Effect on speech emotion classification of feature selection approach using 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 enough 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 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 were performed by utilizing four different publicly accessible databases. Our method obtained 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 Feature Selection Approach Using 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 enough 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 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 were performed by utilizing four different publicly accessible databases. Our method obtained 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 behind naturally interacting with machines because machines can neither understand our emotional state nor our emotional behavior. In previous studies, some modalities have been founded for identifying the emotional state, such as extended text, speech (El Ayadi et al. (2011)), video (Hossain and Muhammad (2019)) , facial expression (Alreshidi and Ullah (2020)), short messages (Sailunaz et al. (2018)), and physiological signals (Qing et al. (2019)). These modalities vary across the applications. The most common modalities in social media are emoticons and short text; video is the most common modality for the gaming system. 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 for a successful SER framework: (i) selecting an excellent emotional database, (ii) useful feature extraction, and (iii) using deep learning algorithms to design accurate classifiers. However, emotional feature extraction is a significant problem in n 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 (MFCC), zero-crossing rate (ZCR), formant frequency, filter bank energy (FBR), linear prediction cepstrum coefficients (LPCC), modulation spectral features (MSFs) and Mel-spectrograms. In (Sezgin et al. (2012)), several distinguishing acoustic features were used to identify the emotion: 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 neighbor (RNN)(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. Those provide for learning the algorithm by combining two functions: the redial base 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 estimates that the combined features have 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 founded in SER to determine the high-level emotion features from utterances to establish the hierarchical representation of speech. The accuracy of handcrafted features is relatively high, and this feature extraction technique always consumes 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 appropriate features that are emotionally powerful must be selected by performing effectively for SER.

Therefore, it is more important to select specific speech features that are not affected by country, speaking style of speaker, culture, and region. Therefore, feature selection (FS) is also essential after extraction and is accompanied by an appropriate classifier to recognize emotions from speech. The summary for FS 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 feature, we follow a convolutional neural network (CNN) (AlexNet). 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 CNN inputs 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 the feature extraction technique 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 extensively used to perform different classification tasks. Additionally, these deep learning models regularly perform extremely better than other shallow CNNs. The main reason is that deep CNNs extract mid-level features from 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 the 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 across the existing handcrafted and deep-learning methods for SD and SI experiments.

The remaining 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)), k-nearest neighbors (KNN) (Kapoor and Thakur (2021)), HMM (Mao et al. (2019)), CNNs (Christy et al. (2020)), Gaussian mixture models (GMM) (Patel et al. (2017)), and MLP. These algorithms have been commonly used to identify emotions. Emotions are categorized into 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 and more than one dimension. Pleasure-arousal-dominance (PAD) is a three-dimensional emotional state model proposed by (Mehrabian (1996)). Different features are essential to identify speech emotion from voice. Spectral features are significant and widely used to classify emotions. AB Kandali et al. introduced an approach to classify emotion-founded MFCCs as the main features and applied the 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 emotion (Demircan and Kahramanli (2014)). SVM, a radial basis function neural network (RBFNN), and an autoassociative neural network (ANNN) have been used to recognize emotions after combining two features, MFCCs and residual phase (RP), from a music database (Nalini and Palanivel (2016)). Chenchah, Farah et al. implemented a SVM and HMM to classify speech emotions after extracting the spectral features from speech signals (Chourasia et al. (2021)). In (C.K. et al. (2017)), particle swarm optimization-based features and high-order statistical features were utilized. The Berlin emotional speech database (EMO-DB) in the experiment and accuracy obtained was between 90% and 99.5%. Paralinguistic features and prosodic features were utilized to detect emotion from speech in (Alonso et al. (2015)). Hossain et al. proposed a cloud-based collaborative media system that uses emotions from speech signals and used standard features such as MFCCs (Hossain.M. Shamim (2014)).

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. In (Poon-Feng et al. (2014)), a generalized discriminant analysis (Gerda) was presented by several Boltzmann machines to analyze and classify the emotions from speech and improve the past reported baseline by traditional approaches. Erik M. Schmidt et al. proposed a regression-based DBN to recognize music emotion and a model based on three hidden layers to learn the emotional features (Han et al. (2014)). Duc Le et al. implemented hybrid classifiers, which were the set of DBNs, HMMs, and attained results on FAU Aibo (Le and Provost (2013)). DNNs were used to generate emotional probabilities into segments [48], 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%.

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)). Accuracy was obtained at approximately 95% using the EMO_DB database. Schmidt et al. used an approach based on linear regression and deep belief networks to identify musical emotion (Schmidt and Kim (2011)). They used the music MoodSwings Lite database and obtained a 5.41% error rate. SVM and DBN were examined utilizing the Chinese academic database (Zhang et al. (2017)). The accuracy using DBNs was 94.5%, and the accuracy of SVM was approximately 85% . In (Fayek et al. (2017)), the authors suggested deep learning approaches. A speech signal spectrogram was used as an IEMOCAP database input. A decision tree was used to identify the emotion from the CASIA

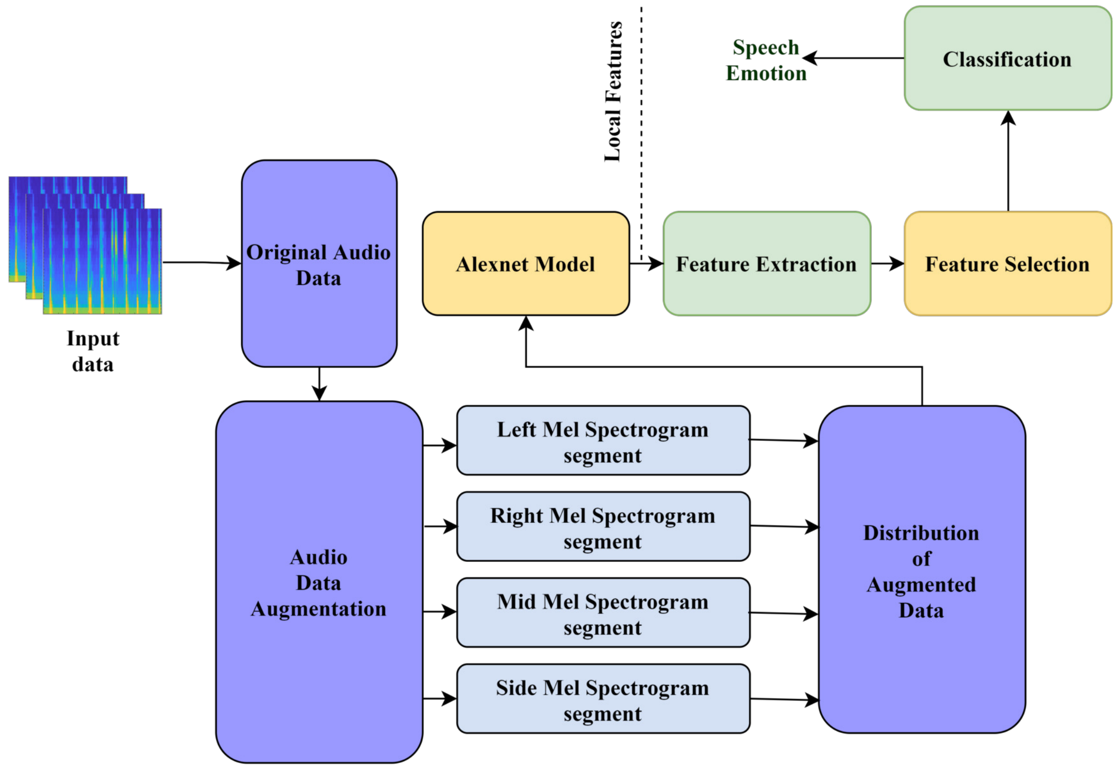

Figure 1. The structure of our proposed model for audio emotion recognition

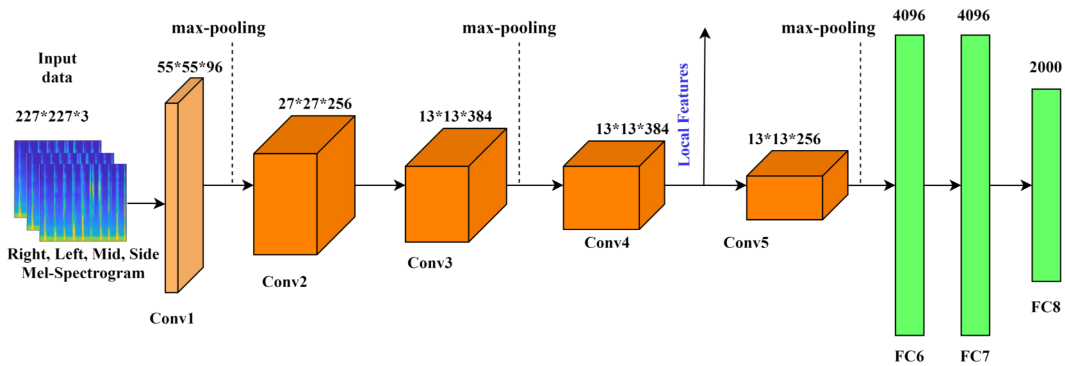

Figure 2. The general architecture of the AlexNet

Chinese emotion corpus in (Tao et al. (2008)) and achieved 89.6% accuracy. 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 performance in recognizing emotions from speech signals. The presented approach obtained the highest accuracy of 99.25% in the IEMOCAP database. 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 Emo-DB and IEMOCAP databases. The attention process was used to develop a dilated CNN and BiLSTM in (Meng et al. (2019)). To identify the 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 obtained accuracy with the Emo-DB database was 84% and 72% with the SAVEE database. Satt A, and S. Rozenberg et al. suggested another efficient convolutional LSTM approach for emotion classification. The introduced model learned spatial patterns and learned

Table 1. Nomenclature

ACRNN	Attention Convolutional Recurrent Neural Network	KNN	K-Nearest Neighbor
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	Multi-layer 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	Redial Base 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 Mixtures 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

spatial spectrogram patterns representing information on the emotional states (Satt et al. (2017)). In (Zhao et al. (2018)), the suggested approach used integrated attention-based 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)) contained a data augmentation technique. The experiment was performed on the IEMOCAP database with four emotions. In (Zeng et al. (2019)), a multitask approach was used for audio analysis. The two different databases were used to extract prosodic and spectral features with an ensemble softmax regression approach (Sun and Wen (2017)). The two different acoustic paralinguistic features set was used in (Haider et al. (2020)). An adversarial data augmentation network was presented in (Yi and Mak (2019)) to create simulated samples to resolve the data scarcity problem. Kernel extreme learning machine (KELM) features were introduced in (Guo et al. (2019)). For the identification of emotional groups, experiments are 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.Energy and pitch were extracted from each audio segment in (Daneshfar et al. (2020); Ververidis and Kotropoulos (2005); Rao et al. (2013)). However, computational costs and a large amount of data are required for deep learning techniques. The majority of current speech emotional databases have a small amount of data. The 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 pre-trained 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. 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. Pretrained networks have many benefits,

Table 2. Detailed description of 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 Recording

including the ability to reduce training time and improve accuracy. They also need fewer training data and deal directly with dynamic variables.

3 PROPOSED METHOD

This section describes the proposed pretrained CNN (AlexNet) algorithm for the SER framework. AlexNet [25] is a pretrained model on the extensive-scale 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 extracted 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 audio input data, (b) low-level feature extraction using AlexNet, (c) feature selection, and (d) classification. In the following, we explain all four steps of our model in detail.

3.1 Creation of 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 1D original audio speech dataset. Then, the generated segments are converted into fixed-size $227 \times 227 \times 3$ input 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 10 ms overlapping. 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 inputs of AlexNet, and these channels are identical to the color RGB image. Therefore, we resize the $64 \times 64 \times 3$ original spectrogram to the new size $227 \times 227 \times 3$. In this case, we can create four (mid, 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 only selects 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 Features 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 as follows.

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 the local regions and form various feature groups. The 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 $3 \times 3 \times 256$. The CL4 contains 384 kernels of size $3 \times 3 \times 192$. For the output value of each CL, the ReLU function is used, which hastens the training process.

3.3.3 Pooling Layer (PL)

After the CLs, the 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 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 the classification. The output of CLs and PLs is given to the fully connected layers. There are tree 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 at recognizing and generalizing the original image pixels. CL4 extracts relevant features from the original pixel values by saving 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 the 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 so that it takes the least time for training, enhances 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 not redundant to any other class features. For this reason, we used entropy-based information theory. The equation of entropy-based information theory defined as:

$$F(E) = -\Sigma S(e_j) log2(S(e_j)) \tag{1}$$

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

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

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

$$I(E/G) = (F(E) - F(E/G)) \tag{3}$$

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

$$SU(E,G) = 2[I(E/G)/F(E) + F(G)] \tag{4}$$

SU balances the information gain's bias toward features with more values by normalizing its value with a range of [0,1]. SU analyzes a pair of features symmetrically. Entropy-based techniques need nominal features. These features can be used to evaluate correlations between continuous features if these features are discretized properly 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 the 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. Extracted features are fed into classification algorithms. CFS usually delete (backward selection) or add (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 for speech emotion recognition.

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 the North American Language. RAVDESS is 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 is recorded in the British English language from 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 is recorded in the German language from five male and five female native 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, neutral, frustration, and others. The IEMOCAP database is recorded by five male and five female professional actors. In this work, we use four (neutral, angry, sadness, happiness) class labels.

4.2 Experimental Setup

In Python language frameworks, all the experiments are completed with version 3.9.0. 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 five different models. One of these is with the max-pooling layer, and the other is without the max-pooling layer. The feature extraction technique is also involved in the proposed method.

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 SD experiments. We use ten-fold cross-validation in our studies. All databases are randomly divided into ten equal complementary subsets with a dividing ratio of 80:20 to train and test the model. Table 3 gives the results achieved from 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 of 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 the Emo-DB.

Table 3. 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 Emo-DB dataset. The FS method obtained 170,465,277 feature maps for the SAVEE, RAVDESS, and IEMOCAP datasets.

The experimental results illustrate significant accuracy improvement by using data resampling and the FS approach. We address the standard deviation and average weighted recall to evaluate the performance and stability of 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 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 of 92.45%, 88.34% with the Emo-DB and RAVDEES datasets. The RF classifier reported the highest accuracy of 93.51% on the Emo-DB datasets

Table 4. 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 5. SI experiments results without FS approach

	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 6. SI experiments results with FS approach

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 7. Comparison of 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	(Chen et al. (2018))	3-D ACRNN	64.74
IEMOCAP	(Meng et al. (2019))	Dilated CNN + BiLSTM	74.96
IEMOCAP	(Satt et al. (2017))	3 Convolution Layers + LSTM	68.00
IEMOCAP	(Zhao et al. (2018))	Attention-BLSTM-FCN	64.00
IEMOCAP	(Etienne et al. (2018))	CNN+LSTM	64.50
IEMOCAP	**Proposed Approach**	**AlexNet+FS+MLP**	**89.12**
IEMOCAP	**Proposed Approach**	**AlexNet+FS+RF**	**86.23**

Table 8. 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	(Sun and Wen (2017))	Ensemble soft-MarginSoftmax (EM-Softmax)	82.40
Emo-DB	(Haider et al. (2020))	eGeMAPs and emobase	76.90
Emo-DB	(Yi and Mak (2019))	OpenSmile Features + ADAN	83.74
Emo-DB	(Mustaqeem et al. (2020))	Redial Based Function Network(RBFN) + Deep BiLSTM	85.57
Emo-DB	(Guo et al. (2019))	Statistical Features and Empirical Features+ KELM	84.49
Emo-DB	(Badshah et al. (2017))	DCNN + DTPM	87.31
Emo-DB	(Meng et al. (2019))	Dilated CNN+ BiLSTM	85.39
Emo-DB	**Proposed Approach**	**AlexNet+FS+MLP**	**92.65**
Emo-DB	**Proposed Approach**	**AlexNet+FS+SVM**	**90.78**
IEMOCAP	(Mustaqeem et al. (2020))	Redial Based Function Network(RBFN) + Deep BiLSTM	72.2
IEMOCAP	Guo et al. (2019)	Statistical Features and Empirical Features+ KELM	57.10
IEMOCAP	(Yi and Mak (2019))	OpenSmile Features + ADAN	65.01
IEMOCAP	(Chen et al. (2018))	Dilated CNN+ BiLSTM	69.32
IEMOCAP	(Xia and Liu (2017))	SP + CNN	64.00
IEMOCAP	(Daneshfar et al. (2020))	IS10 + DBN	64.50
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 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 SVM on the SAVEE database for the SD experiment

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 4 shows that 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 of 88.33% and 91.66% with the SAVEE dataset. As shown in Figure 6, the RAVDESS dataset contains eight emotions, ”anger”, ”calm”, ”fear”, and ”neutral”, which are listed with accuracies of

	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 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 MLP on the IEMOCAP database for the SD experiment

96.32%, 97.65%, 95.54%, and 99.98%, respectively. The IEMOCAP database identified ”anger” with a highest accuracy of 93.23%, while ”happy,” ”sad,” and ”neutral” were recognized with highest accuracies of 83.41%, 91.45%, and 89.65% with the MLP classifier, respectively.

	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 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 MLP on the RAVDESS database for the SI experiment

5.0.2 Speaker-Independent (SI) Experiments

We adopted the single speaker out (SSO) method for SI experiments. One annotator was used for testing, and all other annotators were used for training. In the proposed approach, the IEMOCAP dataset was split into testing and training sessions. By switching all of the testing annotators, the process was repeated, and the average accuracy was achieved from every testing speaker. Table 6 lists the identification results of five classification models for SI experiments without the FS technique. The MLP obtained the highest

	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 SVM on the IEMOCAP database for the SI experiment

accuracy of 88.32% with the Emo-DB dataset. With the SAVE database, MLP obtained the highest accuracy of 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 accuracy of 79.45% and 65.78% with Emo-DB and RAVDESS, respectively. Table 6 represents the outcomes for SI experiments of the feature extraction approach with data resampling and the FS method. The FS and data resampling approach improved the accuracy, according to preliminary results.

We reported the average weighted recall and standard deviation to evaluate an SI experiment's performance and stability utilizing the FS method. The SVM obtained the highest accuracies of 90.78%, 84.00%, 80.94%, and 70.06% for the Emo-DB, IEMOCAP, RAVDESS, and SAVEE databases, respectively, followed by the FS method for SI experiments. However, the MLP achieved the highest accuracies of 92.65%, 80.23%, 82.75%, and 75.38% for the Emo-DB, IEMOCAP, RAVDESS, and SAVEE databases, respectively, followed by the FS method for SI experiments. The confusion matrices of the results obtained for SI experiments are shown in Figs. 7–9 to analyze the individual emotional groups' identification accuracy. The average accuracy achieved with the IEMOCAP and Emo-DB databases is 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 average accuracy with the SAVEE database 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 SVM, while MLP achieved an average accuracy of 80.23%. Figure 9 shows that the average accuracy achieved by SVM with the IEMOCAP database is 84.00%.

Four publicly available databases are used to compare the proposed research. As illustrated in Table 7, 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 SD experiments. The OpenSMILE package was used to extract features in (Özseven (2019)). The obtained accuracies 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 8 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 while 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 with the Emo-DB database were used for SER 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 a limited data. In comparison to (Sun and Wen (2017); Haider et al. (2020)), the suggested technique provides a 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 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 speech emotional databases. Therefore, the main objective of the present research is advanced feature extraction using AlexNet. The proposed CFS approach explored the predictability of every feature. The results showed superior performance of the proposed strategy with four datasets for both SD and SI experiments.

To analyze the classification performance of each emotional group, we displayed 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 discriminant features and eliminating the other poor features. We noticed that the pretrained AlexNet framework is very successful for feature extraction techniques that can be trained with a short number of labeled datasets. The performance of 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 an edge computing environment. We would like to evaluate different deep architectures to enhance the system's performance while 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.

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.

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.

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.

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.

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.

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.

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.