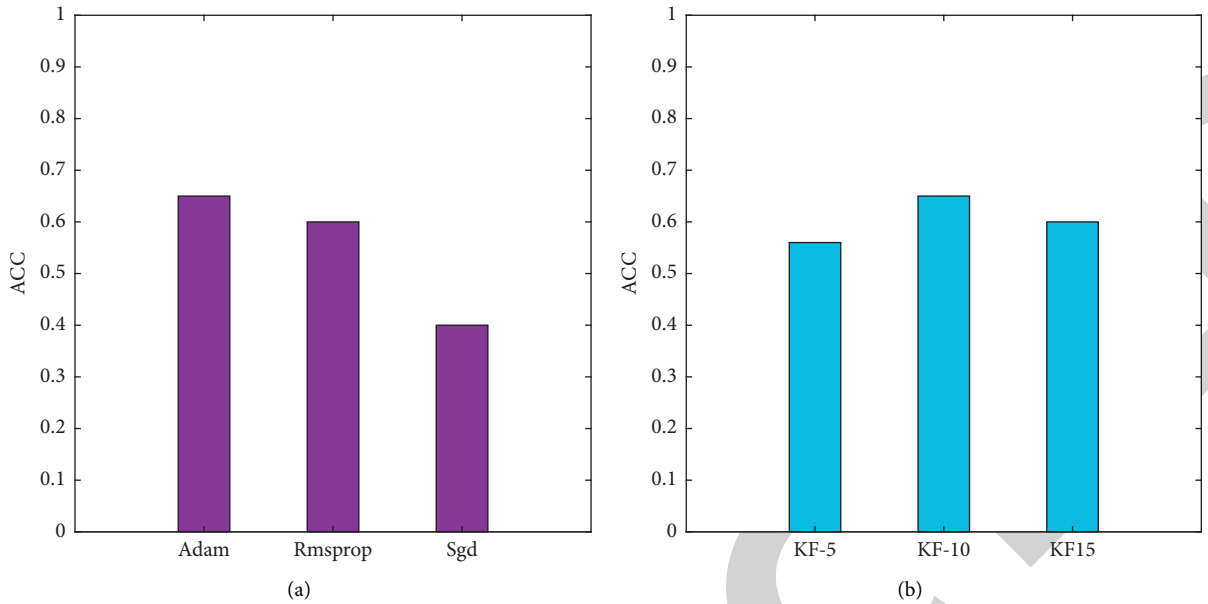

Figure 6: Accuracy under different optimizers and K_fold settings.

Table 6: Recognition accuracy of different models on each dataset.

Model/dataset	Belfast	EMO-DB	CASIA
CNN	0.5821	0.6017	0.5582
LSTM	0.6138	0.5958	0.5404
BiLSTM	0.6312	0.6109	0.5698
CNN-LSTM	0.6284	0.6083	0.5658
DCNN-LSTM	0.6189	0.6226	0.5710
ILSTM	0.6535	0.6490	0.5996

Table 7: Precision of different models on each dataset.

Model/dataset	Belfast	EMO-DB	CASIA
CNN	0.5954	0.6124	0.5668
LSTM	0.6076	0.5896	0.5573
BiLSTM	0.6147	0.6287	0.5702
CNN-LSTM	0.6084	0.6116	0.5694
DCNN-LSTM	0.6291	0.6302	0.5840
ILSTM	0.6387	0.6398	0.5855

the fact that the input at the current moment is related to all previous moments, not just the previous moment. In addition, an attention mechanism is introduced in order to select the feature that can express emotion the most among the many features. These operations enable the model to extract more rich and valuable features for effective classification.

(2) For the EMO-DB dataset, the classification accuracy of CNN is better than that of LSTM. However, the difference between the two is not big. Among several other LSTM-based evolution models, the ILSTM model in this paper still has the highest classification accuracy. However, for this dataset, the advantages of our model are not so obvious.

(3) For the CASIA dataset, the best classification performance is still that of the model in this paper. Compared with CNN, LSTM, BiLSTM, CNN-LSTM, and DCNN-LSTM, the model in this paper is improved by 7.4%, 11.0%, 5.2%, 6.0%, and 5.0% respectively. From this data, it can be seen that the ILSTM model has the highest improvement on the basis of the original LSTM model.

From the experimental data shown in Table 7, the following experimental conclusions can be drawn:

(1) For the Belfast dataset, compared with the data in Table 6, for the CNN model, the accuracy rate is higher than the accuracy rate. This shows that the precision is higher than accuracy. Among several LSTM-based models, ISLTM has the highest accuracy, followed by DCNN-LSTM, and the LSTM is the worst. This shows that different improved models do have to overcome some shortcomings of the traditional LSTM model itself.

(2) For the EMO-DB dataset, the accuracy of ISLTM is improved by 4.5, 8.5, 1.8, 4.6, and 1.5 compared to CNN, LSTM, BiLSTM, CNN-LSTM, and DCNN-LSTM, respectively. Among them, for LSTM, the improvement of the ILSTM model has the largest magnitude. For the DCNN-LSTM model, the accuracy of the algorithm in this paper is improved by a small margin, and the advantages of the model are not obvious.

(3) For the CASIA dataset, the performance gap between our model and other models becomes smaller. This shows that the advantage of this model considering the features between contexts on this dataset is not obvious. The DCNN-LSTM model has almost the same accuracy as the model in this paper.

Table 8: Recall rates of different models on each dataset.

Model/dataset	Belfast	EMO-DB	CASIA
CNN	0.5859	0.6212	0.5587
LSTM	0.6002	0.5964	0.5632
BiLSTM	0.6039	0.6276	0.5687
CNN-LSTM	0.6135	0.6323	0.5754
DCNN-LSTM	0.6198	0.6390	0.5875
ILSTM	0.6444	0.6559	0.6060

Table 9: F1 of different models on each dataset.

Model/dataset	Belfast	EMO-DB	CASIA
CNN	0.5906	0.6168	0.5627
LSTM	0.6039	0.5930	0.5602
BiLSTM	0.6093	0.6281	0.5694
CNN-LSTM	0.6109	0.6218	0.5724
DCNN-LSTM	0.6244	0.6346	0.5857
ILSTM	0.6415	0.6477	0.5956

The recall rate can reflect the comprehensiveness of the model. From the recall rate data shown in Table 8, the following experimental conclusions can be drawn: For the Belfast dataset, the recall rate of the ILSTM model in this paper is compared to CNN, LSTM, BiLSTM, CNN-LSTM, and DCNN. ILSTM is improved by 9.2, 6.6, 5.9, 4.3, and 3.2 respectively. For the EMO-DB dataset, the recall rate of the ILSTM model in this paper is improved by 5.6, 10.0, 4.5, 3.7, and 2.6 compared to CNN, LSTM, BiLSTM, CNN-LSTM, and DCNN-LSTM, respectively. For the CASIA dataset, the recall rate of the ILSTM model in this paper is 8.5, 7.6, 6.6, 5.3, and 3.1 higher than that of CNN, LSTM, BiLSTM, CNN-LSTM, and DCNN-LSTM, respectively. No matter which dataset is used, the recall rate of the model in this paper is at least 2.6 higher than that of any model, which fully proves the comprehensiveness of the model in this paper.

From the experimental data shown in Table 9, the following experimental conclusions can be drawn: For the Belfast dataset, compared with CNN, LSTM, BiLSTM, CNN-LSTM, and DCNN-LSTM, the F1 index of the ILSTM model used is improved by 8.6, 6.2, 5.3, 5.0, and 2.7, respectively. For the EMO-DB dataset, the recall rate of the ILSTM model in this paper is improved by 5.0, 9.2, 3.1, 4.2, and 2.1 compared to CNN, LSTM, BiLSTM, CNN-LSTM, and DCNN-LSTM, respectively. For the CASIA dataset, the recall rate of the ILSTM model in this paper is improved by 5.8, 6.3, 4.6, 4.1, and 1.7 compared with CNN, LSTM, BiLSTM, CNN-LSTM, and DCNN-LSTM, respectively. Overall, the performance of the ILSTM model used in this paper is better than that of other comparative models.

5. Conclusion

Efficient and accurate emotion recognition plays a very important role in the development of human-computer interaction and other fields. Considering that speech is the main way of human-computer interaction, this paper mainly studies emotion recognition from speech data. There are many studies on the application of deep learning models to emotion recognition. In this paper, LSTM is selected as the basic model, and two improvements are made. First, the traditional LSTM algorithm only considers that the input of the previous moment is abandoned. The ILSTM model considers that the input at the current moment is related to not only the previous moment, but also to all previous moments. Therefore, all the features in the speech segment need to be extracted. This way of considering the entire context scene will not lose a lot of information. In addition, in order to select the features that can best express emotion among many features, the model also introduces an attention mechanism. The improved LSTM is tested on three different language speech datasets. The experimental results show that the parameters in the network structure have a great impact on the performance of the emotion recognition system. Selecting an appropriate parameter set can not only improve the performance of the network model, but also greatly reduce the training time of the model. However, although the ISLTM in this paper can improve the classification performance, it also adds more parameters. For different datasets, the parameters will also be different. The determination of parameters is time-consuming. This is also where further optimization is required in the subsequent article.

Data Availability

The labeled dataset used to support the findings of this study is available from the author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was supported by the Anyang Institute of Technology.

References

[1] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[2] M. Ei Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion R ecognition: features, classification schem es and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[3] K. Belhouchette, "Facial recognition to identify emotions: an application of deep learning," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 72, pp. 496–504, 2021.

[4] M. A. Hasnul, N. Azlina, S. Aziz, M. Mohana, and A. A. Aziz, "Electrocardiogram-based emotion recognition systems and their applications in healthcare-A review," *Sensors*, vol. 21, no. 15, p. 5015, 2021.

[5] V. Colonnello, K. Mattarozzi, and P. M. Russo, "Emotion recognition in medical students: effects of facial appearance and care schema activation," *Medical Education*, vol. 53, no. 2, pp. 195–205, 2019.

[6] J. Liu, X. Wu, and X. Wu, "Prototype of educational affective arousal evaluation system based on facial and speech emotion recognition," *International Journal of Information and Education Technology*, vol. 9, no. 9, pp. 645–651, 2019.

[7] J. Oliveira and I. Praca, "On the usage of pre-trained speech recognition deep layers to detect emotions," *IEEE Access*, vol. 9, pp. 9699–9705, 2021.

[8] E. Ostrosi, J.-B. Bluntzer, Z. Zhang, and J. Stjepandić, "Car style-holon recognition in computer-aided design," *Journal of Computational Design and Engineering*, vol. 6, no. 4, pp. 719–738, 2019.

[9] C. E. Williams and K. N. Stevens, "Emotions and speech: some acoustical correlates," *Journal of the Acoustical Society of America*, vol. 52, no. 4B, pp. 1238–1250, 1972.

[10] R. Van Bezooijen, S. A. Otto, and T. A. Heenan, "Recognition of vocal expressions of emotion," *Journal of Cross-Cultural Psychology*, vol. 14, no. 4, pp. 387–406, 1983.

[11] T. Goldbeck, F. Tolkmitt, and K. R. Scherer, *Experimental studies on vocal affect communication*, Psychology press, East Sussex, UK, 1988.

[12] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proceedings of the fourth international conference on spoken language processing. ICSLP'96*, vol. 3, pp. 1970–1973, PA, USA, October 1996.

[13] T. Moriyama and S. Ozawa, "Emotion recognition and synthesis system on speech[C]/Proceedings IEEE international conference on multimedia computing and systems," *IEEE*, vol. 1, pp. 840–844, 1999.

[14] A. Milton, S. Sharmy Roy, and S. Tamil Selvi, "SVM scheme for speech emotion recognition using MFCC feature," *International Journal of Computer Application*, vol. 69, no. 9, pp. 34–39, 2013.

[15] G. Mckeown, M. F. Valstar, and R. Cowie, "The SEMAINE corpus of emotionally coloured character interactions," in *Proceedings of the 2010 EEE International Conference on Multimedia and Expo*, pp. 1079–1084, Singapore, July 2010.

[16] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "CHEAVD: a Chinese natural emotional audio-visual database," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 6, pp. 913–924, 2017.

[17] M. Aghajani, H. Ben Abdessalem, C. Frasson, and C. Frasson, "Voice emotion recognition in real time applications," *Intelligent Tutoring Systems*, vol. 12677, pp. 490–496, 2021.

[18] C. Galand, C. Couturier, G. Platel, and R. Vermot-Gauchy, "Voice-excited predictive coder (VEPC) implementation on a high-performance signal processor," *IBM Journal of Research and Development*, vol. 29, no. 2, pp. 147–157, 1985.

[19] N. Hajj, M. Filo, M. Awad, A. Puglisi, and A. Prados, "Automated composer recognition for multi-voice piano compositions using rhythmic features, n-grams and modified cortical algorithms," *Complex & Intelligent Systems*, vol. 4, no. 1, pp. 55–65, 2018.

[20] H. Kondhalkar and P. Mukherji, "A novel algorithm for speech recognition using tonal frequency cepstral coefficients based on human cochlea frequency map," *Journal of Engineering Science & Technology*, vol. 14, no. 2, pp. 726–746, 2019.

[21] M. Cheffena, "Fall detection using smartphone audio features," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 4, pp. 1073–1080, 2016.

[22] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[23] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[24] J. A. Domínguez-Jiménez, K. C. Campo-Landines, J. C. Martínez-Santos, E. J. Delahoz, and S. H. Contreras-Ortiz, "A machine learning model for emotion recognition from physiological signals," *Biomedical Signal Processing and Control*, vol. 55, Article ID 101646, 2020.

[25] S. Kranti, "Kamble; joydeep Sengupta,Ensemble machine learning-based affective computing for emotion recognition using dual-decomposed EEG signals," *IEEE Sensors Journal*, vol. 22, no. 3, pp. 2496–2507, 2022.

[26] S. Mishra, B. Joshi, R. Paudyal, D. Chaulagain, and S. Shakya, "Deep residual learning for facial emotion recognition," *Mobile Computing and Sustainable Informatics*, vol. 68, pp. 301–313, 2022.

[27] A. Khattak, M. Z. Asghar, M. Ali, and U. Batool, "An efficient deep learning technique for facial emotion recognition," *Multimedia Tools and Applications*, vol. 81, no. 2, pp. 1649–1683, 2022.

[28] J. Zhao, Z. W. Zhang, J. Qiu, L. Shi, Z. Kuang, and W. Jing, "GTSception: a deep learning eeg emotion recognition model based on fusion of global, time domain and frequency domain feature extraction," *Research Square*, 2021.

[29] I. S. Engberg, A. V. Hansen, and O. Andersen, "Design, recording and verification of a danish emotional speech database," in *Proceedings of the 5th European Conference on Speech Communication and Technology*, pp. 1695–1698, Rhodes, Greece, September 1997.

[30] K. Wang, N. An, and B. N. Li, "Speech emotion rec ognition using fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, 2017.

[31] W.-D. Dang, D.-M. Lv, R.-M. Li, Z.-Y. Rui, C. Ma, and Z.-K. Gao, "Multilayer network-based CNN model for emotion recognition," *International Journal of Bifurcation and Chaos*, vol. 32, no. 1, pp. 1–10, 2022.

[32] P. Tahghighi, A. Koochari, and M. Jalali, "Deformable convolutional LSTM for human body emotion recognition," *Pattern Recognition. ICPR International Workshops and Challenges*, vol. 12663, pp. 741–747, 2021.

[33] M. Sajjad and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.

[34] I. E. Livieris, E. Pintelas, and P. Pinteras, "A CNN-LSTM model for gold price time-series forecasting," *Neural Computing & Applications*, vol. 32, pp. 1–10, 2020.

[35] J. Kim and R. A. Saurous, "Emotion recognition from human speech using temporal information and deep learing," in *Proceedings of the 2018 Conference of the International Speech Communication Association*, pp. 937–940, Hyderabad, India, September 2018.