Complex & Intelligent Systems (2022) 8:4237–4249
https://doi.org/10.1007/s40747-022-00713-w

ORIGINAL ARTICLE

A DCRNN-based ensemble classifier for speech emotion recognition in Odia language

Monorama Swain[1] · Bubai Maji[1] · P. Kabisatpathy[2] · Aurobinda Routray[3]

Received: 23 June 2021 / Accepted: 5 March 2022 / Published online: 24 March 2022

Abstract
The Odia language is an old Eastern Indo-Aryan language, spoken by 46.8 million people across India. We have designed an ensemble classifier using Deep Convolutional Recurrent Neural Network for Speech Emotion Recognition (SER). This study presents a new approach for SER tasks motivated by recent research on speech emotion recognition. Initially, we extract utterance-level log Mel-spectrograms and their first and second derivative (Static, Delta, and Delta-delta), represented as 3-D log Mel-spectrograms. We utilize deep convolutional neural networks deep convolutional neural networks to extract the deep features from 3-D log Mel-spectrograms. Then a bi-directional-gated recurrent unit network is applied to express long-term temporal dependency out of all features to produce utterance-level emotion. Finally, we use ensemble classifiers using Softmax and Support Vector Machine classifier to improve the final recognition rate. In this way, our proposed framework is trained and tested on Odia (Seven emotional states) and RAVDESS (Eight emotional states) dataset. The experimental results reveal that an ensemble classifier performs better instead of a single classifier. The accuracy levels reached are 85.31% and 77.54%, outperforming some state-of-the-art frameworks on the Odia and RAVDESS datasets.

Keywords Speech emotion recognition · Deep convolutional neural network · Bi-directional gated recurrent unit · Ensemble classifier

Introduction

In recent years, with the rapid growth in the field of artificial intelligence such as voiceprint, fingerprint, speech emotion recognition, face recognition, and other biometrics systems has attracted more attention by the many researchers [1–3]. With further developments in the processing capability of a computer and the increasing demand for pattern recognition and speech emotion recognition, both of these have been vastly used in the interaction between human-robotics [4, 5, 5, 6, 8]. The information from speech signals carries people's emotional and most natural communication in day-to-day conversations and works. It consists of paralinguistic and linguistic information. Linguistic contains language and contextual information, and paralinguistic gives information related to the emotional state of the speech [7].

Building an SER system is a challenging task. Firstly, the unavailability of speech datasets in different languages is time-consuming work to create a proper speech emotion database. Secondly, the different dataset has built of other regions of the world with their diverse cultural, languages, and speakers with their different speaking styles [8]. Consequently, all of the above variations create difficulties in detecting the emotional state from the speech signal. In addition, recognition of speech emotional systems is independent of hardware equipment. The automobile industry can also have advantages from SER for the many real-time emotion detection tasks. The various techniques have been utilized in

✉ Monorama Swain
mswain@silicon.ac.in

Bubai Maji
bubaimaji51@gmail.com

P. Kabisatpathy
pkabisatpathy@gmail.com

Aurobinda Routray
aroutray@ee.iitkgp.ernet.in

[1] Department of Electronics and Communication Engineering, Silicon Institute of Technology, Bhubaneswar, India

[2] Department of Electronics and Instrumentation, CV Raman College of Engineering, Bhubaneswar, India

[3] Department of Electrical Engineering, Indian Institute of Technology, Kharagpur, India

the pre-processing, feature extraction process, and classification algorithms using several SER datasets [9]. However, several speech-emotion classifier systems and different type features are combined in the literature.

The recognition of the speech emotion system is mainly divided into three sections: speech pre-processing, feature extraction, and classifier model [12]. A robust classification model identifies discriminative emotional feature information as an essential factor in the emotion recognition system [10]. The feature extraction process is the initial step, and many hand-crafted features have been used for SER [11, 14]. In recent years, spectral features have been used more often than hand-crafted features because spectral features can process more high-level emotional information. Due to this advantage, the spectral features give an efficient result compared to other types of features [12, 17]. However, the low-level features are unable to detect the actual emotional state in an utterance. Although, a significant drawback of the SER method is the problems of the feature extraction process because during the process one may lose some important information. So how can we extract as much abundant emotional information from each utterance and train the proper model? That would be the first problem we need to solve. To minimize this issue we are modifying the deep learning method. We employ an ensemble classifier (using Softmax and SVM classifier) based deep learning method. This deep learning method provides a possible solution for the above problem of the feature extraction process for SER. Deep neural network (DNN) is the most common and popular deep learning method, which can extract discriminative features and has shown excellent performance in classification tasks. It has been demonstrated that compared with traditional deep learning methods, DNN achieves better performance.

However, the Gaussian mixture model (GMM) has a problem with the limited training of speech data. On the other hand, the support vector machine (SVM) performs better for recognition tasks than the other classifiers, with limited training data. However, the SVM model does not learn spectral features directly due to spectral features extracted from variable lengths of speech samples [13]. The convolutional neural networks (CNNs) and Recurrent Neural Networks (RNNs) are the two standard deep learning models [14, 15]. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are two basic RNNs which can easily handle time-series data. Out of two the RNNs, LSTM executes with high error rates; however, optimization of GRU is faster with close to LSTM error rates. Basically, CNN is appropriate for data processing of images and realizes the local pattern of data viewing. One or two-layer CNNs performed poorer than the DCNN models [16].

Zhang et al. [8] have also found that 2-D convolution performs better and is the first choice over 1-D convolution for limited data. Accordingly, they may not learn discriminative features well to determine their emotional state. CNN also has some problems because CNN learns high-level features from Low-Level Descriptors (LLDs). The LLDs features are insufficient to extract emotional classification in complex scenarios [8]. Then researchers began to use images like two-dimensional spectrograms as shown in Fig. 1, to extract the right and flexible emotional information relevant to the SER. The horizontal axis defines the information in the time domain. The vertical axis depicts the information in the frequency domain that holds the important information relevant to emotion, and makes it a decent SER system. Due to these advantages, we adopt a deep convolutional neural network that can automatically extract the most emotionally relevant information from the audio sample's spectrogram. Zeng et al. [17] employed deep neural network-based gated Residual Networks (GResNets) and extracted the emotional feature from generated spectrograms on RAVDESS [23] dataset; the accuracy achieved was 65.97%. Badshah et al. [18] also used the DCNN model to extract speech emotion features from spectrograms. Abdel-Hamid et al. [14] implemented CNN-based deep learning model and applied log Mel-spectrograms as an input.

The success of DCNNs motivates us to use DCNNs in the speech emotion recognition field. In this paper, we report a new approach using DCNN and Bi-directional Gated Recurrent Unit with ensemble classifiers (Combine of Softmax and SVM classifiers) as displayed in Fig. 2. First, we extract log Mel-spectrograms and its first and second derivatives with respect to time (static, delta, and delta-delta). Then, we pass all the Mel-spectrogram through a pre-trained DCNN model to extract deep features. In this experiment, we use AlexNet [19], a pre-trained DCNN. After that, all the deep features are applied sequentially as the input of the Bi-GRU model. The Bi-GRU can capture the time–frequency relationship of utterance-level features and extract high-level utterance-level features. Finally, we adopt ensemble classifiers for emotion classification. This experiment was carried out on the Odia database on seven emotional classes and RAVDESS [23] on eight emotional levels of speech signal. Our experimental work reveals that the proposed approach outperforms some previously published results. The main contributions of our works are as follows:

(1) First, we designed a proper speech representation with a DCRNN network using an images such as 3-D log Mel-spectrograms to capture the details of temporal-frequency correlations and assembles a more potent feature learning model.
(2) Secondly, the correlated highest prediction probabilities value in the final prediction vector can confuse the classifier to identify the actual emotional state. Here, we employ an ensemble classifier that can only detect

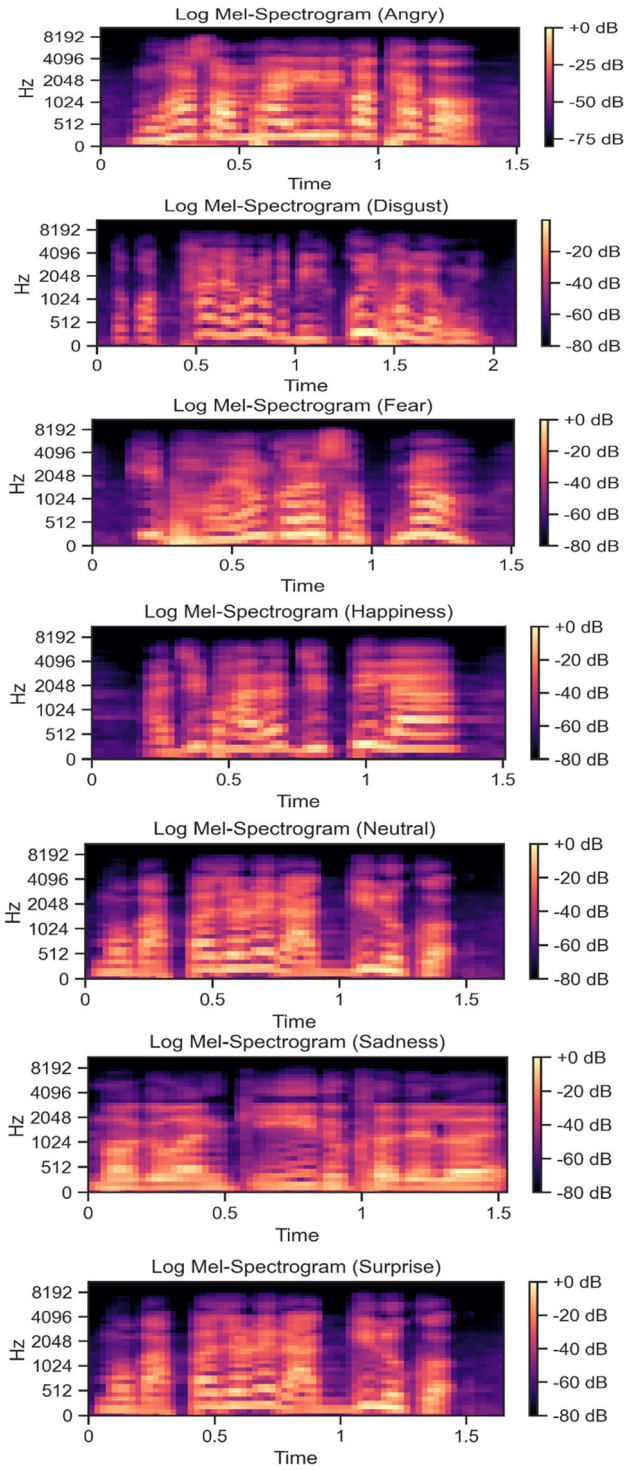

Fig. 1 Log Mel-spectrograms of the seven emotional states in the Odia database

the maximum probability vector between two final prediction vectors of two classifiers, which leads to better performance than a single classifier for speech emotion recognition.

(3) The DCRNN with ensemble classifier model produces an actual prediction vector from confusing two or more than two prediction vectors using discriminative utterance-level features.

The remainder of this paper is formatted as follows. The related works are represented in Section 2. The details of the DCRNN network and classifiers is described in Section 3. Section 4 shows the details of our experimental results, followed by Section 5 with our conclusions.

Related work

The framework of the speech emotion recognition system contains two basic components. The first component is the extraction of the speech features, and the next part is the classifier selection that identified the emotional state from utterances. We discuss in detail the emotion classification strategies followed by the feature extraction process.

Emotional classifier

Classifier plays an essential role for the SER system. Researchers have proposed various deep learning algorithms to represent an efficient classifier to distinguish emotional classes. Some popular emotion classifiers are the K-Nearest Neighbor (KNN) algorithm [20], Hidden Markov Models (HMMs) [21], Gaussian Mixture Models (GMMs) [22], SVM) [23], and Softmax function [24]. In the above classifiers, if the training data is much greater than the number of features ($p > > q$), KNN is better, but for lesser training data, SVM outperforms KNN and GMM. Currently, most researchers use Softmax and SVM classifiers rather than KNN and GMM classifiers, which makes them more popular and reliable. Softmax and SVM are the most useable classifiers in speech-related tasks and the performance difference is usually very small. But sometimes, selected features are not robust enough to design a classifier for speech emotion recognition. So classifiers are trained and tested on the same data. To integrate the merits of classifiers, we ensemble the SVM and Softmax classifiers and evaluate the performances in terms of accuracy for speech emotion recognition.

Feature extraction

The feature extraction process is a primary task for building an SER system. This process can reduce noise from raw audio data and generate highly effective features in learning the emotions from the SER model. Various features were used for SER systems, such as acoustic features, context information, hybrid features, and linguistic features. Among these, acoustic features are mainly used in the emotion recognition domain, containing local and global features [16]. Acoustic features are separated into four groups: spectral features,

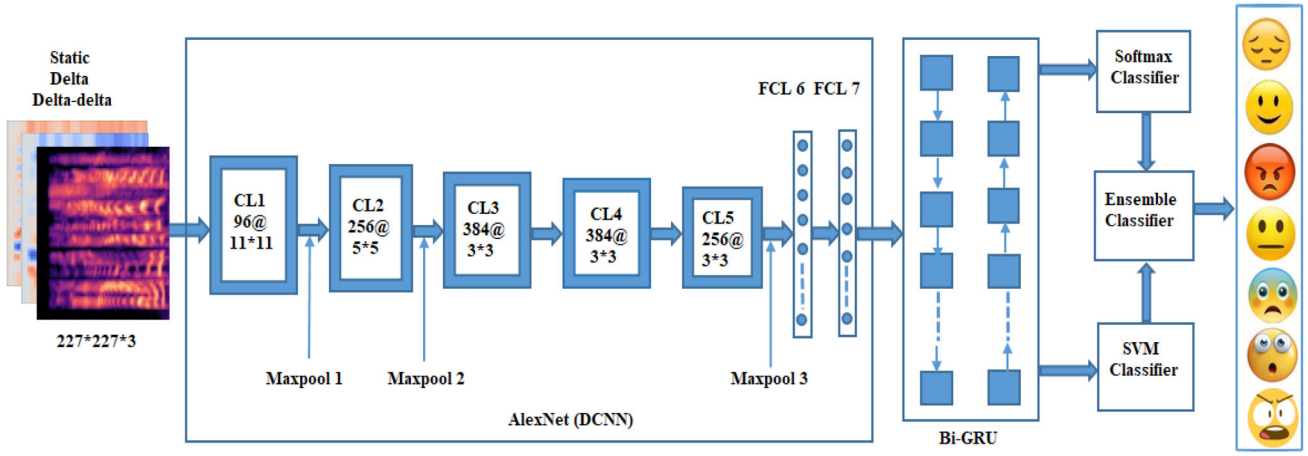

Fig. 2 The framework of our proposed model

prosodic features, voice quality features, and other features. Above all, prosodic and spectral features are used more commonly in SER systems. [31] utilized 79 emotional-related prosodic features. Kachele et al. [25] integrated prosodic features and speech quality information to identify more robust speech emotion systems than prosodic features. To obtain a better result, many traditional linear spectral correlation features have been investigated, such as Zero Crossing Rate (ZCR), Chromagram, Linear Prediction Coefficients (LPC), Root-Means-Square (RMS), Mel-Frequency Cepstrum Coefficients (MFCC), Log-Frequency Power Coefficient (LFPC), Log Mel-spectrogram, and Linear Prediction Cepstrum Coefficients (LPCC) [14, 26], and energy-related features. Prosody features such as Pitch, formant frequency, duration, and loudness were also commonly used [27]. Some voice quality features such as shimmer (amplitude irregularity), jitter (pitch irregularity), fundamental frequency, duration, harmonics-to-noise ratio (HNR), and power are also used [11]. In addition, a combination of prosodic features and voice quality information gives better information about the identification of emotion in comparison with only prosodic features. Some context information has also been studied [28] for emotion recognition. In [28], the authors present an SER system based on cultural information. The authors proved that cross-cultural-based SER performs better than multi-cultural and intra-cultural paradigms. Since these features mentioned above are low-level, they may not contain enough emotional information to identify the subjective emotional state. It may be possible to employ deep learning strategies to learn high-level features that are automatically effective for speech emotion recognition to address this problem.

Proposed methodology

In this section, we present our proposed DCRNN model with a different classifier. The structure of our proposed model is shown in Fig. 2.

First, we extracted the three channels or 3-D log Mel-spectrograms similarly RGB color images from the raw audio samples. Then we created 3-D log Mel-spectrograms which were fed to the deep convolutional neural network. We used pre-trained AlexNet [19] DCNN model to learn deep emotion feature from image such as log Mel-spectrograms. Next, we input the learning features into the Bi-directional gated recurrent unit (Bi-GRU) to extract and obtain two-dimension high-level features, after Bi-GRU, highlighting emotion features. Finally, an ensemble of two classifiers is employed to categorize the utterance-level features for SER. The details of input and output of each part of DCNNs is summarized in subsections below.

Creation of DCRNN input

The spectral features can identify emotional details better in time–frequency correlation and extract high-level information from the spectra using 2-D images. Therefore the researcher gives more attention on spectral features for speech emotion recognition [17, 29]. Generally, durations of speech signals are different, but most of the deep learning models require a fixed size of input. Representations of incomplete feature maps may not detect the correct emotional state of an utterance. To overcome this drawback, we extract 3-D log Mel-spectrograms (Static, Delta, Delta-delta) from the 1-D raw speech signal as inputs to our proposed DCRNN model, to minimize the loss of emotional information.

The process of creation of three-channel utterance level log Mel-spectrograms are as follows. (1) We use Librosa Audio Library [30] to originate the log Mel-spectrograms

from the speech signals under 16 kHz sample rate. (2) Then, we apply first and second-order derivative on the 2-D static log Mel-spectrogram along with the time axis to find the delta and delta-delta log Mel-spectrograms. So it creates itself a 3-D log Mel-spectrogram. The 3-D log Mel-spectrogram of each emotion on the Odia database is shown in Fig. 1. Finally, we resize the log Mel-spectrograms to 277 × 227 × 3 because pre-trained (AlexNet) DCNN requires an input size of 277 × 227 × 3.

Learning extracted feature using deep CNNs

After generating log Mel-spectrograms, we performed pre-trained AlexNet [19] DCNN for feature extraction. The AlexNet is a powerful model capable of achieving competitive performance on challenging small datasets. On the other hand pre-trained networks like VGG, ResNet, and efficient networks are much deeper and has more parameters, which require a very large number of inputs to achieve high performance [31]. We were inspired by the uses [8, 16] of pre-trained AlexNet. In the AlexNet network, the original parameters remain the same, and the layers are used to generate features. The AlexNet deep neural network contains several convolutional layers (CL), dropout, max-pooling layers, fully-connected layer, and the 'Relu' (rectified linear unit) activation function. The detailed description of DCNNs layers as follows.

The AlexNet model accepts with a fixed input size of 277 × 227 × 3. Each convolutional layer consists of several filters, kernel size, non-linear activation function, and padding. The convolutional layer is used to extract local patterns of the input and generates the feature maps. The AlexNet model has five convolutional layers (CL1, CL2, CL3, CL4, and CL5). The CL1, CL2, and CL5 are followed by the max-pooling layer, as shown in Fig. 2. The CL1 layer has a 96 kernel filter and a kernel size of 11 × 11 with a stride number of 4. The size of the CL2 layer is 5 × 5 with 256 kernels and a stride of 1. The CL3 layer has a size of 3 × 3 with 384 kernels connected to the outputs of the CL2 layer, and the CL4 layer has a size 3 × 3 with 384 kernels. The Relu activation function is used in each convolutional layer which increases the training process.

The max-pooling layer reduces the feature maps by utilizing maximum filter activation to get more high-level features. The output from the last max-pooling layers is fed to the fully connected layer. In the AlexNet) [19] model, the fully connected layers are FCL6, FCL7, and FCL8. FCL6 and FCL7 build a 4096-dimensional feature vector, whereas the FCL8 layer contains a 1000-dimensional feature vector because of 1000 types of categories on the Image Net data. We have not used FCL8; the FCL7 produces a 4096-D feature vector connected to the Bi-GRU layer.

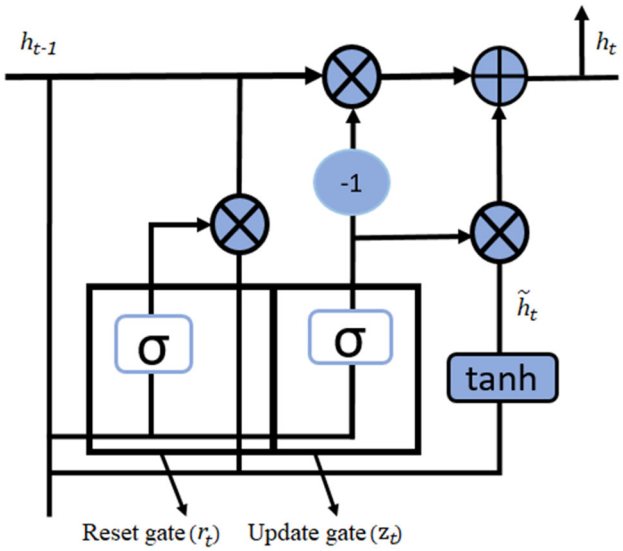

Fig. 3 The diagram of GRU structure

Bi-directional recurrent units

RNN [39] is proposed to solve the problem with time-series data. Most of the DNN, namely, convolutional neural network and multi-layer perceptron neural network are built with the help of weight connection of every layer, and the nodes between every layer are separated. So, nodes are independent of each other. However, in real life, the utterance is a time series with a variable length [32, 33]. Therefore, the previous utterance of the speaker is strongly related to the present utterance, which requires a model that can review the past information and process information of the different length of time-series data. To solve these issues, Hochreiter and Schmidhuber implemented Long Short Term Memory (LSTM) [34]. But the LSTMs take longer time, and require more memory to train than GRUs. And GRUs perform better than LSTMs in small amounts of training data.

Transformer architecture is often another common choice. However, this architecture does not capture the input order information. In [35], the authors point out that the Transformer only starts to outperform CNNs when data is more to train for the classification tasks. However, audio datasets typically do not have large amounts of data, which motivates us towards the use of Bi-GRU. The GRU cell is a particular type of RNN and the modified version of LSTM, as shown in Fig. 3. In the GRU cell, the cell state and the hidden state merge, and the input gate and the forgotten gate are combined and built as an update gate. Output of the last fully connected layer (4096-D) of DCNN network is connected to the Bi-GRU network with a sequences input of $\{x_1, x_2, x_3 \ldots\ldots x_t\}$ and we get output sequentially as $\{y_1, y_2, y_3 \ldots\ldots y_t\}$ by calculating each of input using activations ('Relu') functions

in the network on the basic formulations from time $t = 1$ to $t = T$

From the Fig. 3. We get,

$$h_t = \tilde{h}_t \odot z_t + h_{t-1} \odot (1 - z_t) \tag{1}$$

Here, z_t defines update gate, $\tilde{h}_t$ represent the current value of the present hidden state, $\tilde{h}_t$ denotes the activation value of the current hidden state, and h_{t-1} is the activation value of previous hidden state.

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z), \tag{2}$$

$$\tilde{h}_t = \emptyset(W_h x_t) + r_t \odot (U_h h_{t-1})) + b_h, \tag{3}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r), \tag{4}$$

where, σ defines the sigmoid function, x_t is the input of the GRU cell, $\odot$ defines the element-wise multiplication operation,$\varnothing$ is the tanh function, and W and U are the weight matrix, used during the training operation. The main contribution of r is to control how much information passes through x_t and how much previous information will affected by$\tilde{h}_t$. In essence, the GRU unit can overcome the problem of long-term distance information learning and at the same times also overcomes the problem of gradient dissent.

In this study, we use a speech spectrogram to detect emotion categories using a sequence labelling task. The unidirectional GRU cell cannot handle well a large number of speech samples of different durations. Therefore, Bi-GRU (Bi-direction Gated Recurrent Unit) [39] is utilized to learn present information and the past information of a variable-length of sound samples.

Softmax Classifier

Select an efficient classifier is vital for final classification of emotion. Most of the deep learning model uses Softmax [23] classifier. For example, n possible classes has n nodes in the Softmax layer denoted by c_j.

Where c_j is defined as the discrete probability distribution.

$$\text{Therefore} \quad \sum_{j=1}^{n} c_j = 1. \tag{5}$$

Output of the Softmax functions formula is as follows:

$$c_j = \frac{\exp(d_j)}{\sum_{i=1}^{n} \exp(d_i)}, \tag{6}$$

where, d_j is sum of the input into a Softmax layer which can be define as:

$$d_j = \sum_i h_i W_{ij} \tag{7}$$

Here, d_j is the activation in the second last layer and W_{ij} define the weight associate between the second last layer and Softmax layer. Thus the final prediction class $\widehat{j}$ would be:

$$\widehat{j} = \underset{j}{\text{argmax}} c_j(x) \tag{8}$$

Multi-class SVM classifier

The easiest way to extend SVMs for multiclass classification problems is using the one-vs-all method [23, 36]. For example, in class classification problems, n number of linear SVMs will be independently used, where the data from the other classes form the negative cases. The output representation of the n^{th} SVM is:

$$d_n(x) = W^T x \tag{9}$$

And, the final predicted category is calculated using Eq. (10):

$$\underset{n}{\text{argmax}} d_n(x) \tag{10}$$

The SVM classifier's prediction is the same as the Softmax classifier demonstrated in Eq. (8). The only difference between multiclass SVM and Softmax is in their parameters-weight matrices W. Softmax classifier layer minimizes cross-entropy loss, while SVM classifier tries to find the maximum boundary between the data points concerning the classes.

Ensemble of softmax and SVM classifier

Ensembles classifiers are generally proposed by a combination of two or more classifications. To improve over the best performing classifier, ensemble classifiers must comprise accurate base classifiers [37]. This paper has used the prediction probabilities of Softmax (marked as P^{Softmax}) and SVM (indicated as P^{SVM}) individually. Then we combine Softmax and SVM classifier to ensemble their probability as P^{Ensemble} to predict the final class. We use their maximum probabilities from individual classifiers, as mention in Eq. (11).

$$P^{\text{Ensemble}} = \max\left(P^{\text{Softmax}}, P^{\text{SVM}}\right) \tag{11}$$

This strategy has improved our accuracy significantly. Comparing the results, our method with two different classifiers (SVM and Softmax) shows varied effectiveness. The experimental results show that the DRCNN with ensemble classifier is extremely accurate in identifying emotion.

Experiments

Speech emotional datasets

To illustrate the performances, we tested our model on the Odia dataset and one popular public dataset (RAVDESS). Here, we use RAVDESS [23] dataset to validate our Odia dataset, which is widely used in speech emotion recognition.

The Odia dataset consists of 60 different utterances with seven different emotions: anger, surprise, fear, sadness, happiness, neutral, and disgust [38]. In our previous work, we have used six discrete emotions and a total of 3240 utterances. The dataset is collected from three different Odia dialects (Sambalpuri, Cuttacki, Berhampuri). Each dialect is recorded by six different speakers (three male and three female) whose ages gap between 19 to 40 years. We use ten different Odia emotion sentences, and every sentence was repeated three times in all three dialects. So, in total, 18 (six speakers and three dialects) $\times$ 10 (number of sentences) $\times$ 7 (number of emotion) $\times$ 3 (number of repetition) $=$ 3780 utterances are collected. All the utterances are recorded at a sampling frequency of 8.1 kHz with 16-bit quantization.

The RAVDESS speech emotional corpus [23] was recorded in the English language. The whole speech corpus consists of 1440 emotional audio (.wav) files with eight discrete emotional states: fear, calm, neutral, angry, disgust, sadness, boredom, and happiness. The dataset is completed by twenty-four (twelve male and twelve female) North American professional actors and the average time duration of each audio file is 3 s. The complete recording process was done at a sampling rate of 48 kHz have 16-bit quantization. The details of each emotional state of the RAVDESS and Odia datasets are shown in Table 1.

Table 1 The number of each emotional state of the RAVDESS and Odia datasets

Emotional state	RAVDESS	Odia
Anger	192	540
Disgust	192	540
Fear	192	540
Happiness	192	540
Neutral	96	540
Sadness	192	540
Surprise	192	540
Calm	192	–
Total	1440	3780

Experimental setup

The architecture of our proposed model is illustrated in Fig. 2. The DCRNN model is trained using a batch size of 32 and Adam optimizer [46] with a learning rate of 0.001. We set up 150 epochs for the training of our model. Our experiment is carried out on the TensorFlow, and Keras [39] deep learning platform with computer configuration is on Windows 10 Pro 64-bit operating system, Intel(R) Xeon(R) E-2224 CPU @ 3.4 GHz, NVIDIA QUADRO P620 GPU with 16 GB memory.

The Train-Test Split technique motivates us, from recent years of studies [3, 7, 8]. We split the datasets (Odia dataset and RAVDESS) into 80% training and from the remaining 20% of the data, 5% are used for tuning and 15% are used for test the model. The training samples are divided into multiple train-test splits to overcome the overfitting problem and get more stable results. Therefore, the total samples of datasets are divided into n number of folds. From the n folds, we used $(n-1)$ folds for training, and rest of (one fold) was used for the test and validation set; we set n equal to 5.

Results and analysis

Effects of Bi-GRU layer

It is essential to find out the suitable number of Bi-GRU layers and how many neurons are needed per layer to achieve the best-optimized model. To further investigate, we have studied the effect of different classifiers with and without pre-trained models. So, we first optimize the Bi-GRU layer, which is employed with the output of deep CNNs. The hidden layer and neurons are heavily dependent on the performance of the deep learning model. We train and validate our model with different numbers of Bi-GRU (with 1, 2, 3) layers with varying number of neurons (with 128, 256, 512) on the Odia and RAVDESS datasets using Softmax and SVM classifier. After conducting experiments on various layers, we concluded that $\text{BiGRU}^2{}_{128}$ (two Bi-GRU layers with 128 neurons) on the RAVDESS dataset and $\text{Bi-GRU}^2{}_{256}$ (two Bi-GRU layers with 256 neurons) on the Odia dataset performs better on different classifiers.

Effects of without pre-trained DCNN

We also trained and tested our method without pre-trained DCNN instead of pre-trained DCNN (AlexNet) on the same

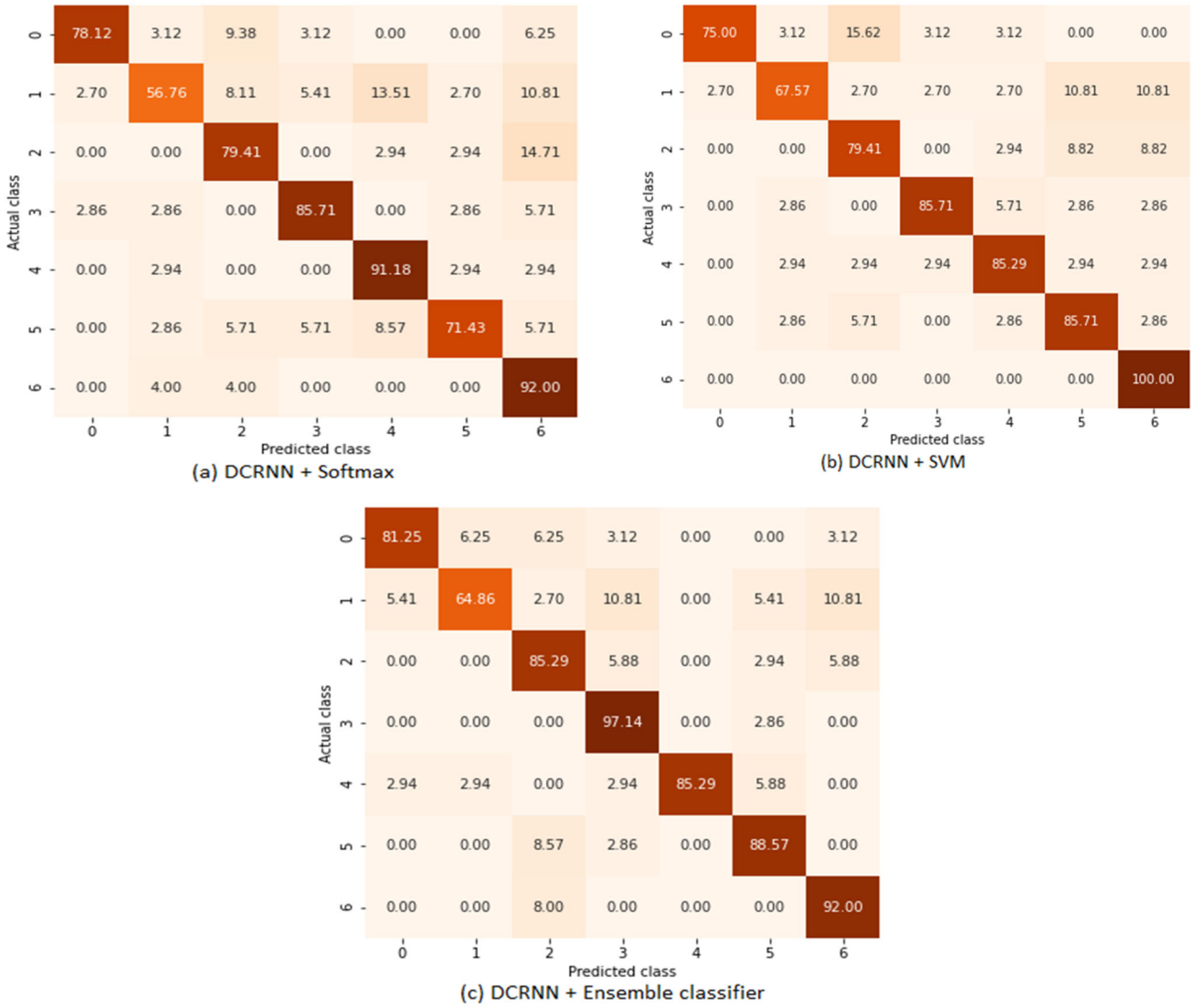

Fig. 4 Confusion matrix of seven class emotions on Odia dataset with different classifiers **a** DCRNN + Softmax, **b** DCRNN + SVM, **c** DCRNN + Ensemble classifier; (0: angry, 1: disgust, 2: fear, 3: happiness, 4: neutral, 5: sadness, 6: surprise)

model of our proposed method. Each parameter of the convolutional layers was randomly initialized with a standard normal distribution. The overall accuracy found without pre-training architecture was 81.24% of Odia dataset and 74.32% of RAVDESS with ensemble classifier. Then, we use the pre-trained ImageNet (AlexNet) model. The pre-trained ImageNet (AlexNet) model showed improved performance by 4.07% and 3.22%, respectively, compared to without the pre-trained model with ensemble classifier as shown in Table 2. The result demonstrates that using a pre-trained DCNN model improves the recognition accuracy and convergence rate.

DCRNN + Ensemble classifier results

Here, we represent the confusion matrix for investigation of the performances of the DCRNN model using Softmax classifier, SVM classifier, and the ensemble classifiers on Odia dataset and RAVDESS dataset shown in Figs. 4 and 5. Figure 4 represents the confusion matrix of the DCRNN model using Softmax classifier on Odia dataset; 'neutral' and 'surprise' achieves the highest recognition rate of 91.18% and 92%, respectively. In comparison 'disgust' is the lowest accuracy rate of 56.76%, and the other four emotions are obtained with accuracies below 90% with an overall accuracy of 81.53%, as illustrated in Table 2.

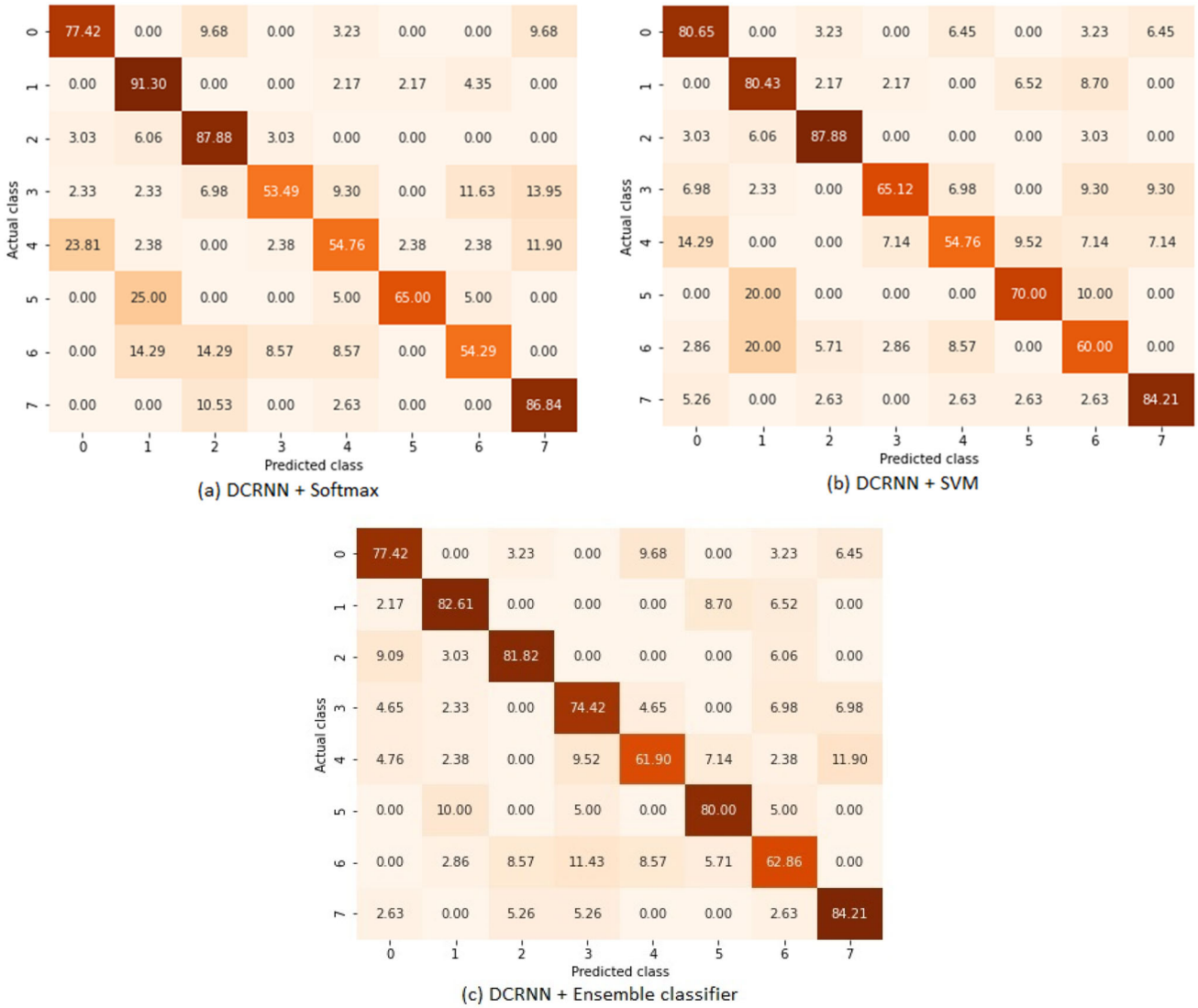

Fig. 5 Confusion matrix of eight class emotions on RAVDESS dataset with different classifiers **a** DCRNN + Softmax, **b** DCRNN + SVM, **c** DCRNN + Ensemble classifier; (0: neutral, 1: calm, 2: happiness, 3: sadness, 4: angry, 5: fear, 6: disgust, 7: surprise)

Table 2 Comparison of accuracy (%) level of DCRNN model at different classifiers with Pre-trained and Without Pre-trained DCNN

Model	Softmax classifier	SVM classifier	Ensemble (Softmax + SVM) classifier
With-out pre-trained (Odia dataset)	76.10	78.94	81.24
With pre-trained (Odia dataset)	**81.53**	**83.62**	**85.31**
With-out pre-trained (RAVDESS)	69.50	70.85	74.32
With Pre-trained (RAVDESS)	**73.52**	**74.91**	**77.54**

Bold indicates the best performance

Figure 4b shows the confusion matrix of the DCRNN model using SVM classifier; 'surprise', 'sadness', and 'disgust' emotions show increased recognition rates of 8% (from 92 to 100%), 13.28% (71.43–85.71%), and 10.81% (56.76–67.57%), while 'angry' and 'neutral' show decreased rates of 3.18% (78.12–75%) and 5.89% (91.18–85.29%). Recognition rates of the two other emotions remain relatively same. The overall recognition rate increased from 81.53 to 83.62%, which shows that the SVM classifier performs slightly better than the Softmax classifier.

And Fig. 4c displays the confusion matrix of the proposed DCRNN model with an ensemble classifier. The confusion matrix illustrates that the recognition rate of 'angry', 'fear', 'happiness', and 'sadness' are 81.25%, 85.29%, 97.14%, and 88.57%, which indicates that 3.13%, 5.88%, 11.43%, and 17.14% as compared to Softmax classifier and an increase of 6.25%, 5.88%, 11.43%, and 2.86% when compared with the SVM classifier. The overall accuracy rate is found 85.31%, which is 3.78% and 1.71% better than the Softmax and SVM classifier.

On the other hand, Fig. 5 shows the performance on the RAVDESS dataset with Softmax, SVM, and an ensemble of Softmax and SVM classifiers with eight emotions. Figure 5a states the performance of the DCRNN model using the Softmax classifier on the RAVDESS dataset. From Fig. 5a, we observed that 'calm', 'happiness', and 'surprise' are recognized well with an accuracy rate of 91.30%, 87.88%, and 86.84%, whereas 'neutral' classified relatively well with a recognition rate of 77.42%. The other four emotions classified less than 70%. The overall accuracy observed is 73.52%, as shown in Table 2. Figure 5b shows the confusion matrix of the DCRNN model using SVM classifier; 'neutral', 'calm', 'happiness', and 'surprise' can be recognized with a recognition rate of 80.65%, 80.43%, 87.88%, and 84.21% respectively. At the same time the other four emotions are indicate a recognition rate below 75%. The overall recognition rate is 74.91%, which shows that our SVM classifier performs better than the Softmax classifier.

Finally, we demonstrate that the recognition rate of ensemble classifier of 'sadness', 'angry', 'fear', and 'disgust' are classified 20.93%, 7.14%, 15%, and 8.57% greater accuracy, than Softmax classifier. The accuracy figures of 'sadness', 'fear', 'angry', and 'disgust' are 9.30%, 10%, 7.14%, 2.86% and 'calm' as 2.18% better performs than SVM classifier as shown in Fig. 5c. On the other side, the accuracy rate of 'happiness', 'calm', and 'surprise' decreases 8.69%, 6.06%, and 2.63%, respectively, compared to the results of the Softmax classifier; 'neutral' and 'happiness' indicate, 3.23% and 6.06% decrease in the recognition rate from the SVM classifier.

The overall average recognition accuracy rate of the ensemble classifier of 77.54%, reveals that it outperforms the Softmax and SVM classifiers by 4.02% and 2.63%, respectively as shown in Table. 2.

Table 3 F1-measure of each emotion for different classifiers on the Odia dataset

F1-Score (Odia)	Softmax classifier	SVM classifier	Ensemble classifier
Angry	83.17	84.31	84.86
Disgust	72.23	76.15	75.33
Fear	78.09	77.29	82.18
Happiness	86.33	88.34	87.49
Neutral	85.57	84.51	92.07
Sadness	77.64	81.44	86.81
Surprise	80.15	82.91	84.37

Table 4 F1-measure of each emotion for different classifiers on the RAVDESS dataset

F1-Score (RAVDESS)	Softmax classifier	SVM classifier	Ensemble classifier
Neutral	71.97	72.21	75.32
Calm	82.62	82.95	84.27
Happiness	75.28	78.40	82.31
Sadness	65.16	70.61	75.05
Angry	61.87	59.83	68.49
Fear	74.61	74.29	73.18
Disgust	62.47	63.15	64.69
Surprise	78.30	78.44	80.17

In addition, we also report the value of the F1-score for each emotional state to calculate the statistical importance of our experimental results; the F1 represents the harmonic mean of precision and recall. Tables 3 and 4 represent the statistical performance on the Odia and RAVDESS databases.

From the results of the F1-score, we demonstrate that each dataset illustrates different issues in recognizing a particular emotional state. On the Odia dataset, 'disgust' is recognized slightly lower than all the other emotions by all the classifiers, whereas 'angry', 'disgust', and 'sadness' perform relatively below all three classifiers on the RAVDESS.

Finally, for further analysis of the effectiveness of the proposed model, the training losses curve of two datasets (Odia and RAVDESS) are represented in Figs. 6 and 7. It can be noticed that training through 150 epochs of two datasets has slight fluctuation in convergence, maybe for the duration of emotion samples were not equal. The average duration of Odia is 4.5 s, and the RAVDESS is 3 s.

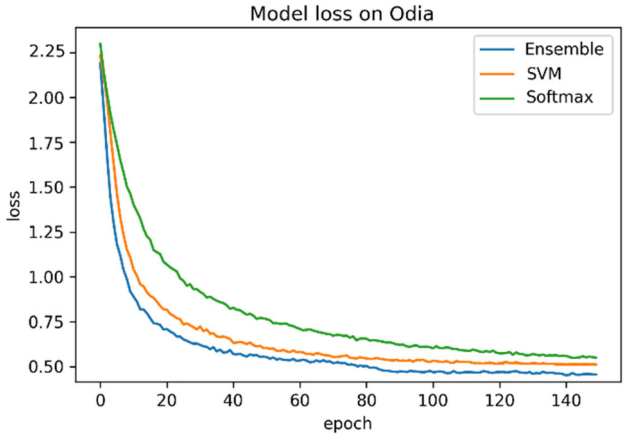

Fig. 6 The model training loss performance of the different classifiers on the Odia

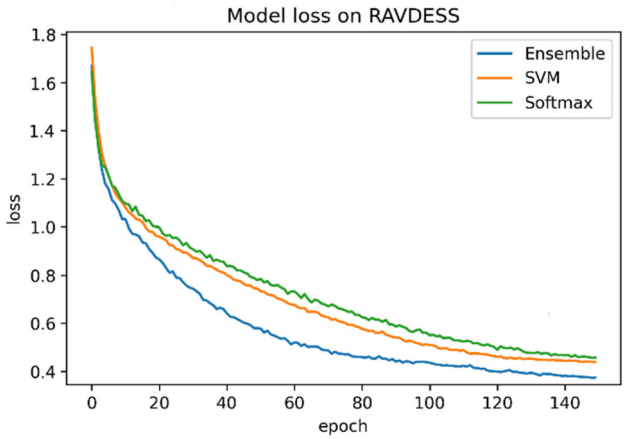

Fig. 7 The model training loss performance of the different classifiers on the RAVDESS

Comparison with recent work

Furthermore, we compare the result of our proposed method with several recent studies as well. Table 5 demonstrates a comparison between the results of our proposed method and previously public results with the model and features on the two datasets. Distinctly, on the RAVDESS dataset, our proposed method clearly expresses better with the performance level of 17.44%, 11.57%, 8.14%, 4.04%, and 5.93% collated with [3, 7, 17, 40, 41].

Our proposed method on the Odia dataset achieves 18.61% and 10.72% better accuracy compared to [38]. They used only prosodic features such as pitch, energy, format, etc., associated with SVM and GMM classifiers.

Conclusion and future work

Automatic speech emotion recognition becomes challenging because it is imperative to identify which features are prominent for speech emotion recognition tasks. This work has been motivated from the fact that we need to build more accurate emotion classification for regional languages for easy translation from one language to another. We have selected a deep learning method to build an SER system for the Odia language, because accuracy is the main factor and all emotional parts are not well understood.

This study is inspired by the effect of 3-D log Mel-spectrograms (like RGB color representation) for verifying emotional classes from the speech signal. We propose a new deep CRNN with an ensemble of two (Softmax and SVM) classifiers. A pre-trained deep CNNs ImageNet (AlexNet) model produces a deep high-level feature from the speech

Table 5 Accuracy (%) comparison with some state-of-the-art model with features

Database	Feature	Model and Classifier	Accuracy (%)
	PCA, NMF-1, Prosodic	Quadratic SVM [40]	60.10%
	Spectrogram	GResNet [17]	65.97%
RAVDESS	eGeMAPS, Supervector, Log-spectrogram, F0, MFCCs, FB128	Bi-LSTM + CNN + CapsNet [41]	69.40%
	3-D Log Mel-spectrogram	DCNN + CFS + MLP [7]	73.50%
	MFCCs, Spectrogram, Chromagram, Contrast, Tonnetz	Deep CNN [3]	71.61%
	3-D Log Mel-spectrogram	**Ours** (DCRNN + Ensemble classifier)	**77.54%**
	10 Prosodic	SVM [38]	74.59%
Odia	10 Prosodic	GMM [38]	66.70%
	3-D Log Mel-spectrogram	**Ours** (DCRNN + Ensemble classifier)	**85.31%**

Bold indicates the best performance

spectrogram. After that, the output of deep CNNs is learned by Bi-GRU to avoid long-term dependency and gives utterance level features. Finally, we get the classification results using the maximum probabilities of the two classifiers. Our experimental results show an 85.31% and 77.54% overall classification rate on the seven classes Odia dataset and eight classes RAVDESS dataset, which reveals outperforms others. In the future, we would like to further investigate our created Odia dataset with more data, and also other datasets from different languages applying our proposed framework. We plan also to implement the average mode of predictions by adding more multi-class classifiers. Further, we would like to extend our work by adopting different acoustic speech features with the modern techniques like using the transformer based model, to achieve more stable and reliable results.

Acknowledgements This research study is supported by the Department of Science and Technology (DST), Grant no. DST/ICPS/CLUSTER/Data Science/2018/General, India. We also thank Prof. J. Talukdar for improving the quality of this paper.

Declarations

Conflict of interest The authors have no conflicts of interest.

References

1. Khokher R, Singh RC, Kumar R (2015) Footprint recognition with principal component analysis and independent component analysis. Macromol Symp 347(1):16–26. https://doi.org/10.1002/masy.201400045
2. Mittal S, Agarwal S, Nigam MJ (2018) Real time multiple face recognition: a deep learning approach. In: Proceedings of the 2018 international conference on digital medicine and image processing, ACM, pp 70–76. https://doi.org/10.1145/3299852.3299853
3. Issa D, Demirci MF, Yazici A (2020) Speech emotion recognition with deep convolutional neural networks. Biomed Signal Process Control 59:101894. https://doi.org/10.1016/j.bspc.2020.101894
4. Le BV, Lee S (2014) Adaptive hierarchical emotion recognition from speech signal for human-robot communication. In: 2014 10th International conference on intelligent information hiding and multimedia signal processing, IEEE, pp 807–810. https://doi.org/10.1109/IIH-MSP.2014.204
5. Rázuri JG, Sundgren D, Rahmani R, Larsson A, Cardenas AM, Bonet I (2015) Speech emotion recognition in emotional feedback for human-robot interaction. Int J Adv Res Artif Intell 4(2):20–27
6. Ramakrishnan S, El Emary IMM (2013) Speech emotion recognition approaches in human computer interaction. Telecommun Syst 52:1467–1478. https://doi.org/10.1007/s11235-011-9624-z
7. Sui X, Zhu T, Wang J (2017) Speech emotion recognition based on local feature optimization. J Univ Chin Acad Sci 34(4):431–438
8. Mustafa MB, Yusoof MAM, Don ZM, Malekzadeh M (2018) Speech emotion recognition research: an analysis of research focus. Int J Speech Tech 21(1):137–156. https://doi.org/10.1007/s10772-018-9493-x
9. Farooq M, Hussain F, Baloch NK, Raja FR, Yu H, Zikria YB (2020) Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. Sensors 20(21):6008. https://doi.org/10.3390/s20216008
10. Zhang H, Gou R, Shang J, Shen F, Wu Y, Dai G (2021) Pre-trained deep convolution neural network model with attention for speech emotion recognition. Front Physiol 12:643202. https://doi.org/10.3389/fphys.2021.643202
11. Arano KA, Gloor P, Orsenigo C, Vercellis C (2021) When old meets new: emotion recognition from speech signals. Cogn Comput 13:771–783. https://doi.org/10.1007/s12559-021-09865-2
12. Lu G, Yuan L, Yang W, Yan J, Li H (2018) Speech emotion recognition based on long-term and short-term memory and convolutional neural network. J Nanjing Inst Posts Telecomm 38(5):63–69. https://doi.org/10.14132/j.cnki.1673-5439.2018.05.009
13. Sun L, Zou B, Fu S, Chen J, Wang F (2019) Speech emotion recognition based on DNN-decision tree SVM model. Speech Commun 115:29–37
14. Ayadi ME, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn 44(3):572–587
15. Swain M, Routray A, Kabisatpathy P (2018) Databases, features and classifiers for speech emotion recognition: a review. Int J Speech Technol 21(1):93–120
16. Wang ZQ, Tashev I (2017) Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In: 2017 IEEE international conference on acoustics, speech, and signal processing (ICASSP), pp 5150–5154
17. Jiang P, Fu H, Tao H, Lei P, Zhao L (2019) Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. IEEE Access 7:90368–90377. https://doi.org/10.1109/ACCESS.2019.2927384
18. Hu H, Xu M, Wu W (2007) GMM supervector based SVM with spectral features for speech emotion recognition. In: 2007 IEEE international conference on acoustics, speech, and signal processing (ICASSP), pp 413–416. https://doi.org/10.1109/ICASSP.2007.366937
19. Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 22(10):1533–1545
20. Shewalkar A, Nyavanandi D, Ludwig SA (2019) Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM AND GRU. JAISCR 9(4):235–245. https://doi.org/10.2478/jaiscr-2019-0006
21. Zhang S, Zhang S, Huang T, Gao W (2017) Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Trans Multimedia 20(6):1576–1590. https://doi.org/10.1109/TMM.2017.2766843
22. Zeng Y, Mao H, Peng D, Yi Z (2017) Spectrogram based multi-task audio classification. Multimed Tools Appl, pp 1–18
23. Livingstone SR, Russo FA (2018) The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5):e0196391

24. Badshah AM, Ahmad J, Rahim N, Baik SW (2017) Speech emotion recognition from spectrograms with deep convolutional neural network. In: 2017 International conference on platform technology and service (PlatCon), pp 1–5. https://doi.org/10.1109/PlatCon.2017.7883728
25. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105
26. Pao TL, Chen YT, Yeh JH, Cheng YM, Lin YY (2007) A comparative study of different weighting schemes on KNN-based emotion recognition in mandarin speech. Int Conf Adv Intell Comput Theories App. https://doi.org/10.1007/978-3-540-74171-8_101
27. Nwe TL, Foo SW, De Silva LC (2003) Speech emotion recognition using hidden markov models. Speech Commun 41(4):603–623
28. Ververidis D, 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 (ICME), Netherlands, pp 1500–1503
29. Tang Y (2015) Deep learning using linear support vector machines. arXiv:1306.0239
30. Schuller B, Rigoll G, Lang M (2004) Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: 2004 IEEE International conference on acoustics, speech, and signal processing (ICASSP), pp 1-577
31. Zhou Y, Sun Y, Zhang J, Yan Y (2009) Speech emotion recognition using both spectral and prosodic features. In: 2009 International conference on information engineering and computer science (ICIECS), Wuhan, China, pp 1–4. https://doi.org/10.1109/ICIECS.2009.5362730
32. Kachele M, Zharkov D, Meudt S, Schwenker F (2014) Prosodic, spectral and voice quality feature selection using a long-term stopping criterion for audio-based emotion recognition. 2014 22nd international conference on pattern recognition (ICPR). Stockholm, Sweden, pp 803–808
33. Pan Y, Shen P, Shen L (2005) Feature extraction and selection in speech emotion recognition. In: IEEE (AVSS) conference on advanced video and signal based surveillance, Como, Italy, pp 64–69
34. Petrushin VA (2000) Emotion recognition in speech signal: experimental study, development, and application. In: 6th International Conference on Spoken Language Processing, Beijing, China, pp 222–225
35. Quiros-Ramirez MA, Onisawa T (2015) Considering cross-cultural context in the automatic recognition of emotion. Int J Mach Learn Cyber 6(1):119–127
36. Chen M, He X, Yang J, Zhang H (2018) 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process Lett 25(10):1440–1444
37. McFee B, Raffel C, Liang D, Ellis DPW, McVicar M, Battenberg E, Nieto O (2015) librosa: audio and music signal analysis in python. In: proceedings of the 14th Python in Science Conference, pp 18–25
38. Dua M, Shakshi SR et al (2021) Deep CNN models-based ensemble approach to driver drowsiness detection. Neural Comput Appl 33:3155–3168. https://doi.org/10.1007/s00521-020-05209-7
39. Zhu Z, Dai W, Hu Y, Li J (2020) Speech emotion recognition based on Bi-GRU and Focal Loss. Pattern Recog Lett 140:358–365
40. Xiao Z, Xu X, Zhang H, Szczerbicki E (2021) A new multi-process collaborative architecture for time series classification. Knowl Based Syst 220:1–11
41. Xiao Z, Xu X, Xing H, Luo S, Dai P, Zhan D (2021) RTFN: a robust temporal feature network for time series classification. Inf Sci 571:65–86
42. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
43. Gong Y, Chung YA, Glass J (2021) AST: audio spectrogram transformer. arXiv:2104.01778
44. Duan K, Keerthi SS, Chu W, Shevade SK, Poo AN (2003) Multi-category classification by soft-max combination of binary classifiers. In: Proceedings of the 4th international conference on multiple classifier systems, MCS'03, Springer, Berlin, pp 125–134. https://doi.org/10.1007/3-540-44938-8_13
45. Morrison D, Wang R, De Silva LC (2007) Ensemble methods for spoken emotion recognition in call-centres. Speech Commun 49(2):98–112. https://doi.org/10.1016/j.specom.2006.11.004
46. Swain M, Routray A, Kabisatpathy P, Kundu JN (2016) Study of prosodic feature extraction for multidialectal Odia speech emotion recognition. In: IEEE region 10 conference (TENCON), pp 1644–1649
47. Kingma DP, Ba JL (2017) ADAM: A method for stochastic optimization. arXiv:1412.6980
48. Geron A (2017) Hands-on machine learning with Scikit-Learn and Tensor-Flow: concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc, USA
49. Shegokar P, Sircar P (2016) Continuous wavelet transform based speech emotion recognition. In: Proceedings of the 10th international conference on signal processing and communication systems, pp 1–8. https://doi.org/10.1109/ICSPCS.2016.7843306
50. Jalal MA, Loweimi E, Moore RK, Hain T (2019) Learning temporal clusters using capsule routing for speech emotion recognition. In: Proceedings of the INTERSPEECH 2019, Graz, Austria, pp 1701–1705. https://doi.org/10.21437/Interspeech.2019-3068