Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 171 (2020) 12–21

Procedia Computer Science

www.elsevier.com/locate/procedia

Third International Conference on Computing and Network Communications (CoCoNet'19)

An End-to-End Model for Detection and Assessment of Depression Levels using Speech

Srimadhur N.S, Lalitha S

Department of Electronics and Communication Engineering, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India

Abstract

In this work, individuals with psychological unit of depression are detected using speech samples. Spectrogram based convolutional neural networks and end to end convolutional neural network models are implemented to achieve the task. Parameter tuning has been performed by choosing different sub-parameters of convolutional neural network. Speech samples from audio visual emotion challenge (AVEC) 2016 DAIC-Woz dataset are utilized for validating the models. Experimental analysis has shown that performance of end to end model is ahead of spectrogram based model and baseline models by an efficiency of 13%. Further, the proposed model has been applied to estimate the severity level of depression using the PHQ-8 scores of speech samples which has been never attempted using speech samples.

Keywords: Depression detection, Spectrograms, severity level, convolutional neural networks;

1. Introduction

Depression is psychological disorder identified by an extensive and continuous low mood, led by a low self-assurance along with loss of interest in usually enjoyable activities. Depression is the main cause of more than two-thirds of suicides every year [1]. If left untreated will have serious effect on individuals, their families and society. Despite the high pervasiveness of depression, existing diagnostic techniques depend extremely upon verbal reports of patients, with the help of relatives or friends and experience of clinicians so there is chance of subjective biases due to which there tends to be inconsistent results at different situations [2].So it is extremely important to go for automatic

10.1016/j.procs.2020.04.003

depression detection. While research into biological markers for depression has showed many positive results, such as low serotonin levels, no biomarker determines to be depressive state. While biomarkers remain difficult to find, important developments have been made in using computing and signal processing as a diagnostic tool. These frameworks depend either on speech processing methods where study is based on features identified with prosodics, the vocal tract, combination of acoustic low level features belonging to prosodic and spectral features and parameters extracted from glottal waveform in frequency and time domains [3]-[4]. Classification of depressed speech is based on feature selection methods on combination of different features.[5] Research analysis has been done on facial and body expressions where non-verbal behaviors of depressed persons using both automatic feature discriminators and manual feature extractors are investigated on combination of facial and vocal sounds [6]-[7]. Analysis has been made on eye gaze features extracted from face videos using active performance models for a binary classification task, mean distance between eyelids was appreciably smaller and the mean period of blinks was considerably longer in depressed persons [8].

The studies mainly analyzed speech signals, video signals or combination of both of them. In general discriminative features are extracted form speech or image and classifiers are built on top of it for classification. In this work depression detection using speech carried out in three stages which are preprocessing, feature extraction and classification. In pre-processing stage silence regions are removed from speech signal and to divide into frames windowing technique is used. In feature extraction stage different features like prosodic, spectral, temporal, MFCC, LPCC, PLP are extracted and from them discriminative features are picked up manually [9]. Artificial neural network, Hidden Markov Models, Support Vector Machine and Gaussian Mixture Models were used as classifiers [10].More recently deep learning methods have showed their capacity in many audio and video based applications. These methods learn discriminative features through multiple layers and performed better than traditional methods.

In the remainder of paper, Section 2 provided related work done in this area. Following this proposed methodology of this study is presented in Section 4. Experimental setup is discussed in section 4 and results are analyzed in section 5. Finally, future directions are proposed in section 6.

2. Related Work

So far, many automatic depression detection methods using speech signals have been investigated. In the first stage acquisition of voice data is crucial and significant, different kinds of speech types reading, interview and picture description are the common ways. Hailiang Long et al. [9] examined discriminative power of the different speech types by extracting acoustic features like short time energy, intensity, loudness, formant frequencies, shimmer, jitter, zero crossing rate etc for analysing the speech signals. Features were extracted using openSMILE software, introduced multiple classifier system using SVM as a classifier in which combined the results of all features for predicting the classes and obtained 78.02% prediction accuracy. In another study by S. Alghowinem et al. [10], 60 real world subject recordings has been taken through interview for examination. Acoustic and linguistic features has been extracted through openSMILE software and classified them using multiple classifiers like Gaussian Mixture Models (GMM), Support Vector Machines (SVM), Multilayer perceptron neural networks(MLP) and Hierarchical fuzzy signature(HFS). Among all features combination of SVM and GMM has given best performance compared to other classifiers. Different audio features has been analysed for classification of depressed speech utilizing i-vectors for representing in audio information in low dimensional space by author Paula Lopez-Otero in this paper[11]. MFCCs, Shifted Delta Cepstrum(SDC) features, Energy, Spectral features, Prosodic features are examined for classification and Gaussian mixture model(GMM) used as a classifier and verified results on AVEC 2013 dataset. An accuracy of 70% is achieved with spectral features.

Due to limited work in depression, an investigation is performed into speech features and classifiers for various speech applications Veena narayanan et al. [12] has investigated on different conditions of stress recognition from speech which is another work releated to psycholological state. Five different stress condition have been detected using Interspeech 2010 features on Indian context database through Naïve bayes, simple logistic, sequential minimal optimization and decision tree classifiers and achieved better performance than previous works [13]. In another study by S.Lalitha et al. [15] Spectrograms of speech emotions are extracted, discriminative features are learnt through convolutional neural networks, results are validated on Berlin corpus database and achieved average accuracy of 84.3%. A number of studies have also been proposed for end to end deep learning model for speech applications like

speech emotion recognition where speech samples are directly given to the deep learning model and discriminative features are learnt then Long short term memory (LSTM) is used to model the temporal features and classification has been performed without need of separate classifiers [16].

In the reported works appreciable performances have been achieved however manual hand-picked features have been utilized which requires subject knowledge and used shallow architectures. In this paper [14] author proposed an innovative deep learning method where Convolutional neural networks (CNN) and Long short term memory (LSTM) are used for learning features and classifying the speech into depressed and non-depressed states, pre-processing is performed by removing silence regions, spectrograms and MFCC coefficients which are low level features were calculated and experiments are performed on AVEC 2016 DAIC-WOZ dataset and achieved F1 score of 0.52 for depressed and 0.7 for non-depressed.

Various bottleneck issues do exists with the existing works on speech based depression detection like hand-picked features for classification which requires expert subject knowledge and human efforts, requirement of additional classifiers for classification purpose, speaker and language dependency. This work is an attempt to address the stated issues. In this work, an investigation is carried out on depression detection using spectrogram based convolutional neural network and end to end convolutional neural network models. Parameter tuning has been performed and comparative analysis has been carried out between two models and best model has been chosen for categorizing the depression state.

3. Proposed Methodology

Fig.1 depicts the flow diagram of Spectrogram based convolutional neural network and end to end convolutional neural network models. Models consist of four modules Database preparation, Pre-processing, Model design cum training and classification which are discussed as follows.

3.1 Dataset Preparation

In this work, AVEC 2016 DAIC-WOZ database which is introduced by DeVault et al.[17] based on interviews with clinically depressed patients and normal persons is utilized. This database consists of speech samples of 189 participants among which 146 were of non-depressed and 43 of depressed persons. This dataset was used in the Audio/Visual Emotion Challenge (AVEC) 2016 [18].

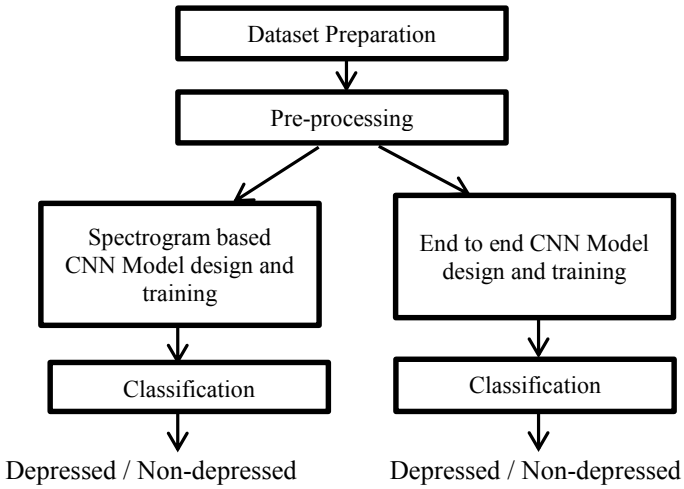
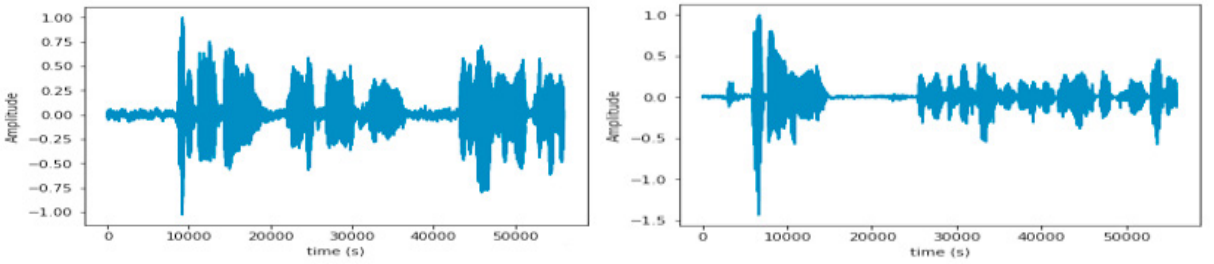


Fig.1. Flow diagram of proposed depression detection models

Fig.2(a) and 2(b) depicts the input speech samples of non-depressed and depressed states. The non-speaker part has been removed and segmented the samples into speech samples of 7 sec duration each using NCH wavepad sound editor which is software developed by NCH software[19] and re-sampled at 8 KHz maintaining consistent frame rate

with a resolution of 16 bits/sample. This has been done to for balancing the samples between depressed and non-depressed speech states. Total samples considered are 2240 of which 1107 are of non-depressed and 1133 of depressed samples. From the input samples it is clear that non-depressed speech sample's intensity is spread across the time where as in depressed speech it is concentrated in short intervals of time and intensity of speech signal is more in non-depressed speech when compared to depressed speech samples.



(a) (b)

Fig.2. Input speech samples of (a) Non-depressed state (b) depressed state

3.2 Pre-processing

Data samples are divided into training and test samples exclusively to overcome under fitting of the model. Training and test samples arrays are re-shaped into a 2D array in which each row corresponds to a sample and size of columns corresponds to number of channels(number of training or test samples) which is given as input to convolutional neural network.

3.3 *Spectrogram based convolutional neural network model*

Architecture of this model is shown in Fig.3. In this model Spectrograms are given as input to the convolutional neural network (CNN) and low level features are extracted from spectrograms where spectrogram is log scale plot of Short time fourier transform (STFT). CNN takes gray scale image of spectrogram as a input, kernel slides over the image and learns the patterns of depressed and non-depressed samples. Initially model learns the vertical lines, edges and as it goes deeper it learns more discriminative features. Here four kernels of sizes 3*3, 1*3 and 3*1 are used in three convolution layers respectively. Max pooling layers are used after each convolution layers to reduce the dimensionality of the feature maps and reduce redundancy, drop out layers with 25% probability and Regularization are used to generalize the model. Generated feature maps converted to single array using flattening layer and given to dense layers. Architecture of model is shown in Fig.3.

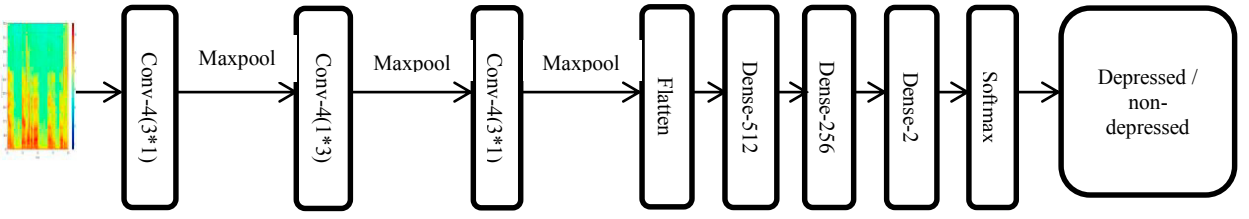


Fig. 3. Spectrogram based convolutional neural network model for depression detection

3.4 End to end convolutional neural network model design

The key operation of our technique is convolution,

$$(g * h)(t) = \sum_{k=-T}^{T} g(t).h(t-k) \quad (1)$$

Where g(t) defines kernel function, which operates on raw speech samples h (t).To reduce the dimensionality of samples, used max-pooling operations. Pooling size (P) can be selected based on kernel size (K) and rate of overlapping (R) using the relation given below

$$R = \frac{K-1}{K+P-1} \quad (2)$$

The overlapping rate (R) generally has to be less than 1 and is commonly considered as 0.5. Stride can be used to reduce dimensions but striding produces worse performance than using max-pooling. In max-pooling it uses necessary information and discard redundant information, whole information is considered when using striding. Max-pooling rate has to be less than 0.5 to avoid extracting same features for consecutive frames.

3.4.1 Model design

Proposed model, which is illustrated in Fig. 4, is explained below.

Input. After pre-processing the raw speech samples to normalized values, segmented it into 7 s sequences and given them as input. At 8 Khz this corresponds to 56000 sized input vector.

Temporal convolution. Investigated by using 64 filters with kernel sizes of 8, 16, 24,32 to extract the information from raw signal. Based on kernels sizes of 8, 16, 24,32 applied max pooling sizes of 10, 18, 26,34 respectively to reduce the frame rate of signal and keeping discriminative features. Max pooling sizes are calculated using equation (2).

Temporal convolution2. In the deeper layers wanted to extract large number of higher level abstractions. So time domain convolutions with 128 kernels are considered of sizes 6, 12, 18. Based on kernels sizes of 6, 12, 18 max pooling sizes of 8, 14, 20 are selected respectively to reduce the frame rate of signal and keeping overlapping rate around 0.5.

Temporal convolution3. The last convolutional layer should provide higher level features so investigated it by using 256 filters of kernel sizes of 6, 12, and 18. Max pooling sizes of 8, 14, and 20 for kernel sizes of 6, 12, and 18 respectively are applied to keep the overlapping rate below 0.5.

Non-linearity and striding. For a powerful network non-linearity is needed so weighted sum of inputs has been passed to the non-linear activation function and used ReLU function for it. In general stride value taken as 1. Padding is appending zeros to maintain dimensionality same means preserving the size of feature maps.

Dropout. Due to high number of parameters our model contains used drop out layer after every pooling layer with 0.5 probabilities. This generalizes the model.

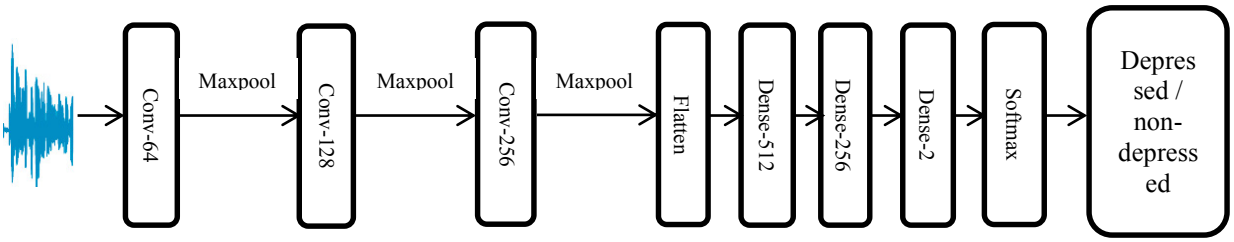


Fig. 4. Proposed end to end Convolutional neural network for depression detection using speech

Fully connected layers. Output from convolution layers is converted to 1d array using flattening layer and is given to fully connected layers in which error gets calculated through loss function and considered categorical cross entropy model as a loss function in this work. Once forward propagation is completed, error is calculated and back-propagation starts where weights are updated till the error gets reduced.

Softmax decision. This is where classification based on the calculated probability of the predicted classes takes place.

3.5 Back propagation and convolutional neural networks weight updation

Output vector of particular layer is given by:

$$O(x) = \sum_{x'} w_{x'}^{l}\ f(o_{x-x'}^{l-1}) + b_x^l \quad (3)$$

where w is weights of the kernels, a is input activation vector which is output of activation function of previous layer output, b is the bias vector of layer l. Weights of kernels are updated after processing of each batch and after each

iteration based on cost error and after each updation, error gets reduced.

For a total of N predictions, the predicted outputs a_p^L and their corresponding target values t_p the mean squared error if given by :

$$E = \frac{1}{2}\sum_{p=1}^{N}\left(t_p - a_p^l\right)^2 \quad (4)$$

For the back propagation two updates are performed, for the weights and deltas. For updating weights gradient components for each weight can be calculated by applying chain rule.

$$\frac{\partial E}{\partial w_x^l} = \sum_{x\prime} \frac{\partial E}{\partial o_{x\prime}^l}\frac{\partial o_{x\prime}^l}{\partial w_x^l}$$

$$= \sum_{x\prime} \delta_{x\prime}^l \frac{\partial o_{x\prime}^l}{\partial w_x^l} \quad (5)$$

By substituting eqn (3) in (5) and solving will gives :

$$\frac{\partial o_{x\prime}^l}{\partial w_x^l} = \frac{\partial}{\partial w_x^l}\left(\sum_{x\prime\prime} w_{x\prime\prime}^l f(o_{x\prime-x\prime\prime}^{l-1}) + b^l\right) \quad (6)$$

After expanding summation in eqn (6) results in zero values for all except for $x'' = x$ and gives us:

$$\frac{\partial o_{x\prime}^l}{\partial w_x^l} = \mathrm{f}(o_{x'-x}^{l-1}) \quad (7)$$

Substituting eqn (7) in (5) gives us :

$$\frac{\partial E}{\partial w_x^l} = \delta_x^l * \mathrm{f}(rot_{180^o}(o_x^{l-1})) \quad (8)$$

Where deltas $\delta_x^l = \delta_x^{l+1} * w_x^{l+1} f'(o_x^l)$ which implies that:

$$\frac{\partial E}{\partial w_x^l} = (\delta_x^{l+1} * w_x^{l+1} f'(o_x^l)) * \mathrm{f}(rot_{180^o}(o_x^{l-1})) \quad (9)$$

Weights of convolutional layers are updated using equation(9) during backpropagation stage and reduces error.

4. Experimental Background

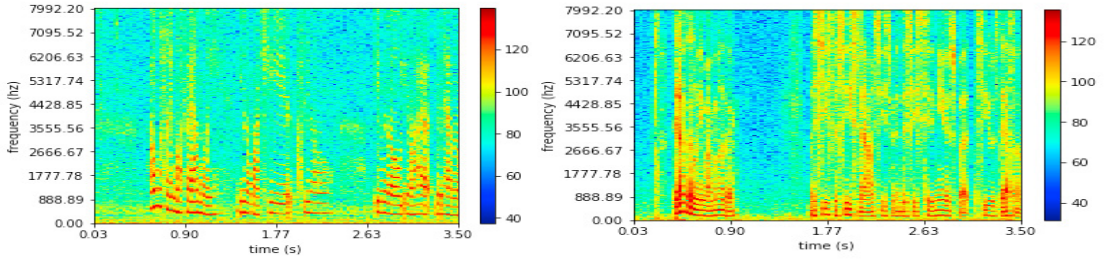
Experiments are performed on Python programming language utilizing Spyder tool [20]. Five cross validation has been performed by dividing Data set into Training and testing samples, labels are prepared as this is supervised learning and these are saved in npz format. Saved npz files are given as input to the models during training and measured performance using training accuracy and is evaluated on testing samples. Considered 20 iterations during training and adam optimizer is considered for training the model throughout all experiments with a fixed learning rate of 10^{-3}, $\beta 1=0.9$, $\beta_2=0.999$ with the batch size of 32 [21]. To prevent the effect of over fitting utilized regularization and in particular drop out layer after max pooling layers with a probability of 0.5 [22]. Model performance is evaluated on accuracy, recall, precision and F-measure parameters. Accuracy is number of samples that are predicted correctly among all classes. Recall is true positive rate which is defined by ratio of True positive samples to True positive and false negative samples. Precision is percentage of results which are relevant and is defined by True positive divided by total positive predictions. F-score is the trade-off between recall and precision that is harmonic mean of recall and precision [24].

5. Results and Discussion

Depression detection using speech is implemented in two models they are:
(i) Spectrogram based convolutional neural networks
(ii) End to end convolutional neural networks

5.1 Spectrogram based convolutional neural network model:

This model takes spectrogram images of data set as an input and each spectrogram image is of size 456*834 pixels as shown in Fig.5. From the spectrograms it is observed that intensity of speech is concentrated more at lower frequencies and low at higher frequencies for non-depressed samples whereas for depressed speech samples high frequency components also exists with higher intensities and intensity is presented more in short periods of time intervals. CNN uses these kind of features from spectrograms to classify samples into different states. Implemented two models by varying the number of convolutional layers. Model1 consists of two convolutional layers of four kernels each with kernel sizes of 3*3 and 1*3 respectively.Model2 consists of three convolutional layers with 4 kernels each with kernel sizes of 3*3,1*3 and 3*1 respectively. The selected kernel sizes are applied in such a way that in first layer edges patterns of images are detected and smoothening of image is done to identify the high frequency components and in consequent layers features along horizontally and vertically are learnt over the spectrograms. Performance characteristics of models are shown in Table 1. Model2 shows better performance when compared to Model1.



(a) (b)

Fig.5. Spectrogram of speech signal for (a) non-depressed state (b) depressed state

Table 1 Performance characteristics of both models and baseline model

Sl No.	Model	Class	Recall(%)	Precision(%)	F-Score(%)	Overall Accuracy(%)
1	Model1	Depressed	80	56	66	59.2
		Non-depressed	39	67	49	
2	Model2	Depressed	77	58	66	61.32
		Non-depressed	47	67	55	

Accuracy of depressed state samples is 80% whereas accuracy of non-depressed samples is 39% in Model1. In the case of model2 accuracy of depressed state is 77% and for non-depressed state it is 47%. Overall accuracy of Model1 is 59.2 and for Model2 it is 61.32. Achieved F-score for Model2 is 66% for depressed and 55% for non-depressed state. Performance of this model with lesser number of kernels is comparable to the model proposed by Di Huang et al.[14] where depression detection is implemented on convolutional neural network(CNN) and Long short term memory(LSTM) with 32 kernels and obtained F-score of 0.52 for depressed and 0.7 for non-depressed state.Main reason for lesser accuracy is because of variability in the volume during recording of data samples and normalized the data samples to reduce this affect but there is no changes in the spectrogram of data samples so this spectrogram based convolutional neural networks for depression detection is ineffective to variance of speaker volume To overcome this effect features has to be learnt directly from the raw speech which is illustrated in next section.

5.2 *End to end CNN model:*

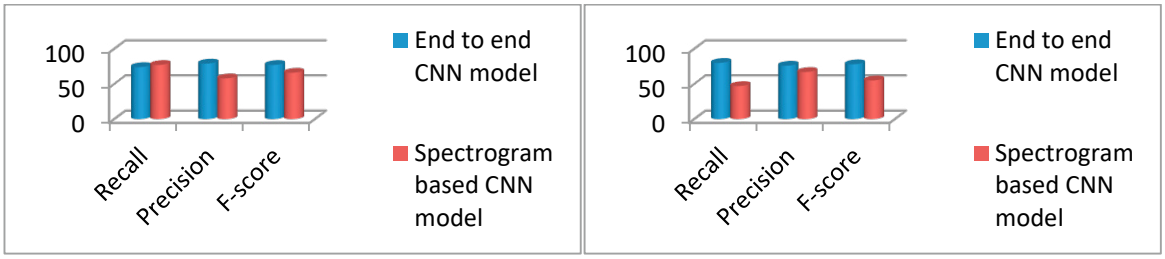
This model is built with three convolutional layers of 64 kernels,128 kernels and 256 kernels respectively. Number of kernels and number of convolutional layers are chosen in such a way that there is tradeoff between learning of discriminative features and generalization of model by not over training the samples where over training leads to overfitting in which prediction accuracies are less for test samples. Experiments have been conducted by varying the

kernel sizes of three layers of convolutional layers. Model1 is designed with kernel sizes of 8,6,6; Model2 with kernel sizes of 16,6,6; Model3 with kernel sizes of 16,12,12; Model4 with kernel sizes of 24,12,12; Model5 with kernel sizes of 32,12,12; Model6 with kernel sizes of 32,18,12.Kernel sizes are chosen by considering the size of input samples because if kernel size is more then it will miss indepth feature of speech signal. Also in the first convolutional layer as the input is raw waveform kernel size of filter has to be high so that it will learn the time domain features effectively. As go higher levels need small kernels to learn abstract features from the feature maps generated from previous layers. Model configurations for proposed models and performance characteristics are tabulated in Table 2.Model 6 has shown best performance among the models proposed. Accuracy of depressed samples is 74.64% and for non-depressed samples it is 80.62% with overall accuracy of 74.64%. From this it is concluded that as kernel sizes increases accuracy in prediction increases. Comparative analyses of two proposed models are shown through bar graph in Fig.6. End to end convolutional neural network for depression detection achieved better performance than the spectrogram based convolutional neural network model for both the classes.

Table 2 Model configurations and performance characteristics of end to end convolutional neural networks

Sl.No	Model	Layer	No of kernels	Kernel size	Max pooling size	Class	Recall(%)	Precision(%)	F-Score(%)
1	Model1	Layer1	64	8	10	Depressed	54	59	57
		Layer2	128	6	8				
		Layer3	256	6	8	Non-depressed	64	59	61
2	Model2	Layer1	64	16	18	Depressed	86	60	71
		Layer2	128	6	8				
		Layer3	256	6	8	Non-depressed	45	77	57
3	Model3	Layer1	64	16	18	Depressed	96	61	75
		Layer2	128	12	14				
		Layer3	256	12	14	Non-depressed	40	93	56
4	Model4	Layer1	64	24	26	Depressed	69	65	67
		Layer2	128	12	14				
		Layer3	256	12	14	Non-depressed	63	68	66
5	Model5	Layer1	64	32	34	Depressed	75	74	74
		Layer2	128	12	14				
		Layer3	256	12	14	Non-depressed	61	83	70
6	Model6	Layer1	64	32	34	Depressed	74	79	77
		Layer2	128	18	20				
		Layer3	256	12	14	Non-depressed	80	76	78

Results of Existing works on depression detection using AVEC dataset are tabulated in Table 3. In this work Xingchen Ma et al.[14] proposed an innovative deep learning method where Convolutional neural networks (CNN) and Long short term memory (LSTM) are used for learning features and classifying the speech into depressed and non-depressed states, pre-processing is performed by removing silence regions, spectrograms and MFCC coefficients which are low level features were calculated and experiments are performed on AVEC 2016 DAIC-WOZ dataset and achieved F1 score of 0.52 for depressed and 0.7 for non-depressed. The author Jingwen Zhang proposed AdaBoost framework and collaborative representation (AdaBoost-CRC) for depression detection[23].The AdaBoost framework improved the recognition accuracy of CRC for the positive samples in the case of data insufficiency and imbalance. Mel Frequency coefficients (MFCC's) are extracted and performed classification using adaboost CRC classifier. Results are validated on AVEC 2013 dataset and obtained accuracy of 66.66% and F-score of 58.35%.



(a) (b)

Fig.6. Bar graph showing comparison of performance metrics of spectrogram based CNN model and proposed end to end model (a) depressed state (b) non-depressed state

Table 3 Existing system on Depression detection using AVEC dataset

Sl.No	Authors	Methodology	F-Score(%)
1	Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, Yunhong Wang [14] (2016)	Spectrogram based convolutional neural networks and long short term memory(LSTM)	52
2	Jingwen Zhang, Haochen Yin, Jinfang Wang, Shuxin Luan, Chang Liu [23] (2018)	AdaBoost framework and collaborative representation (AdaBoost-CRC)	58.35
3	Proposed Work	End to End CNN Model	78

Comparison of existing work with proposed work indicates that improvement in accuracy by 8% and F-score by 19%. Till now classification is has been performed to classify classes into depressed and non-depressed states. In this work additionally categorization of depressed states is proposed and is implemented on Model 6 which is illustrated in next section.

5.3 End to end CNN model for different levels of depressed states

Here using the PHQ-8 scores of speech samples from dataset have classified depression into three stages first stage (PHQ-8 score- 10 to 13), intermediate stage (PHQ8 score- 14 to 16), final stage (PHQ8 score- 17 to 20).Number of samples utilized are non-depressed (1172 samples), depressed 1st stage (895 samples), depressed intermediate stage (253 samples) and depressed final stage (170 samples).Performed experiments on Model 6 as mentioned in Table II. The performance characteristics of this model is tabulated in Table 4. Performances of non-depressed state and first stage depressed state are better when compared to other two states with accuracy of 92% for non-depressed, 62% for depressed first stage where as it is 34 % for depressed intermediate stage and 8% for depressed final stage and primary reason for this is number of data samples available for depressed intermediate and final states are less when compared to other two states. So Model is underfitted. To achieve higher performance requires more data samples for each class.

Table 4 Performance characteristics of the three staged depressed end to end model

Sl.No	Class type	Recall(%)	Precision(%)	F-Score(%)
1	Non-depressed	92	65	76
2	Depressed 1st stage	62	81	7
3	Depressed intermediate stage	34	69	45
4	Depressed final stage	8	75	15

6. Conclusion and Future Work

An investigation is carried out on depression detection using spectrogram based convolutional neural network and end to end convolutional neural network models. Parameter tuning has been performed and comparative analysis has been carried out between two models and best model has been chosen for categorizing the depression state. Validation has been performed on AVEC 2016 DAIC-woz dataset. Performance of end to end model is better than the baseline

models and spectrogram based convolutional neural network model. Future work includes optimization of the model through selection of best sub parameters of convolutional layer, max-pooling layer to achieve higher accuracy rates and to generalize the model. Dataset for depression needs to be enhanced and to include datasets from different languages for implementing language independent models. Also work needs to be done on multi-scale convolutional network to learn features at different time scale and frequencies from speech samples.

References

[1] US Department of Health and Human Services, Healthy People 2010:Understanding and improving health, vol. 2, US Government Printing Office, Washington, DC, 2000.

[2] J C Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie and D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," J. Neurolinguist., vol. 20, pp. 50-64, 2007.

[3] E. Moore, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech," IEEE Trans. on Biom. Eng., vol. 55, no. 1, pp. 96–107, 2008.

[4] L. S. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents," in 2010 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), 2010, pp. 5154–5157.

[5] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An Investigation of Depressed Speech Detection: Features and Normalization," in Interspeech2011, 2011, pp. 2997–3000.

[6] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Ying, N. Minh Hoai, M. T. Padilla, Z. Feng, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in 3rd Int. Conf. on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009., pp. 1–7.

[7] S. Scherer, G. Stratou, M. Mahmoud, and J. Boberg, "Automatic Behavior Descriptors for Psychological Disorder Analysis," IEEE Conf. on Automatic Face and Gesture Recognition 2013, p. NA, 2013.

[8] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Eye Movement Analysis for Depression Detection," in 2013 IEEE Int. Conf. on Image Processing ICIP2013, Melbourne, Australia, 15-18 Sep 2013, 2013.

[9] Hailiang Long, Zhenghao Guo, Xia Wu, Bin Hu, Zhenyu Liu, Hanshu Cai,"Detecting Depression in Speech: Comparison and Combination between Different Speech Types", IEEE International Conference on Bioinformatics and Biomedicine, 2017.

[10] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Tom Gedeon, Michael Breakspear, Gordon Parker, "A Comparative study of different classifiers for detecting depression from spontaneous speech",ICASSP, 2013.

[11] Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo, "A Study of Acoustic Features for the Classification of Depressed Speech", International Convention on Information and Communication Technology, Electronics and Microelectronics, 2014.

[12] Veena Narayanan, S Lalitha, Deepa Gupta,"Stress Recognition using Auditory Features for Psychotherapy in Indian Context", ICCSP,2018.

[13] Veena Narayanan, S Lalitha, Deepa Gupta,"An epitomization of stress recognition from speech signal",International Journal of Engineering & Technology, 7 (2.27) (2018) 61-68

[14] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, Yunhong Wang,"DepAudioNet: An Efficient Deep Model for Audio based Depression Classification," AVEC' 16 Proceedings of the 6th International Workshop, 2016.

[15] S.Lalitha, Shikha tripathi, Deepa gupta,"Enhanced speech emotion detection using deep neural networks", International Journal of Speech Technology, 2018.

[16] Panagiotis Tzirakis, Jiehao Zhang, Bj¨orn W. Schuller, "END-TO-END SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORKS",ICASSP, 2018.

[17] Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D, Marsella S, Traum DR. The Distress Analysis Interview Corpus of human and computer interviews. InLREC 2014 May (pp. 3123-3128)

[18] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D.Lalanne,M.Torres,S.Scherer,G.Stratou,R.Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotionrecognitionworkshopandchallenge," inAVEC Workshop. ACM, 2016, pp. 3–10.

[19] https://www.nch.com.au/wavepad/index.html

[20] https://www.spyder-ide.org/

[21] A.I. Diveev, S.V. Konstantinov,E.A. Sofronova,"A Comparison of Evolutionary Algorithms and Gradient-based Methods for the Optimal Control Problem," International Conference on Control, Decision and Information Technologies,2018.

[22] N. Srivastava, G. E Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," Journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] Jingwen Zhang, Haochen Yin, Jinfang Wang, Shuxin Luan, Chang Liu, "Severe Major Depression Disorders Detection using AdaBoost-Collaborative Representation Classification Method", International Conference on Sensing, Diagnostics, Prognostics, and Control,2018.

[24] G. Canbek, S. Sagiroglu, T. T. Temizel, and N. Baykal, "Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights," in 2017 International Conference on Computer Science and Engineering (UBMK), 2017, pp. 821–826.