# Effect on speech emotion classification of a feature selection approach using a convolutional neural network

**Ammar Amjad** [1] , **Lal Khan** [1] , **Hsien-Tsung Chang** Corresp. [1, 2, 3]

[1] Department of Computer Science and Information Engineering, Chang Gung University, Taoyuan, Taiwan

[2] Department of Physical Medicine and Rehabilitation, Chang Gung Memorial Hospital, Taoyuan, Taiwan

[3] Artificial Intelligence Research Center, Chang Gung University, Taoyuan, Taiwan

Corresponding Author: Hsien-Tsung Chang
Email address: smallpig@widelab.org

Speech emotion recognition (SER) is a challenging issue because it is not clear which features are effective for classification. Emotionally related features are always extracted from speech signals for emotional classification. Handcrafted features are mainly used for emotional identification from audio signals. However, these features are not sufficient to correctly identify the emotional state of the speaker. The advantages of a deep convolutional neural network (DCNN) are investigated in the proposed work. A pretrained framework is used to extract the features from speech emotion databases. In this work, we adopt the feature selection (FS) approach to find the discriminative and most important features for SER. Many algorithms are used for the emotion classification problem. We use the random forest (RF), decision tree (DT), support vector machine (SVM), multilayer perceptron classifier (MLP), and k-nearest neighbors (KNN) to classify seven emotions. All experiments are performed by utilizing four different publicly accessible databases. Our method obtains accuracies of 92.02%, 88.77%, 93.61%, and 77.23% for Emo-DB, SAVEE, RAVDESS, and IEMOCAP, respectively, for speaker-dependent (SD) recognition with the feature selection method. Furthermore, compared to current handcrafted feature-based SER methods, the proposed method shows the best results for speaker-independent SER. For EMO-DB, all classifiers attain an accuracy of more than 80\% with or without the feature selection technique.

# Effect on Speech Emotion Classification of a Feature Selection Approach Using a Convolutional Neural Network

**Ammar Amjad**[1], **Lal Khan**[1], **and Hsien-Tsung Chang**[1,2,3]

[1]Department of Computer Science and Information Engineering, College of Engineering, Chang Gung University, Taoyuan, Taiwan
[2]Department of Physical Medicine and Rehabilitation, Chang Gung Memorial Hospital, Taoyuan, Taiwan
[3]Artificial Intelligence Research Center, Chang Gung University, Taiwan

Corresponding author:
Hsien-Tsung Chang[1,2,3]

Email address: smallpig@widelab.org

## ABSTRACT

Speech emotion recognition (SER) is a challenging issue because it is not clear which features are effective for classification. Emotionally related features are always extracted from speech signals for emotional classification. Handcrafted features are mainly used for emotional identification from audio signals. However, these features are not sufficient to correctly identify the emotional state of the speaker. The advantages of a deep convolutional neural network (DCNN) are investigated in the proposed work. A pretrained framework is used to extract the features from speech emotion databases. In this work, we adopt the feature selection (FS) approach to find the discriminative and most important features for SER. Many algorithms are used for the emotion classification problem. We use the random forest (RF), decision tree (DT), support vector machine (SVM), multilayer perceptron classifier (MLP), and k-nearest neighbors (KNN) to classify seven emotions. All experiments are performed by utilizing four different publicly accessible databases. Our method obtains accuracies of 92.02%, 88.77%, 93.61%, and 77.23% for Emo-DB, SAVEE, RAVDESS, and IEMOCAP, respectively, for speaker-dependent (SD) recognition with the feature selection method. Furthermore, compared to current handcrafted feature-based SER methods, the proposed method shows the best results for speaker-independent SER. For EMO-DB, all classifiers attain an accuracy of more than 80% with or without the feature selection technique.

## 1 INTRODUCTION

Recently, there has been much progress in artificial intelligence. However, we are still far short of interacting naturally with machines because machines can neither understand our emotional state nor our emotional behavior. In previous studies, some modalities have been proposed for identifying emotional states, such as extended text (Khan et al. (2021)), speech (El Ayadi et al. (2011)), video (Hossain and Muhammad (2019)), facial expressions (Alreshidi and Ullah (2020)), short messages (Sailunaz et al. (2018)), and physiological signals (Qing et al. (2019)). These modalities vary across applications. The most common modalities in social media are emoticons and short text; video is the most common modality for gaming systems. Electroencephalogram signal-based emotion classification methods have also been introduced recently (Liu et al. (2020); Bazgir et al. (2018); Suhaimi et al. (2020)); however, the use of electroencephalogram signals is invasive and annoying for people.

Due to some inherent advantages, speech signals are the best source for affective computing. Speech signals can be obtained more economically and readily than other biological signals. Therefore, most researchers have focused on automatic speech emotion recognition (SER). There are numerous applications for identifying emotional persons, such as interactions with robots, entertainment, cardboard systems, commercial applications, computer games, audio surveillance, call centers, and banking.

Three main issues should be addressed to obtain a successful SER framework: (i) selecting an excellent emotional database, (ii) performing useful feature extraction, and (iii) using deep learning algorithms to design accurate classifiers. However, emotional feature extraction is a significant problem in an SER framework. In prior studies, many researchers have suggested significant features of speech, such as energy, intensity, pitch, standard deviation, cepstrum coefficients, Mel-frequency cepstrum coefficients (MFCCs), zero-crossing rate (ZCR), formant frequency, filter bank energy (FBR), linear prediction cepstrum coefficients (LPCCs), modulation spectral features (MSFs) and Mel-spectrograms. In (Sezgin et al. (2012)), several distinguishing acoustic features were used to identify emotions: spectral, qualitative, continuous, and Teager energy operator-based (TEO) features. Thus, many researchers have suggested that the feature set comprises more speech emotion information (Rayaluru et al. (2019)). However, combining feature sets complicates the learning process and enhances the possibility of overfitting. In the last five years, researchers have presented many classification algorithms, such as the hidden Markov model (HMM) (Mao et al. (2019)), support vector machine (SVM) (Kurpukdee et al. (2017)), deep belief network (DBN) (Shi (2018)), K-nearest neighbors (KNN) (Zheng et al. (2020)) and bidirectional long short-term memory networks (BiLSTMs) (Mustaqeem et al. (2020)). Some researchers have also suggested different classifiers; in the brain emotional learning model (BEL) (Mustaqeem et al. (2020)), a multilayer perceptron (MLP) and adaptive neuro-fuzzy inference system are combined for SER. The multikernel Gaussian process (GP) (Chen et al. (2016b)) is another proposed classification strategy with two related notions. These provide for learning in the algorithm by combining two functions: the radial basis function (RBF) and the linear kernel function. In (Chen et al. (2016b)), the proposed system extracted two spectral features and used these two features to train different machine learning models. The proposed technique estimated that the combined features had high accuracy, above 90 percent on the Spanish emotional database and 80 percent on the Berlin emotional database. Han et al. adopted both utterance- and segment-level features to identify emotions.

Some researchers have weighted the advantages and disadvantages of each feature. However, no one has identified which feature is the best feature among feature categories (El Ayadi et al. (2011); Sun et al. (2015); Anagnostopoulos et al. (2015)). Many deep learning models have been proposed in SER to determine the high-level emotion features of utterances to establish a hierarchical representation of speech. The accuracy of handcrafted features is relatively high, and this feature extraction technique always requires manual labor (Anagnostopoulos et al. (2015); Chen et al. (2016a, 2012)). The extraction of handcrafted features usually ignores the high-level features. However, the best and most appropriate features that are emotionally powerful must be selected by effective performance for SER.

Therefore, it is more important to select specific speech features that are not affected by country, speaking style of the speaker, culture, or region. Feature selection (FS) is also essential after extraction and is accompanied by an appropriate classifier to recognize emotions from speech. A summary of FS is presented in (Kerkeni et al. (2019)). Both feature extraction and FS effectively reduce computational complexity, enhance learning effectiveness, and reduce the storage needed. To extract the local features, we use a convolutional neural network (CNN) (AlexNet). The CNN automatically extracts the appropriate local features from the augmented input spectrogram of an audio speech signal. When using CNNs for the SER system, the spectrogram is frequently used as the CNN input to obtain high-level features. In recent years, numerous studies have been presented, such as (Abdel-Hamid et al. (2014); Krizhevsky et al. (2017)). The authors used a CNN model for feature extraction of audio speech signals. Recently, deep learning models such as AlexNet (Li et al. (2021)), VGG (Simonyan and Zisserman (2015)), and ResNet (He et al. (2015)) have been used extensively to perform different classification tasks. Additionally, these deep learning models regularly perform much better than shallow CNNs. The main reason is that deep CNNs extract mid-level features from the input data using multilevel convolutional and pooling layers.

The main contributions of this paper are as follows: 1). In the proposed study, AlexNet is used to extract features for a speech emotion recognition system. 2). A feature selection approach is used to enhance the accuracy of SER. 3). The proposed approach performs better than existing handcrafted and deep-learning methods for SD and SI experiments.

The rest of the paper is organized as follows: Part 2 reviews the previous work in SER related to this paper's current study. A detailed description of the emotional dataset used in the presented work and the proposed method for FS and the classifier are discussed in Part 3. The results are discussed in Part 4. Part 5 contains the conclusion and outlines future work.

# 2 BACKGROUND

In this study, five different machine learning algorithms are used for emotion recognition tasks. There are two main parts of SER. One part is based on distinguishing feature extraction from audio signals. The second part is based on selecting a classifier that classifies emotional classes from speech utterances.

## 2.1 Speech Emotion Recognition Using Machine Learning Approaches

Researchers have used different machine learning classifiers to identify emotional classes from speech: SVM (Sezgin et al. (2012)), random forest (RF) (Noroozi et al. (2017)), Gaussian mixture models (GMMs) (Patel et al. (2017)), HMMs (Mao et al. (2019)), CNNs (Christy et al. (2020)), k-nearest neighbors (KNN) (Kapoor and Thakur (2021)) and MLP. These algorithms have been commonly used to identify emotions. Emotions are categorized using two approaches: categorical and dimensional approaches. Emotions are classified into small groups in the categorical approach. Ekman (Ekman (1992)) proposed six basic emotions: anger, happiness, sadness, fear, surprise, and disgust. In the second category, emotions are defined by axes with a combination of several dimensions (Costanzi et al. (2019)). Different researchers have described emotions relative to one or more dimensions. Pleasure-arousal-dominance (PAD) is a three-dimensional emotional state model proposed by (Mehrabian (1996)). Different features are essential in identifying speech emotions from voice. Spectral features are significant and widely used to classify emotions. A decision tree was used to identify emotions from the CASIA Chinese emotion corpus in (Tao et al. (2008)) and achieved 89.6% accuracy. AB Kandali et al. introduced an approach to classify emotion-founded MFCCs as the main features and applied a GMM as a classifier (Kandali et al. (2009)). Milton, A. et al. presented a three-stage traditional SVM classifying different Berlin emotional datasets (Milton et al. (2013)). VB Waghmare et al. adopted spectral features (MFCCs) as the main feature and classified emotions from the Marathi speech dataset (Waghmare et al. (2014)). Demircan, S. et al. extracted MFCC features from the Berlin EmoDB database. They used the KNN algorithm to recognize speech emotions (Demircan and Kahramanli (2014)). The Berlin emotional speech database (EMO-DB) was used in the experiment, and the accuracy obtained was between 90% and 99.5%. Hossain et al. proposed a cloud-based collaborative media system that uses emotions from speech signals and uses standard features such as MFCCs (Hossain.M. Shamim (2014)). Paralinguistic features and prosodic features were utilized to detect emotions from speech in (Alonso et al. (2015)). SVM, a radial basis function neural network (RBFNN), and an autoassociative neural network (AANN) were used to recognize emotions after combining two features, MFCCs and the residual phase (RP), from a music database (Nalini and Palanivel (2016)). SVMs and DBNs were examined utilizing the Chinese academic database (Zhang et al. (2017)). The accuracy using DBNs was 94.5%, and the accuracy of the SVM was approximately 85%. In (C.K. et al. (2017)), particle swarm optimization-based features and high-order statistical features were utilized. Chourasia et al. implemented an SVM and HMM to classify speech emotions after extracting the spectral features from speech signals (Chourasia et al. (2021)).

## 2.2 Speech Emotion Recognition Using Deep Learning Approaches

Low-level handcrafted features are very useful in distinguishing speech emotions. With many successful deep neural network (DNN) applications, many experts have started to target in-depth emotional feature learning. Schmidt et al. used an approach based on linear regression and deep belief networks to identify musical emotions (Schmidt and Kim (2011)). They used the MoodSwings Lite music database and obtained a 5.41% error rate. Duc Le et al. implemented hybrid classifiers, which were a set of DBNs and HMMs, and attained good results on FAU Aibo (Le and Provost (2013)). Deng et al. presented a transfer learning feature method for speech emotion recognition based on a sparse autoencoder. Several databases were used, including the eNTERFACE and EMO-DB databases (Deng et al. (2013)). In (Poon-Feng et al. (2014)), a generalized discriminant analysis method (Gerda) was presented with several Boltzmann machines to analyze and classify emotions from speech and improve the previous reported baseline by traditional approaches. Erik M. Schmidt et al. proposed a regression-based DBN to recognize music emotions and a model based on three hidden layers to learn emotional features (Han et al. (2014)).

Trentin et al. proposed a probabilistic echo-state network-based emotion recognition framework that obtained an accuracy of 96.69% using the WaSep database (Trentin et al. (2015)). More recent work introduced deep retinal CNNs (DRCNNs) in (Niu et al. (2017)), which showed good performance in recognizing emotions from speech signals. The presented approach obtained the highest accuracy, 99.25%, in the IEMOCAP database. In (Fayek et al. (2017)), the authors suggested deep learning approaches. A

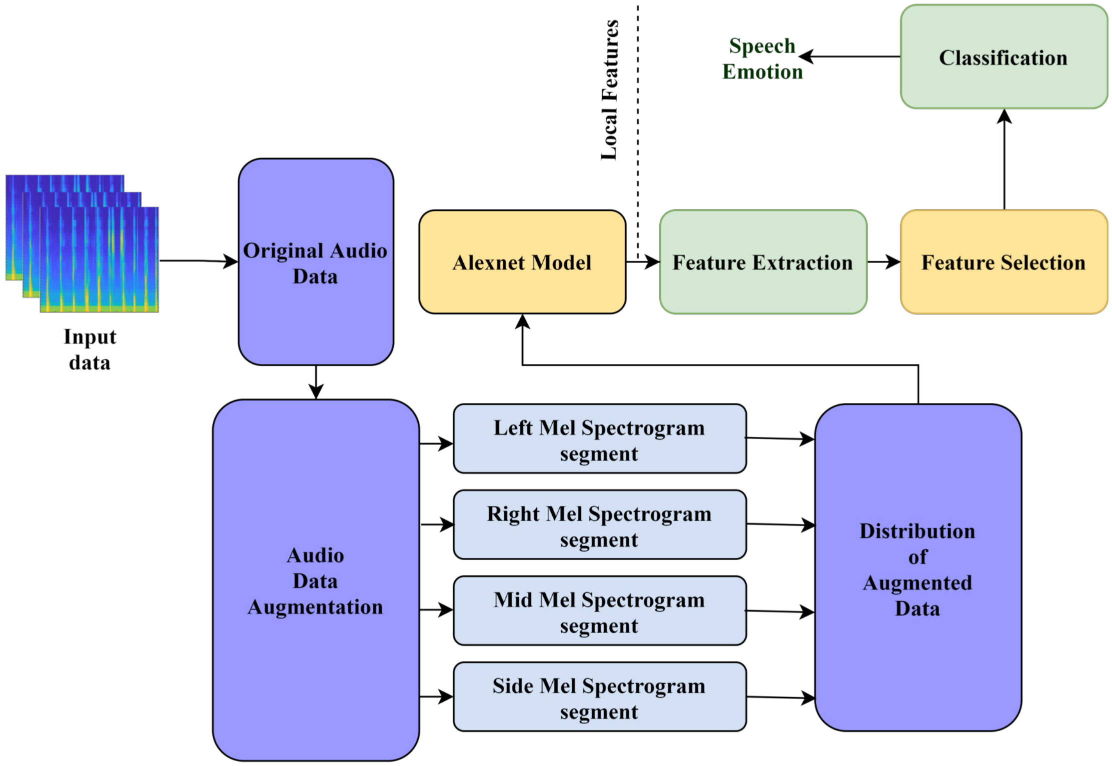


**Figure 1.** The structure of our proposed model for audio emotion recognition

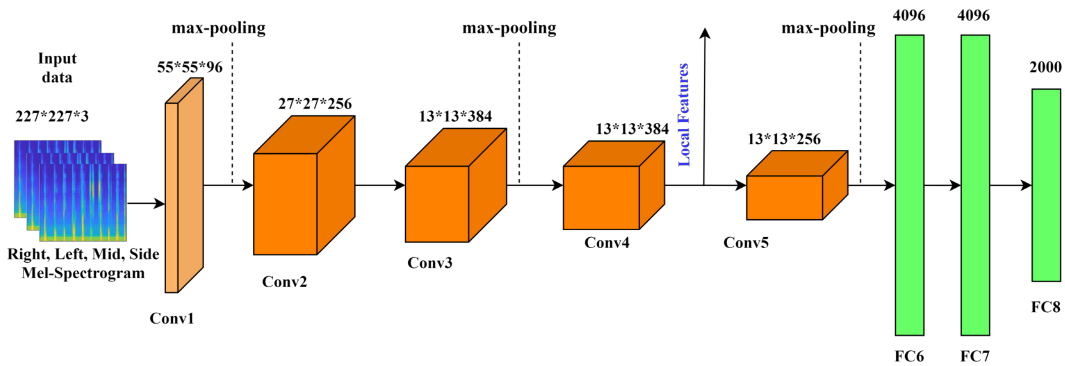


**Figure 2.** The general architecture of AlexNet

speech signal spectrogram was used as an input. The signal may be represented in terms of time and frequency. The spectrogram is a fundamental and efficient way to describe emotional speech impulses in the time-frequency domain. It has been used with particular effectiveness for voice and speaker recognition and word recognition (Stolar et al. (2017)). Satt A and S. Rozenberg et al. suggested another efficient convolutional LSTM approach for emotion classification. The introduced model learned spatial patterns and spatial spectrogram patterns representing information on the emotional states (Satt et al. (2017)). The experiment was performed on the IEMOCAP database with four emotions. Two different databases were used to extract prosodic and spectral features with an ensemble softmax regression approach (Sun and Wen (2017)). For the identification of emotional groups, experiments were performed on the two different datasets. A CNN was used in (Fayek et al. (2017)) to classify four emotions from the IEMOCAP database: happy, neutral, angry, and sad. In (Xia and Liu (2017)), multitasking learning was used to obtain activation and valence data for speech emotion detection using the DBN model. IEMOCAP was used in the experiment to identify the four emotions. However, high computational costs and a large amount of data are required for deep learning techniques. The majority of current speech emotional

**Table 1.** Nomenclature

| | | | |
|---|---|---|---|
| ACRNN | Attention Convolutional Recurrent Neural Network | KNN | K-Nearest Neighbors |
| BEL | Brain Emotional Learning | LPCC | Linear Predictive Cepstral Coefficients |
| BiLSTM | Bidirectional Long Short-Term Memory | MFCC | Mel Frequency Cepstral Coefficients |
| CNN | Convolutional Neural Network | MLP | Multilayer Perceptron |
| CL | Convolutional Layer | MSF | Modulation Spectral Features |
| CNN | Convolutional Neural Network | PAD | Pleasure-Arousal-Dominance |
| CFS | Correlation-Based Feature Selection | PL | Pooling Layer |
| DBN | Deep Belief Network | RBFNN | Radial Basis Function Neural Network |
| DCNN | Deep Convolutional Neural Network | RBF | Radial Basis Function |
| DNN | Deep Neural Network | RF | Random Forest |
| DRCNN | Deep Retinal CNNs | RP | Residual Phase |
| DT | Decision Tree | RNN | Recurrent Neural Network |
| FS | Feature Selection | SAVEE | Surrey Audio-Visual Expressed Emotion |
| FCL | Fully Connected Layer | SD | Speaker-Dependent |
| FBR | Filter Bank Energy | SI | Speaker- Independent |
| GMM | Gaussian Mixture Model | SVM | Support Vector Machine |
| GP | Gaussian Process | SER | Speech Emotion Recognition |
| HMM | Hidden Markov Model | TEO | Teager Energy Operator |
| KELM | Kernel Extreme Learning Machine | ZCR | Zero-Crossing Rate |

databases have a small amount of data. Deep learning model approaches are insufficient for training with large-scale parameters. A pretrained deep learning model is used based on the above studies. In (Badshah et al. (2017)), a pretrained DCNN model was introduced for speech emotion recognition. The outcomes were improved with seven emotional states. In (Badshah et al. (2017)), the authors suggested a DCNN accompanied by discriminant temporal pyramid matching. DNNs were used to divide emotional probabilities into segments (Gu et al. (2018)), which were utilized to create utterance features; these probabilities were fed to the classifier. The IEMOCAP database was used in the experiment, and the obtained accuracy was 54.3%. In (Zhao et al. (2018)), the suggested approach used integrated attention with a fully convolutional network (FCN) to automatically learn the optimal spatiotemporal representations of signals from the IEMOCAP database. The hybrid architecture proposed in (Etienne et al. (2018)) included a data augmentation technique. In (Wang and Guan (2008); Zhang et al. (2018)), the fully connected layer (FC7) of AlexNet was used for the extraction process. The results were evaluated on four different databases. In (Guo et al. (2018)), an approach for SER that combined phase and amplitude information utilizing a CNN was investigated. In (Chen et al. (2018)), a three-dimensional convolutional recurrent neural network including an attention mechanism (ACRNN) was introduced. The identification of emotion was evaluated using the Emo-DB and IEMOCAP databases. The attention process was used to develop a dilated CNN and BiLSTM in (Meng et al. (2019)). To identify speech emotion, 3D log-Mel spectrograms were examined for global contextual statistics and local correlations. The OpenSMILE package was used to extract features in (Özseven (2019)). The accuracy obtained with the Emo-DB database was 84%, and it was 72% with the SAVEE database. Pretrained networks have many benefits, including the ability to reduce the training time and improve accuracy. Kernel extreme learning machine (KELM) features were introduced in (Guo et al. (2019)). An adversarial data augmentation network was presented in (Yi and Mak (2019)) to create simulated samples to resolve the data scarcity problem. Energy and pitch were extracted from each audio segment in (Ververidis and Kotropoulos (2005); Rao

**Table 2.** Detailed description of the datasets

| Datasets | Speakers | Emotions | Languages | Size |
|---|---|---|---|---|
| RAVDESS | 24 Actors (12 male, 12 female) | eight emotions ( calm, neutral, angry, happy, fear, surprise, sad, disgust ) | North American English | 7356 files (total size: 24.8 GB). |
| SAVEE | 4 (male) | seven emotions (sadness, neutral, frustration, happiness, disgust ,anger, surprise) | British English | 480 utterances (120 utterances per speaker) |
| Emo-DB | 10 (5 male, 5 female) | seven emotions (neutral, fear, boredom, disgust, sad, angry, joy) | German | 535 utterances |
| IEMOCAP | 10 (5 male, 5 female) | nine emotions (surprise, happiness, sadness, anger, fear, excitement, neutral, frustration and others) | English | 12 hours of recordings |

**Table 3.** Categories of emotional speech databases, their features, and some examples of each category.

| | Simulated | Semi Natural |
|---|---|---|
| Description | generated by trained and experienced actors delivering the same sentence with different degrees of emotion | created by having individuals read a script with a different emotions |
| Single emotion at a time | yes | yes |
| Widely used | yes | no |
| Copyrights and privacy protection | yes | yes |
| Includes contextual information | no | yes |
| Includes situational information | no | yes |
| Emotions that are separate and distinct | yes | no |
| Numerous emotions | yes | yes |
| Simple to model | yes | no |
| Numerous emotions | yes | yes |
| Examples | EMO-DB, RAVDESS | IEMOCAP , SAVEE |

et al. (2013); Daneshfar et al. (2020)). They also needed fewer training data and could deal directly with dynamic variables. Two different acoustic paralinguistic feature sets were used in (Haider et al.

(2020)). An implementation of real-time voice emotion identification using AlexNet was described in (Lech et al. (2020)). When trained on the Berlin Emotional Speech (EMO-DB) database, the presented method obtained an average accuracy of 82%.

# 3 PROPOSED METHOD

This section describes the proposed pretrained CNN (AlexNet) algorithm for the SER framework. We fine-tune the pretrained model (Krizhevsky et al. (2017)) on the created image-like Mel-spectrogram segments. We do not train our own deep CNN framework owing to the limited emotional audio dataset. Furthermore, computer vision experiments (Ren et al. (2016); Campos et al. (2017)) have depicted that fine-tuning the pretrained CNNs on target data is acceptable to relieve the issue of data insufficiency. AlexNet is a model pretrained on the extensive ImageNet dataset, containing a wide range of different labeled classes, and uses a shorter training time. AlexNet (Krizhevsky et al. (2017)) comprises five convolution layers, three max-pooling layers, and three fully connected layers. In the proposed work, we extract the low-level features from the fourth convolutional layer (CL4).

The architecture of our proposed model is displayed in Figure 1. Our model comprises four processes: (a) development of the audio input data, (b) low-level feature extraction using AlexNet, (c) feature selection, and (d) classification. Below, we explain all four steps of our model in detail.

## 3.1 Creation of the Audio Input

In the proposed method, the Mel-spectrogram segment is generated from the original speech signal. We create three channels of the segment from the original 1D audio speech dataset. Then, the generated segments are converted into fixed-size $227 \times 227 \times 3$ inputs for the proposed model. Following (Zhang et al. (2018)), 64 Mel-filter banks are used to create the log Mel-spectrogram, and each frame is multiplied by a 25 ms window size with a 10 ms overlap. Then, we divide the log Mel spectrogram into fixed segments by using a 64-frame context window. Finally, after extracting the static segment, we calculate the regression coefficients of the first and second order around the time axis, thereby generating the delta and double-delta coefficients of the static Mel spectrogram segment. Consequently, three channels with $64 \times 64 \times 3$ Mel-spectrogram segments can be generated as the inputs of AlexNet, and these channels are identical to the color RGB image. Therefore, we resize the original $64 \times 64 \times 3$ spectrogram to the new size $227 \times 227 \times 3$. In this case, we can create four (middle, side, left, and right) segments of the Mel spectrogram, as shown in Figure 2.

## 3.2 Emotion Recognition Using AlexNet

In the proposed method, CL4 of the pretrained model is used for feature extraction. The CFS feature selection approach is used to select the most discriminative features. The CFS approach selects only very highly correlated features with output class labels. The five different classification models are used to test the accuracy of the feature subsets.

## 3.3 Feature Extraction

In this study, feature extraction is performed using a pretrained model. The original weight of the model remains fixed, and existing layers are used to extract the features. The pretrained model has a deep structure that contains extra filters for every layer and stacked CLs. It also includes convolutional layers, max-pooling layers, momentum stochastic gradient descent, activation functions, data augmentation, and dropout. AlexNet uses a rectified linear unit (ReLU) activation function. The layers of the network are explained below.

### *3.3.1 Input Layer*

This layer of the pretrained model is a fixed-size input layer. We resample the Mel spectrogram of the signal to a fixed size $227 \times 227 \times 3$.

### *3.3.2 Convolutional Layer (CL)*

The convolutional layer is composed of convolutional filters. Convolutional filters are used to obtain many local features in the input data from local regions to form various feature groups. AlexNet contains five CLs, in which three layers follow the max-pooling layer. CL1 includes 96 kernels with a size of $11 \times 11 \times 3$, zero padding, and a stride of 4 pixels. CL2 contains 256 kernels, each of which is $5 \times 5 \times 48$ in size and includes a 1-pixel stride and a padding value of 2. The CL3 contains 384 kernels of size

$3 \times 3 \times 256$. CL4 contains 384 kernels of size $3 \times 3 \times 192$. For the output value of each CL, the ReLU function is used, which speeds up the training process.

#### 3.3.3 Pooling Layer (PL)

After the CLs, a pooling layer is used. The goal of the pooling layer is to subsample the feature groups. The feature groups are obtained from the previous CLs to create a single data convolutional feature group from the local areas. Average pooling and max-pooling are the two basic pooling operations. The max-pooling layer employs maximum filter activation across different points in a quantified frame to produce a modified resolution type of CL activation.

#### 3.3.4 Fully Connected Layers (FCLs)

Fully connected layers incorporate the characteristics acquired from the PL and create a feature vector for classification. The output of the CLs and PLs is given to the fully connected layers. There are three fully connected layers in AlexNet: FC6, FC7, and FC8. A 4096-dimensional feature map is generated by FC6 and FC7, while FC8 generates 1000-dimensional feature groups.

Feature maps can be created using FCLs. These are universal approximations, but fully connected layers do not work fully in recognizing and generalizing the original image pixels. CL4 extracts relevant features from the original pixel values by preserving the spatial correlations inside the image. Consequently, in the experimental setup, features are extracted from the CL4 employed for SER. A total of 64,896 features are obtained from CL4. Certain features are followed by an FS method and pass through a classification model for identification.

### 3.4 Feature Selection

The discriminative and related features for the model are determined by feature selection. FS approaches are used with several models to minimize the training time and enhance the ability to generalize by decreasing overfitting. The main goal of feature selection is to remove insignificant and redundant features.

### 3.5 Correlation-Based Measure

We can identify an excellent feature if it is related to the class features and is not redundant with respect to any other class features. For this reason, we use entropy-based information theory. The equation of entropy-based information theory is defined as:

$$F(E) = -\Sigma S(e_j) log2(S(e_j)). \quad (1)$$

The entropy of E after examining the values of G is defined in the equation below:

$$F(E/G) = -\Sigma S(g_k)\Sigma S(e_j/g_k) log2(S(e_j/g_k)) \quad (2)$$

$S(e_j)$ denotes the probability for all values of E, whereas $S(e_j/g_k)$ denotes the probabilities of E when the values of G are specified. The percentage by which the entropy of E decreases reflects the irrelevant information about E given by G, which is known as information gain. The equation of information gain is given below:

$$I(E/G) = (F(E) - F(E/G)). \quad (3)$$

If $I(E/G)$ ¿ $I(H/G)$, then we can conclude that feature G is much more closely correlated to feature E than to feature H. We possess one more metric, symmetrical uncertainty, which indicates the correlation between features, defined by the equation below:

$$SU(E,G) = 2[I(E/G)/F(E) + F(G)]. \quad (4)$$

SU balances the information gain bias toward features with more values by normalizing its value to the range [0,1]. SU analyzes a pair of features symmetrically. Entropy-based techniques need nominal features. These features can be used to evaluate the correlations between continuous features if these features are discretized properly.

We use the correlation feature-based approach (CFS) Wosiak and Zakrzewska (2018) in the proposed work based on the previously described techniques. It evaluates a subset of features and selects only highly correlated discriminative attributes. CFS ranks the features by applying a heuristic correlation evaluation function. It estimates the correlation within the features. CFS drops unrelated features that have limited similarity with the class label. The CFS equation is as follows:

$$FS = \max_{Sk} \frac{r_{cf1} + r_{cf2} + r_{cf3} + .... + r_{cfk}}{\sqrt{k + 2(r_{f1f2} + .... + r_{fifj} + .... + r_{fkfk-1})}}, \tag{5}$$

where k represents the total number of features, $r_{cfi}$ represents the classification correlation of the features, and $r_{fifj}$ represents the correlation between features. The extracted features are fed into classification algorithms. CFS usually deletes (backward selection) or adds (forward selection) one feature at a time.

### 3.6 Classification Methods

The discriminative features provide input to the classifiers for emotion classification. In the proposed method, five different classifiers, KNN, RF, decision tree, MLP, and SVM, are used to evaluate the performance of speech emotion recognition.

### 3.7 Support Vector Machine (SVM)

SVMs are used for binary regression and classification. They create an optimal higher-dimensional space with a maximum class margin. SVMs identify the support vectors $v_j$, weights $w_{fj}$, and bias b to categorize the input information. For classification of the data, the following expression is used:

$$sk(v, v_j) = (\rho v^e v_j + k)^z. \tag{6}$$

In the above equations, k is a constant value, and b represents the degree of the polynomial. For a polynomial $\rho$ ¿ zero:

$$v = (\Sigma_{i=0}^{n} w_{fj} sk(v_j, v) + b. \tag{7}$$

In the above equation, sk represents the kernel function, v is the input, vj is the support vector, wfj is the weight, and b is the bias. In our study, we utilize the polynomial kernel to translate the data into a higher-dimensional space.

### 3.8 k-Nearest Neighbors (KNN)

This classification algorithm keeps all data elements. It identifies the most comparable N examples and employs the target class emotions for all data examples based on similarity measures. In the proposed study, we fixed N = 10 for emotional classification. The KNN method finds the ten closest neighbors using the Euclidean distance, and emotional identification is performed using a majority vote.

### 3.9 Random Forest (RF)

An RF is a classification and regression ensemble learning classifier. It creates a class of decision trees and a meaningful indicator of the individual trees for data training. The RF replaces each tree in the database at random, resulting in unique trees, in a process called bagging. The RF splits classification networks based on an arbitrary subset of characteristics per tree.

### 3.10 Multilayer Perceptron (MLP)

MLPs are neural networks that are widely employed in feedforward processes. They consist of multiple computational levels. Identification issues may be solved using MLPs. They use a supervised back-propagation method for classifying occurrences. The MLP classification model consists of three layers: the input layer, the hidden layers, and the output layer. The input layer contains neurons that are directly proportional to the features. The degree of the hidden layers depends on the overall degree of the emotions in the database. It features dimensions after the feature selection approach. The number of output neurons in the database is equivalent to the number of emotions. The sigmoid activation function utilized in this study is represented as follows:

$$p_i = \frac{1}{1 + e^- qi} \tag{8}$$

In the above equation, the state is represented by pi, whereas the entire weighted input is represented by qi. When using the Emo-DB database, there is only one hidden layer in the MLP. It has 232 neurons. When using the SAVEE database, there is only one hidden layer in the MLP, and it comprises 90 neurons. The MLP contains a single hidden layer, and 140 neurons are present in the IEMOCAP database. In comparison, one hidden layer and 285 neurons are present in the RAVDESS dataset. The MLP is a two-level architecture; thus, identification requires two levels: training and testing. The weight values are set throughout the training phase to match them to the particular output class.

# 4 EXPERIMENTS

## 4.1 Datasets

This experimental study contains four emotional speech databases, and these databases are publicly available.

- **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS):** RAVDESS is an audio and video database consisting of eight acted emotional categories: calm, neutral, angry, surprise, fear, happy, sad, and disgust, and these emotions are recorded only in North American English. RAVDESS was recorded by 12 male and 12 female professional actors.

- **Surrey Audio-Visual Expressed Emotion (SAVEE):** The SAVEE database contains 480 emotional utterances. The SAVEE database was recorded in British English by four male professional actors with seven emotion categories: sadness, neutral, frustration, happiness, disgust, anger, and surprise.

- **Berlin Emotional Speech Database (Emo-DB):** The Emo-DB dataset contains 535 utterances with seven emotion categories: neutral, fear, boredom, disgust, sad, angry, and joy. The Emo-DB emotional dataset was recorded in German by five male and five female native-speaker actors.

- **Interactive Emotional Dyadic Motion Capture (IEMOCAP):** The IEMOCAP multispeaker database contains approximately 12 hours of audio and video data with seven emotional states, surprise, happiness, sadness, anger, fear, excitement, and frustration, as well as neutral and other states. The IEMOCAP database was recorded by five male and five female professional actors. In this work, we use four (neutral, angry, sadness, and happiness) class labels.

## 4.2 Experimental Setup

All the experiments are completed in version 3.9.0 of the Python language framework. Numerous API libraries are used to train the five distinct models. The framework uses Ubuntu 20.04. The key objective is to implement an input data augmentation and feature selection approach for the five different models. The feature extraction technique is also involved in the proposed method.

### *4.2.1 Anaconda*

Anaconda is the best data processing and scientific computing platform for Python. It already includes numerous data science and machine learning libraries. Anaconda also includes many popular visualization libraries, such as matplotlib. It also provides the ability to build a different environment with a few unique libraries to carry out the task.

### *4.2.2 Keras*

The implementation of our model for all four datasets was completed from scratch using Keras. It makes it extremely simple for the user to add and remove layers and activate and utilize the max-pooling layer in the network.

### *4.2.3 Librosa*

Librosa (McFee et al. (2015)) is a basic Python library used for this research. Librosa is used to examine the audio signal recordings. The four (side, middle, left, and right) segments of the Mel spectrogram were obtained through Librosa.

# 5 EXPERIMENTAL RESULTS AND ANALYSIS

### 5.0.1 Speaker-Dependent (SD) Experiments

The performance of the proposed SER system is assessed using benchmark databases for the SD experiments. We use ten-fold cross-validation in our studies. All databases are divided randomly into ten equal complementary subsets with a dividing ratio of 80:20 to train and test the model. Table 4 gives the results achieved by five different classifiers utilizing the features extracted from CL4 of the model. The SVM achieved 92.11%, 87.65%, 82.98%, and 79.66% accuracies for the Emo-DB, RAVDESS, SAVEE and IEMODB databases, respectively. The proposed method reported the highest accuracy of 86.56% on the Emo-DB database with KNN. The MLP classifier obtained 86.75% accuracy for the IEMOCAP database. In contrast, the SVM reported 79.66% accuracy for the IEMOCAP database. The MLP classifier reported the highest accuracy, 91.51%, on the Emo-DB database. The RF attained 82.47% accuracy on the Emo-DB database, while DT achieved 80.53% accuracy on Emo-DB.

**Table 4.** Standard deviation and weighted average recall of the SD experiments without FS

| | SVM | RF | KNN | MLP | DT |
|---|---|---|---|---|---|
| RAVDESS | 87.65±1.79 | 78.65±4.94 | 78.15±3.39 | 80.67±2.89 | 76.28±3.24 |
| SAVEE | 82.98±4.87 | 78.38±4.10 | 79.81±4.05 | 81.13±3.63 | 69.15±2.85 |
| Emo-DB | 92.11±2.29 | 82.47±3.52 | 86.56±2.78 | 91.51±2.09 | 80.53±4.72 |
| IEMODB | 79.66±4.44 | 80.93±3.75 | 74.33±3.37 | 86.75±3.64 | 67.25±2.33 |

Table 4 represents the results of the FS approach. The proposed FS technique selected 460 distinguishing features out of a total of 64,896 features for the Emo-DB dataset. The FS method obtained 170,465,277 feature maps for the SAVEE, RAVDESS, and IEMOCAP datasets.

The experimental results illustrate a significant accuracy improvement by using data resampling and the FS approach. We consider the standard deviation and average weighted recall to evaluate the performance and stability of the SD experiments using the FS approach. The SVM classifier reached 93.61% and 96.02% accuracy for RAVDESS and Emo-DB, respectively, while the obtained accuracies were 88.77% and 77.23% for SAVEE and IEMOCAP, respectively, through the SVM. The MLP classifier obtained 95.80% and 89.12% accuracies with the Emo-DB and IEMOCAB databases, respectively.

The KNN classifier obtained the highest accuracy, 92.45% and 88.34%, with the Emo-DB and RAVDEES datasets. The RF classifier reported the highest accuracy, 93.51%, on the Emo-DB dataset and 86.79% accuracy on the SAVEE dataset with the feature selection approach. The results of the confusion matrix were used to evaluate the identification accuracy of the individual emotional labels. Table 5 shows that the SVM obtained better recognition accuracy than the other classification models with the FS method. As shown in Figure 5, the SVM recognized ”frustration” and ”neutral” with the highest accuracies, 88.33% and 91.66%, with the SAVEE dataset. As shown in Figure 6, the RAVDESS dataset contains eight emotions, including ”anger”, ”calm”, ”fear”, and ”neutral”, which are listed with accuracies of 96.32%, 97.65%, 95.54%, and 99.98%, respectively. The IEMOCAP database identified ”anger” with the highest accuracy of 93.23%, while ”happy,” ”sad,” and ”neutral” were recognized with the highest accuracies of 83.41%, 91.45%, and 89.65% with the MLP classifier, respectively.

### 5.0.2 Speaker-Independent (SI) Experiments

We adopted the single-speaker-out (SSO) method for the SI experiments. One annotator was used for testing, and all other annotators were used for training. In the proposed approach, the IEMOCAP dataset

**Table 5.** Standard deviation and weighted average recall of the SD experiments with FS

| Database | SVM | RF | KNN | MLP | DT |
|---|---|---|---|---|---|
| RAVDESS | 93.61±1.32 | 85.21±3.55 | 88.34±2.67 | 84.50±2.23 | 78.45±2.67 |
| SAVEE | 88.77±2.45 | 86.79±2.96 | 83.45±3.21 | 85.45±3.12 | 75.68±3.82 |
| Emo-DB | 96.02±1.07 | 93.51±2.21 | 92.45±2.45 | 95.80±2.34 | 79.13±4.01 |
| IEMODB | 77.23±2.66 | 86.23±2.54 | 82.78±2.17 | 89.12±2.57 | 72.32±1.72 |

**Table 6.** Standard deviation and weighted average recall of the SI experiment results without FS

| | SVM | RF | KNN | MLP | DT |
|---|---|---|---|---|---|
| RAVDESS | 75.34±2.58 | 65.78±2.32 | 69.12±2.20 | 71.01±2.84 | 67.41±2.37 |
| SAVEE | 63.02±3.21 | 59.66±3.79 | 71.81±3.81 | 65.18±2.05 | 59.55±2.23 |
| Emo-DB | 87.65±2.56 | 79.45±2.11 | 75.30±2.19 | 88.32±2.67 | 76.27±2.35 |
| IEMODB | 61.85±3.20 | 60.11±4.20 | 55.47±2.96 | 63.18±1.62 | 54.69±3.72 |

**Table 7.** Standard deviation and weighted average recall of the SI experiment results with FS

| Database | SVM | RF | KNN | MLP | DT |
|---|---|---|---|---|---|
| RAVDESS | 80.94±2.17 | 76.82±2.16 | 75.57±3.29 | 82.75±2.10 | 76.18±1.33 |
| SAVEE | 70.06±3.33 | 65.55±2.42 | 60.58±3.84 | 75.38±2.74 | 63.69±2.22 |
| Emo-DB | 90.78±2.45 | 85.73±2.58 | 81.32±2.12 | 92.65±3.09 | 78.21±3.47 |
| IEMODB | 84.00±2.76 | 78.08±2.65 | 76.44±3.88 | 80.23±2.77 | 75.78±2.25 |

**Table 8.** Comparison of the SD experiments with existing methods.

| Database | Reference | Feature | Accuracy(%) |
|---|---|---|---|
| RAVDESS | (Bhavan et al. (2019)) | Spectral Centroids, MFCC and MFCC derivatives | 75.69 |
| RAVDESS | **Proposed Approach** | **AlexNet+FS+RF** | **86.79** |
| RAVDESS | **Proposed Approach** | **AlexNet+FS+SVM** | **88.77** |
| SAVEE | (Özseven (2019)) | OpenSmile Features | 72.39 |
| SAVEE | **Proposed Approach** | **AlexNet+FS+RF** | **86.79** |
| SAVEE | **Proposed Approach** | **AlexNet+FS+SVM** | **88.77** |
| Emo-DB | (Guo et al. (2018)) | Amplitude spectrogram and phase information | 91.78 |
| Emo-DB | (Chen et al. (2018)) | 3-D ACRNN | 82.82 |
| Emo-DB | (Meng et al. (2019)) | Dilated CNN + BiLSTM | 90.78 |
| Emo-DB | (Özseven (2019)) | OpenSMILE features | 84.62 |
| Emo-DB | (Bhavan et al. (2019)) | Spectral Centroids, MFCC and MFCC derivatives | 92.45 |
| Emo-DB | **Proposed Approach** | **AlexNet+FS+MLP** | **95.80** |
| Emo-DB | **Proposed Approach** | **AlexNet+FS+SVM** | **96.02** |
| IEMOCAP | (Satt et al. (2017)) | 3 Convolution Layers + LSTM | 68.00 |
| IEMOCAP | (Chen et al. (2018)) | 3-D ACRNN | 64.74 |
| IEMOCAP | (Zhao et al. (2018)) | Attention-BLSTM-FCN | 64.00 |
| IEMOCAP | (Etienne et al. (2018)) | CNN+LSTM | 64.50 |
| IEMOCAP | (Meng et al. (2019)) | Dilated CNN + BiLSTM | 74.96 |
| IEMOCAP | **Proposed Approach** | **AlexNet+FS+MLP** | **89.12** |
| IEMOCAP | **Proposed Approach** | **AlexNet+FS+RF** | **86.23** |

**Table 9.** Comparison of SI experiments with existing methods.

| Database | Reference | Feature | Accuracy (%) |
|---|---|---|---|
| RAVDESS | **Proposed Approach** | **AlexNet+FS+MLP** | **82.75** |
| RAVDESS | **Proposed Approach** | **AlexNet+FS+SVM** | **80.94** |
| SAVEE | (Sun and Wen (2017)) | Ensemble soft-MarginSoftmax (EM-Softmax) | 51.50 |
| SAVEE | (Haider et al. (2020)) | eGeMAPs and emobase | 42.40 |
| SAVEE | **Proposed Approach** | **AlexNet+FS+MLP** | **75.38** |
| SAVEE | **Proposed Approach** | **AlexNet+FS+SVM** | **70.06** |
| Emo-DB | (Badshah et al. (2017)) | DCNN + DTPM | 87.31 |
| Emo-DB | (Sun and Wen (2017)) | Ensemble soft-MarginSoftmax (EM-Softmax) | 82.40 |
| Emo-DB | (Yi and Mak (2019)) | OpenSmile Features + ADAN | 83.74 |
| Emo-DB | (Guo et al. (2019)) | Statistical Features and Empirical Features+ KELM | 84.49 |
| Emo-DB | (Meng et al. (2019)) | Dilated CNN+ BiLSTM | 85.39 |
| Emo-DB | (Haider et al. (2020)) | eGeMAPs and emobase | 76.90 |
| Emo-DB | (Lech et al. (2020)) | AlexNet | 82.00 |
| Emo-DB | (Mustaqeem et al. (2020)) | Radial Basis Function Network( RBFN) + Deep BiLSTM | 85.57 |
| Emo-DB | **Proposed Approach** | **AlexNet+FS+MLP** | **92.65** |
| Emo-DB | **Proposed Approach** | **AlexNet+FS+SVM** | **90.78** |
| IEMOCAP | (Xia and Liu (2017)) | SP + CNN | 64.00 |
| IEMOCAP | (Chen et al. (2018)) | Dilated CNN+ BiLSTM | 69.32 |
| IEMOCAP | Guo et al. (2019) | Statistical Features and Empirical Features+ KELM | 57.10 |
| IEMOCAP | (Yi and Mak (2019)) | OpenSmile Features + ADAN | 65.01 |
| IEMOCAP | (Daneshfar et al. (2020)) | IS10 + DBN | 64.50 |
| IEMOCAP | (Mustaqeem et al. (2020)) | Radial Basis Function Network( RBFN) + Deep BiLSTM | 72.2 |
| IEMOCAP | **Proposed Approach** | **AlexNet+FS+MLP** | **89.12** |
| IEMOCAP | **Proposed Approach** | **AlexNet+FS+RF** | **86.23** |

| | anger | fear | sad | neutral | boredom | disgust | happy |
|---|---|---|---|---|---|---|---|
| anger | **93.42** | 1.68 | 0 | 0 | 0 | 0 | 4.88 |
| fear | 3.52 | **94.81** | 0 | 0.55 | 0.55 | 0 | 0.55 |
| sad | 0 | 0.55 | **98.88** | 0 | 0.55 | 0 | 0 |
| neutral | 0 | 0 | 0 | **97.45** | 2.53 | 0 | 0 |
| boredom | 0 | 0 | 0.55 | 2.87 | **96.56** | 0 | 0 |
| disgust | 0 | 0 | 0 | 0.65 | 0.55 | **98.78** | 0 |
| happy | 4.88 | 0.55 | 0 | 0.55 | 0 | 0.55 | **93.45** |

**Figure 3.** Confusion matrix obtained by the SVM on the Emo-DB database for the SD experiment

| | anger | surprise | sad | neutral | frustration | disgust | happy |
|---|---|---|---|---|---|---|---|
| anger | **91.32** | 0 | 1.67 | 0 | 0 | 1.67 | 5.32 |
| surprise | 3.00 | **89.63** | 0 | 0.44 | 0.44 | 0.44 | 6.03 |
| sad | 0.44 | 0 | **87.20** | 9.00 | 0 | 2.90 | 0.44 |
| neutral | 0.55 | 0 | 0.44 | **92.45** | 0.53 | 6.99 | 0.55 |
| frustration | 0.44 | 0 | 0.44 | 0.44 | **97.78** | 0.44 | 0.44 |
| disgust | 0.44 | 0 | 6.74 | 8.34 | 0 | **81.90** | 2.56 |
| happy | 10 | 0.44 | 0.44 | 0.44 | 5.54 | 2.56 | **80.56** |

**Figure 4.** Confusion matrix obtained by the SVM on the SAVEE database for the SD experiment

was split into testing and training sessions. By switching all of the testing annotators, the process was repeated, and the average accuracy was obtained for every testing speaker. Table 7 lists the identification results of five classification models for the SI experiments without the FS technique. The MLP obtained the highest accuracy, 88.32%, with the Emo-DB dataset. With the SAVE database, MLP obtained the highest accuracy, 65.18%. The SVM achieved the highest accuracy of 87.65% with Emo-DB and 75.34% with the RAVDESS database. The random forest achieved the highest accuracies, 79.45% and 65.78%,

| | anger | surprise | sad | neutral | fear | disgust | happy | calm |
|---|---|---|---|---|---|---|---|---|
| anger | **96.32** | 0 | 0.44 | 0.56 | 0.44 | 0.75 | 1.47 | 0 |
| surprise | 1.25 | **93.11** | 0.30 | 2.55 | 0.98 | 0.44 | 1.35 | 0 |
| sad | 2.31 | 1.12 | **85.20** | 3.32 | 0.32 | 3.41 | 1.85 | 2.45 |
| neutral | 0 | 0 | 0 | **99.98** | 0 | 0 | 0 | 0 |
| fear | 0.55 | 1.24 | 0.34 | 0.55 | **95.54** | 0 | 1.22 | 0.20 |
| disgust | 4.44 | 1.45 | 1.56 | 0.98 | 0 | **90.78** | 0.78 | 0.78 |
| happy | 1.98 | 3.29 | 1.78 | 2.51 | 1.25 | 0.44 | **88.61** | 0.12 |
| calm | 0 | 0 | 0.66 | 1.23 | 0.44 | 0 | **0** | **97.65** |

**Figure 5.** Confusion matrix obtained by the SVM on the RAVDESS database for the SD experiment

| | anger | neutral | sad | happy |
|---|---|---|---|---|
| anger | **93.23** | 4.07 | 1.23 | 1.45 |
| neutral | 2.52 | **89.65** | 5.03 | 2.78 |
| sad | 1.07 | 3.79 | **91.45** | 3.67 |
| happy | 1.54 | 16.57 | 1.88 | **83.41** |

**Figure 6.** Confusion matrix obtained by the MLP on the IEMOCAP database for the SD experiment

with Emo-DB and RAVDESS, respectively. Table 7 represents the outcomes for the SI experiments with the feature extraction approach with data resampling and the FS method. The FS and data resampling approach improved the accuracy, according to the preliminary results.

We report the average weighted recall and standard deviation to evaluate the SI experiment's performance and stability utilizing the FS method. The SVM obtained the highest accuracies, 90.78%, 84.00%,

| | anger | surprise | sad | neutral | frustration | disgust | happy |
|---|---|---|---|---|---|---|---|
| **anger** | **94.22** | 0.22 | 0.22 | 0 | 0.44 | 2.44 | 2.44 |
| **surprise** | 9.14 | **70** | 2.44 | 0.44 | 10.54 | 2.44 | 4.98 |
| **sad** | 2.44 | 0 | **85.33** | 5.77 | 1.78 | 2.22 | 2.44 |
| **neutral** | 0.22 | 0.44 | 4.46 | **90.66** | 0.22 | 3.76 | 0.22 |
| **frustration** | 2.44 | 11.54 | 4.98 | 2.44 | **69.08** | 2.44 | 7.06 |
| **disgust** | 2.44 | 0.22 | 8.72 | 16.33 | 4.78 | **58.77** | 8.72 |
| **happy** | 19.71 | 8.72 | 2.44 | 0.44 | 10.90 | 0.44 | **57.33** |

**Figure 7.** Confusion matrix obtained by the SVM on the RAVDESS database for the SI experiment

| | anger | surprise | sad | neutral | fear | disgust | happy | calm |
|---|---|---|---|---|---|---|---|---|
| **anger** | **91.35** | 2.58 | 0.75 | 0 | 0.44 | 1.78 | 1.61 | 1.47 |
| **surprise** | 7.45 | **80.55** | 5.37 | 0 | 0.98 | 1.66 | 3.43 | 0.54 |
| **sad** | 6.23 | 1.86 | **72.10** | 6.77 | 1.78 | 1.75 | 1.88 | 7.61 |
| **neutral** | 0 | 2.65 | 2.66 | **84.97** | 0 | 1.45 | 2.65 | 5.60 |
| **fear** | 2.38 | 2.39 | 1.10 | 0.44 | **90.56** | 1.10 | 1.45 | 0.56 |
| **disgust** | 3.45 | 1.15 | 1.98 | 0.44 | 0.78 | **88.62** | 1.15 | 2.41 |
| **happy** | 5.78 | 5.26 | 4.54 | 0.44 | 5.78 | 1.56 | **75.34** | 1.28 |
| **calm** | 0.33 | 1.56 | 2.98 | 0 | 0 | 0.33 | **0** | **94.78** |

**Figure 8.** Confusion matrix obtained by the MLP on the RAVDESS database for the SI experiment

80.94%, and 70.06%, for the Emo-DB, IEMOCAP, RAVDESS, and SAVEE databases, respectively, followed by the FS method in the SI experiments. However, the MLP achieved the highest accuracies, 92.65%, 80.23%, 82.75%, and 75.38%, for the Emo-DB, IEMOCAP, RAVDESS, and SAVEE databases, respectively, followed by the FS method in the SI experiments. The confusion matrices of the results obtained for the SI experiments are shown in Figs. 7-9 to analyze the individual emotional groups' identification accuracies. The average accuracies achieved with the IEMOCAP and Emo-DB databases

| | anger | neutral | sad | happy |
|---|---|---|---|---|
| anger | **88.54** | 3.78 | 2.19 | 5.47 |
| neutral | 5.88 | **72.12** | 17.77 | 4.21 |
| sad | 1.35 | 16.99 | **77.64** | 4.00 |
| happy | 6.83 | 19.54 | 8.77 | **64.84** |

**Figure 9.** Confusion matrix obtained by the SVM on the IEMOCAP database for the SI experiment

were 78.90% and 85.73%, respectively. The RAVDESS database contains eight emotion categories, three of which, ”calm”, ”fear”, and ”anger,” were identified with accuracies of 94.78%, 91.35%, and 84.60%, respectively, by the MLP. In contrast, the other five emotions were identified with less than 90.00% accuracy, as represented in Figure 8. The MLP achieved an average accuracy with the SAVEE database of 75.38%. With the SAVEE database, ”anger,” ”neutral,” and ”sad” were recognized with accuracies of 94.22%, 90.66%, and 85.33%, respectively, by the MLP classifier. IEMOCAP achieved an average accuracy of 84.00% with the SVM, while the MLP achieved an average accuracy of 80.23%. Figure 9 shows that the average accuracy achieved by the SVM with the IEMOCAP database is 84.00%.

Four publicly available databases are used to compare the proposed method. As illustrated in Table 8, the developed system outperformed (Guo et al. (2018); Chen et al. (2018); Meng et al. (2019); Özseven (2019); Bhavan et al. (2019)) on the Emo-DB dataset for the SD experiments. The OpenSMILE package was used to extract features in (Özseven (2019)). The accuracies obtained with the SAVEE and Emo-DB databases were 72% and 84%, respectively. In comparison to (Chen et al. (2018); Meng et al. (2019); Satt et al. (2017); Zhao et al. (2018)), the proposed method performed well on the IEMOCAP database. The models in (Chen et al. (2018); Meng et al. (2019); Etienne et al. (2018)) are computationally complex and require extensive periods of training. In the proposed method, AlexNet is used for the extraction process, and the FS technique is applied. The FS approach reduced the classifier’s workload while also improving efficiency. When using the RAVDESS database, the suggested technique outperforms (Zeng et al. (2019); Bhavan et al. (2019)) in terms of accuracy.

Table 9 illustrates that the suggested approach outperforms (Meng et al. (2019); Sun and Wen (2017); Haider et al. (2020); Yi and Mak (2019); Guo et al. (2019); Badshah et al. (2017); Mustaqeem et al. (2020)) for SI experiments using the Emo-DB database. The authors extracted low-level descriptor feature emotion identification and obtained accuracies with the Emo-DB database of 82.40%, 76.90%, and 83.74%, respectively, in (Sun and Wen (2017); Haider et al. (2020); Yi and Mak (2019)). Different deep learning methods were used for SER with the Emo-DB database in (Meng et al. (2019); Guo et al. (2019); Badshah et al. (2017); Mustaqeem et al. (2020)). In comparison to other speech emotion databases, the SAVEE database is relatively small. The purpose of using a pretrained approach is that it can be trained effectively with limited data. In comparison to (Sun and Wen (2017); Haider et al. (2020)), the suggested technique provides better accuracy with the SAVEE database. When using the IEMOCAP database, the proposed methodology outperforms (Yi and Mak (2019); Guo et al. (2019); Xia and Liu (2017); Daneshfar et al. (2020); Mustaqeem et al. (2020); Meng et al. (2019)). The classification results of the proposed scheme show a significant improvement over current methods. With the RAVDESS database, the proposed approach achieved 73.50 percent accuracy.

## 6 CONCLUSIONS AND FUTURE WORK

In this research, the primary emphasis was on learning discriminative and important features from advanced emotional speech databases. Therefore, the main objective of the present research was advanced feature extraction using AlexNet. The proposed CFS approach explored the predictability of every feature. The results showed the superior performance of the proposed strategy with four datasets in both SD and SI experiments.

To analyze the classification performance of each emotional group, we display the results in the form of confusion matrices. The main benefit of applying the FS method is to reduce the abundance of features by selecting the most discriminative features and eliminating the poor features. We noticed that the pretrained AlexNet framework is very successful for feature extraction techniques that can be trained with a small number of labeled datasets. The performance in the experimental studies empowers us to explore the efficacy and impact of gender on speech signals. The proposed model is also useful for multilanguage databases for emotion classification.

In future studies, we will perform testing and training techniques using different language databases, which should be a useful evaluation of our suggested technique. We will test the proposed approach in the cloud and in an edge computing environment. We would like to evaluate different deep architectures to enhance the system's performance when using spontaneous databases.

## REFERENCES

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(10):1533–1545.

Alonso, J. B., Cabrera, J., Medina, M., and Travieso, C. M. (2015). New approach in quantification of emotional intensity from the speech signal: emotional temperature. *Expert Systems with Applications*, 42(24):9554–9564.

Alreshidi, A. and Ullah, M. (2020). Facial emotion recognition using hybrid features. *Informatics*, 7(1).

Anagnostopoulos, C.-N., Iliou, T., and Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177.

Badshah, A. M., Ahmad, J., Rahim, N., and Baik, S. W. (2017). Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 International Conference on Platform Technology and Service (PlatCon)*, pages 1–5.

Bazgir, O., Mohammadi, Z., and Habibi, S. A. H. (2018). Emotion recognition with machine learning using eeg signals. In *2018 25th National and 3rd International Iranian Conference on Biomedical Engineering (ICBME)*, pages 1–5.

Bhavan, A., Chauhan, P., Hitkul, and Shah, R. R. (2019). Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems*, 184:104886.

Campos, V., Jou, B., and i Nieto, X. G. (2017). From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction.

Chen, L., Mao, X., Xue, Y., and Cheng, L. L. (2012). Speech emotion recognition: Features and classification models. *Digital Signal Processing*, 22(6):1154–1160.

Chen, L., Mao, X., and Yan, H. (2016a). Text-independent phoneme segmentation combining egg and speech data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(6):1029–1037.

Chen, M., He, X., Yang, J., and Zhang, H. (2018). 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444.

Chen, S.-H., Wang, J.-C., Hsieh, W.-C., Chin, Y.-H., Ho, C.-W., and Wu, C.-H. (2016b). Speech emotion classification using multiple kernel gaussian process. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–4.

Chourasia, M., Haral, S., Bhatkar, S., and Kulkarni, S. (2021). Emotion recognition from speech signal using deep learning. In Hemanth, J., Bestak, R., and Chen, J. I.-Z., editors, *Intelligent Data Communication Technologies and Internet of Things*, pages 471–481, Singapore. Springer Singapore.

Christy, A., Vaithyasubramanian, S., Jesudoss, A., and Praveena, M. D. A. (2020). Multimodal speech emotion recognition and classification using convolutional neural network techniques. *International Journal of Speech Technology*, 23(2):381–388.

C.K., Y., Hariharan, M., Ngadiran, R., Adom, A. H., Yaacob, S., Berkai, C., and Polat, K. (2017). A new hybrid pso assisted biogeography-based optimization for emotion and stress recognition from speech signal. *Expert Systems with Applications*, 69:149–158.

Costanzi, M., Cianfanelli, B., Saraulli, D., Lasaponara, S., Doricchi, F., Cestari, V., and Rossi-Arnaud, C. (2019). The effect of emotional valence and arousal on visuo-spatial working memory: Incidental emotional learning and memory for object-location. *Frontiers in psychology*, 10:2587–2587. 31803120[pmid].

Daneshfar, F., Kabudian, S. J., and Neekabadi, A. (2020). Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and gaussian elliptical basis function network classifier. *Applied Acoustics*, 166:107360.

Demircan, S. and Kahramanli, H. (2014). Feature extraction from speech data for emotion recognition. *Journal of Advances in Computer Networks*, 2:28–30.

Deng, J., Zhang, Z., Marchi, E., and Schuller, B. (2013). Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 511–516.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L., and Schmauch, B. (2018). Cnn+lstm architecture for speech emotion recognition with data augmentation. *Workshop on Speech, Music and Mind 2018*.

Fayek, H. M., Lech, M., and Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68. Advances in Cognitive Engineering Using Neural Networks.

Gu, Y., Chen, S., and Marsic, I. (2018). Deep mul timodal learning for emotion recognition in spoken language. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Guo, L., Wang, L., Dang, J., Liu, Z., and Guan, H. (2019). Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine. *IEEE Access*, 7:75798–75809.

Guo, L., Wang, L., Dang, J., Zhang, L., Guan, H., and Li, X. (2018). Speech emotion recognition by combining amplitude and phase information using convolutional neural network. In *Proc. Interspeech 2018*, pages 1611–1615.

Haider, F., Pollak, S., Albert, P., and Luz, S. (2020). Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods.

Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Hossain, M. S. and Muhammad, G. (2019). Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78.

Hossain.M. Shamim, M. (2014). Cloud-based collaborative media service framework for healthcare. *International Journal of Distributed Sensor Networks*, 10(3):858712.

Kandali, A. B., Routray, A., and Basu, T. K. (2009). Vocal emotion recognition in five native languages of assam using new wavelet features. *International Journal of Speech Technology*, 12(1):1.

Kapoor, P. and Thakur, N. (2021). Emotion recognition using q-knn: A faster knn approach. In Gupta, D., Khanna, A., Bhattacharyya, S., Hassanien, A. E., Anand, S., and Jaiswal, A., editors, *International Conference on Innovative Computing and Communications*, pages 759–768, Singapore. Springer Singapore.

Kerkeni, L., Serrestou, Y., Raoof, K., Mbarki, M., Mahjoub, M. A., and Cleder, C. (2019). Automatic speech emotion recognition using an optimal combination of features based on emd-tkeo. *Speech Communication*, 114:22–35.

Khan, L., Amjad, A., Ashraf, N., Chang, H.-T., and Gelbukh, A. (2021). Urdu sentiment analysis with

deep learning methods. *IEEE Access*, 9:97803–97812.
Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90.
Kurpukdee, N., Koriyama, T., Kobayashi, T., Kasuriya, S., Wutiwiwatchai, C., and Lamsrichan, P. (2017). Speech emotion recognition using convolutional long short-term memory neural network and support vector machines. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1744–1749.
Le, D. and Provost, E. M. (2013). Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 216–221.
Lech, M., Stolar, M., Best, C., and Bolia, R. (2020). Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. *Frontiers in Computer Science*, 2:14.
Li, S., Wang, L., Li, J., and Yao, Y. (2021). Image classification algorithm based on improved AlexNet. *Journal of Physics: Conference Series*, 1813(1):012051.
Liu, J., Wu, G., Luo, Y., Qiu, S., Yang, S., Li, W., and Bi, Y. (2020). Eeg-based emotion classification using a deep neural network and sparse autoencoder. *Frontiers in Systems Neuroscience*, 14:43.
Mao, S., Tao, D., Zhang, G., Ching, P. C., and Lee, T. (2019). Revisiting hidden markov models for speech emotion recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6715–6719.
McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python.
Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.
Meng, H., Yan, T., Yuan, F., and Wei, H. (2019). Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access*, 7:125868–125881.
Milton, A., Roy, S. S., and Selvi, S. (2013). Svm scheme for speech emotion recognition using mfcc feature. *International Journal of Computer Applications*, 69:34–39.
Mustaqeem, Sajjad, M., and Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE Access*, 8:79861–79875.
Nalini, N. and Palanivel, S. (2016). Music emotion recognition: The combined evidence of mfcc and residual phase. *Egyptian Informatics Journal*, 17(1):1–10.
Niu, Y., Zou, D., Niu, Y., He, Z., and Tan, H. (2017). A breakthrough in speech emotion recognition using deep retinal convolution neural networks. *CoRR*, abs/1707.09917.
Noroozi, F., Sapiński, T., Kamińska, D., and Anbarjafari, G. (2017). Vocal-based emotion recognition using random forests and decision tree. *International Journal of Speech Technology*, 20(2):239–246.
Patel, P., Chaudhari, A., Pund, M. A., and Deshmukh, D. (2017). Speech emotion recognition system using gaussian mixture model and improvement proposed via boosted gmm. *IRA-International Journal of Technology & Engineering*, 7:56–64.
Poon-Feng, K., Huang, D.-Y., Dong, M., and Li, H. (2014). Acoustic emotion recognition based on fusion of multiple feature-dependent deep boltzmann machines. In *The 9th International Symposium on Chinese Spoken Language Processing*, pages 584–588.
Qing, C., Qiao, R., Xu, X., and Cheng, Y. (2019). Interpretable emotion recognition using eeg signals. *IEEE Access*, 7:94160–94170.
Rao, K. S., Koolagudi, S. G., and Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology*, 16(2):143–160.
Rayaluru, A., Bandela, S. R., and T, K. K. (2019). Speech emotion recognition using feature selection with adaptive structure learning. In *2019 IEEE International Symposium on Smart Electronic Systems (iSES) (Formerly iNiS)*, pages 233–236.
Ren, S., He, K., Girshick, R., Zhang, X., and Sun, J. (2016). Object detection networks on convolutional feature maps.
Sailunaz, K., Dhaliwal, M., Rokne, J., and Alhajj, R. (2018). Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):28.
Satt, A., Rozenberg, S., and Hoory, R. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. In *Proc. Interspeech 2017*, pages 1089–1093.

Schmidt, E. M. and Kim, Y. E. (2011). Learning emotion-based acoustic features with deep belief networks. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 65–68.

Sezgin, M. C., Gunsel, B., and Kurt, G. K. (2012). Perceptual audio features for emotion detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2012(1):16.

Shi, P. (2018). Speech emotion recognition based on deep belief network. In *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, pages 1–5.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition.

Stolar, M. N., Lech, M., Bolia, R. S., and Skinner, M. (2017). Real time speech emotion recognition using rgb image classification and transfer learning. In *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–8.

Suhaimi, N. S., Mountstephens, J., and Teo, J. (2020). Eeg-based emotion recognition: A state-of-the-art review of current trends and opportunities. *Computational Intelligence and Neuroscience*, 2020:8875426.

Sun, Y. and Wen, G. (2017). Ensemble softmax regression model for speech emotion recognition. *Multimedia Tools and Applications*, 76(6):8305–8328.

Sun, Y., Wen, G., and Wang, J. (2015). Weighted spectral features based on local hu moments for speech emotion recognition. *Biomedical Signal Processing and Control*, 18:80–90.

Tao, J., Liu, F., Zhang, M., and Jia, H. (2008). Design of speech corpus for mandarin text to speech.

Trentin, E., Scherer, S., and Schwenker, F. (2015). Emotion recognition from speech signals via a probabilistic echo-state network. *Pattern Recognition Letters*, 66:4–12. Pattern Recognition in Human Computer Interaction.

Ververidis, D. and Kotropoulos, C. (2005). Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm. In *2005 IEEE International Conference on Multimedia and Expo*, pages 1500–1503.

Waghmare, V. B., Deshmukh, R., Shrishrimal, P., and Janvale, G. (2014). Development of isolated marathi words emotional speech database. *International Journal of Computer Applications*, 94:19–22.

Wang, Y. and Guan, L. (2008). Recognizing human emotional state from audiovisual signals*. *IEEE Transactions on Multimedia*, 10(5):936–946.

Wosiak, A. and Zakrzewska, D. (2018). Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis. *Complexity*, 2018:2520706.

Xia, R. and Liu, Y. (2017). A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Transactions on Affective Computing*, 8(1):3–14.

Yi, L. and Mak, M.-W. (2019). Adversarial data augmentation network for speech emotion recognition. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 529–534.

Zeng, Y., Mao, H., Peng, D., and Yi, Z. (2019). Spectrogram based multi-task audio classification. *Multimedia Tools and Applications*, 78(3):3705–3722.

Zhang, S., Zhang, S., Huang, T., Gao, W., and Tian, Q. (2018). Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3030–3043.

Zhang, W., Zhao, D., Chai, Z., Yang, L. T., Liu, X., Gong, F., and Yang, S. (2017). Deep learning and svm-based emotion recognition from chinese speech for smart affective services. *Softw. Pract. Exper.*, 47(8):1127–1138.

Zhao, Z., Zheng, Y., Zhang, Z., Wang, H., Zhao, Y., and Li, C. (2018). Exploring spatio-temporal representations by integrating attention-based bidirectional-lstm-rnns and fcns for speech emotion recognition. In *Proc. Interspeech 2018*, pages 272–276.

Zheng, C., Wang, C., and Jia, N. (2020). An ensemble model for multi-level speech emotion recognition. *Applied Sciences*, 10(1).

Özseven, T. (2019). A novel feature selection method for speech emotion recognition. *Applied Acoustics*, 146:320–326.