Hybrid CNN-SVM classifier for efficient depression detection system

Afef Saidi,Slim Ben Othman, Slim Ben Saoud
Advanced Systems Lab - Tunisia Polytechnic School, University of Carthage, BP 743 - 2078 - La Marsa, Tunisia
afef.saidi@ept.run.tn, slim.benothman@ept.rnu.tn, slim.bensaoud@gmail.com

2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET) | 978-1-7281-6356-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/IC_ASET49463.2020.9318302

***Abstract*— Depression is a serious debilitating mental disorder affecting people from all ages all over the world. The number of depression cases increases annually in a continuous way. Due to the complexity of traditional techniques based on clinical diagnosis, there is a need for an automatic detection system of the depression. In this paper we present a novel audio-based approach to automatically detect depression using hybrid model. This model combines convolutional neural networks (CNN) and support vector machines (SVM), where SVM takes the place of the fully connected layers in CNN. In this proposed model, the features are automatically extracted using CNN and the classification is done using the SVM classifier. This approach was evaluated using DAIC-WOZ dataset provided by AVEC 2016 depression analysis sub-challenge. Experimental results showed that our hybrid model achieved an accuracy of 68% which outperform the CNN model (58.57%). Compared to the previous audio-based works using the same DAIC-WOZ dataset, our work showed a significant improvement in terms of accuracy, precision and recall.**

***Keywords*— Depression , CNN, SVM, hybrid model, Classification**

I. INTRODUCTION

According to the World Health Organization (WHO), depression is a common mental illness which affect about 300 million people worldwide (4.4% of the world's population) [1]. The number of affected people by depression is in a continuing increase [2]. A large number of these people abstain from seeking treatment either due to the high costs or to a lack of awareness for the signs of depression.

Traditional methods for depression diagnostic usually are based on clinical interviews that require an extensive assistance of experts and the active involvement of the depressed people [3].Therefore, several research efforts are seeking for the development of an automated detection system for depression. These approaches are based on machine learning which rely on feature engineering for its success.

In this paper we propose a novel approach to classify whether a person is depressed or not using only audio data. This approach presents an alternative solution based on hybrid CNN-SVM model to improve the classifier's performance. In this hybrid model, CNN works as a trainable feature extractor and SVM performs as a classifier.

In order to evaluate the performance of such approach a comparison between the proposed hybrid CNN-SVM model and CNN model is investigated in this paper.

The rest of this paper is organized as follows: A related work for the depression detection is reviewed in section 2. The basic concepts of CNN, SVM and our proposed hybrid CNN-SVM model for depression detection are introduced in Section 3. The experimental evaluation of our proposed approach and the results are given in Section 4. Finally, conclusions are presented in Section 5.

II. BACKGROUND AND RELATED WORK

Some studies have investigated depression detection by means of hand-crafted features. These features are the key success of the traditional classifiers where the features are manually extracted from raw data. The design of these features is often a tedious task as it needs the test and selection to obtain a good accuracy [4].

Tasnim et al. [5] explored the depression detection using different machine learning algorithms (Random Forests, SVM, Gradient Boosting Trees, Deep Neural Network) and the acoustic features in voice. A set of 2268 audio features was extracted from AVEC 2013 dataset. These features were reduced to 791 features using Principal Component Analysis (PCA). Using these features for classification task, it was found that deep neural network gives the best results. While for regression task, the best results were given by Random Forest. In [1], Stepanov et al. aimed to extract features from different modalities (speech, text, video) to estimate the depression severity. The obtained results showed that comparing to other modalities, behavioral characteristic features extracted from speech gives the lowest mean absolute error (MAE). The work in [3] explored two approaches for feature extraction for the depression detection based on hand-crafted features (statistic features) and CNN model. For depression detection and severity estimation, both techniques outperform the previous works.

On the other hand, other approaches have relied on deep neural networks models to extract features. This technique aims to extract the most discriminative features from the input data in an implicit and automatic way.

In [6], Acharya et al. presented a depression detection model using EEG signal and CNN model. The evaluation of the proposed approach using EEG signals from the left and right hemisphere achieved respectively accuracies of 93.54% and 95.49 %. Despite the fact that using EEG signal for depression detection may give high performance, the process of data collection and processing can be tedious and complex. Also,

978-1-7281-6356-7/20/$31.00 ©2020 IEEE

considering that relevant insights may be afforded from audio data [5] other researches have used it to detect depression. Yang et al. [7] proposed a hybrid framework for the classification and estimation of depression from audio, video and text descriptors. This framework is based on the fusion of DCNN-DNN models. The evaluation of the proposed model using the development and test set achieved respectively a MAE with 2.477, 4.359 which are better than the baseline results. The authors in [8] suggested an audio-based method for depression classification named DepAudioNet. This method combined CNN and Long Short-Term Memory (LSTM) to capture the short term and long-term temporal and spectral correlations of the audio representation. Experimental results showed that the suggested approach achieved a superior result to the baseline (1.2× improvement in F1-score).

III. Proposed Depression Dection System

In this section, we first briefly introduce our approach. Then we present the theory of CNN and SVM classifiers. Lastly, we describe our proposed hybrid CNN–SVM model for the depression detection.

A. Approach overview

While other researches focus on semantic features in the audio signals to predict depression [9], this work addresses the detection of depression from acoustic features in speech (such as pitch, rhythm, intensity, articulation, etc.).

To analyze acoustic features of speech, a first step is the segmentation of the person's speech. This step allows to remove silence, noise and other speakers.

In order to extract acoustic features in speech, there are some hand-crafted methods such as Mel-frequency cepstral coefficients (MFCCs), chroma vectors, zero crossing rate, etc. These techniques extract short-term and mid-term audio features. However, these features provide lower level representations of speech [10]. Hence, some features in the speech of depressed subjects can be not detected.

Unlike these methods, spectrograms provide a high-level representation of speech. In this work we aim to represent the audio data via a spectrogram which will be fed as input to CNNs as it requires images.

To obtain these spectrograms, a short-time Fourier transform (STFT) is performed on the segmented audio data. This will result on a matrix representation of a spectrogram where the frequency and time are represented in the vertical and horizontal axis, respectively. The frequency intensity at a particular time is represented by a value in the matrix.

After this data preprocessing, the input data (spectrograms) is given to CNN as an automatic feature extractor.

SVM was used instead of the fully connected layers as a classifier. It received the extracted features from CNN as a feature vector to perform the classification.

The overall process of our proposed approach is detailed in Fig. 1.

Fig. 2 (b) shows the general workflow of our proposed approach to detect the depression. This approach is compared with the baseline model (a) based on CNN.

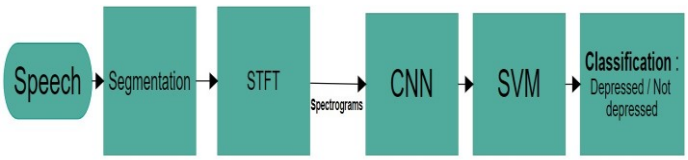

Fig. 1. Block diagram of our approach

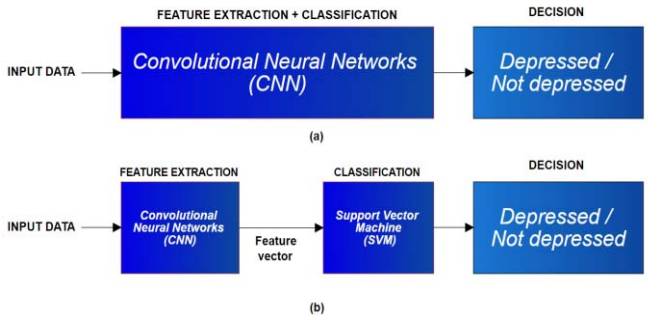

Fig. 2. Workflow of the baseline model (a) , (b) General workflow of the proposed approach

B. Convolutional Neural Networks

Convolutional Neural Networks (CNN) is composed of two parts: an automatic feature extraction part and a classification part.

A typical architecture of CNN consists mainly of three types of layers [11].

- Convolution layers: in this layer a convolution operation to the input data is done using convolution kernels (filters) resulting a feature map.
- Pooling layers: the goal of this step is to keep as much relevant features as possible while reducing the spatial dimensions of the input for the next convolutional layer. The most used pooling functions are Max and Average (max / avg pooling).
- Fully connected layers: it converts the 2D feature maps into a 1D feature vector after several iteration of convolution and pooling. This vector will classify the extracted features.

As shown in Fig. 3, the architecture of the CNN model used in this work consists of 7 layers: 2 convolutional layers (CV) with max-pooling (MP) and 3 fully connected layers (FC). The CNN architecture was inspired from the work in [12].

Each input is an image with a dimension of 513x125 which represent 4 seconds of audio.

The two convolution layers consists of 32 filters of size 3x3 resulting in a 32 feature maps after each layer followed by a ReLU activation function.

To reduce the dimensionality of the feature maps, each convolution layer is succeeded by a max-pooling layer

respectively with (4x3) filter, (1x3) stride and (1x3) filter, (1x3) stride.
Subsequently, three dense layers followed by a dropout layer of 0.5 are used. These dense layers have respectively 512, 256 and 512 neurons.
At last, to predict whether the output class is depressed or not depressed a softmax function is utilized.
The CNN model was trained with a batch size of 64 and using adadelta optimizer.

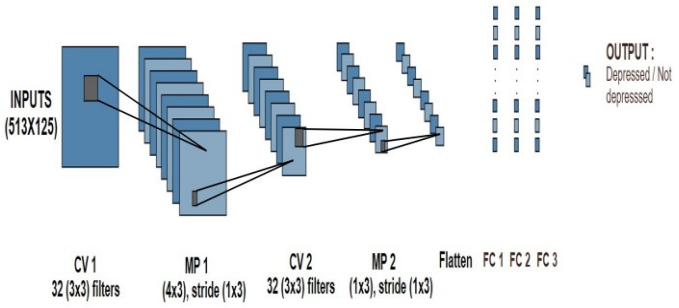

Fig. 3. CNN architecture

C. Support Vector Machines

SVM is a binary classification algorithm (for binary classification problems) and a form of linear classifiers. It can solve problems with small samples due to its ability of generalization [13].
The principle of SVM is to find a linear separator of two data classes or hyperplane with the maximum width (also called margin) using a nonlinear transformation Φ (equation 1). This margin is the distance between the separation boundary and the closet data points. These latter are called support vectors, they are used for the determination of the hyperplane (Fig. 4).
This transformation was carried out by kernel functions like:

- Linear kernel
- Polynomial kernel
- Sigmoid kernel
- RBF kernel (Radial Basis Function)

$$f(x) = WT\,\Phi(x) + b \qquad (1)$$

Where:

- W ϵ Rn,
- b ϵ R,
- Φ(x) is a feature map.

As the performance of SVM model depends on its hyper-parameters among them the kernel choice, in our work we used the Grid search method to find the best combination of these hyper-parameters to build our SVM model.

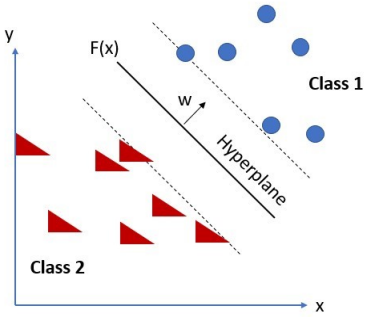

Fig. 4. Principle of SVM

D. Hybrid CNN-SVM model

In this section, we describe the architecture of our depression detection system based on hybrid CNN-SVM (Fig. 5).

After the training of CNN, the fully connected layers are replaced by SVM classifier to perform classification. SVM take the outputs from CNN's hidden units as a feature vector for the training step. The classification step is performed by SVM on the test set using the automatically extracted features.

The process of the CNN-SVM combined model can be summarized as follows:

1. The input data is fed to the CNN classifier for the training.
2. A feature vector can be automatically extracted, after the training of CNN.
3. The fully connected layers of CNN are replaced by SVM classifier which will be trained using the automatically extracted feature vectors.
4. In the test step, a given input map is delivered to CNN to obtain a test feature vector.
5. The classification is done by SVM classifier using the given test feature vector.

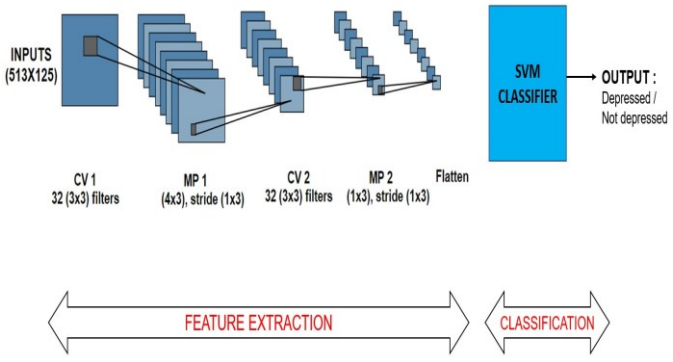

Fig. 5. Architecture of the hybrid CNN-SVM model for depression detection

IV. Experimental evaluation

To assess the effectiveness of our proposed depression detection system, we perform our experiments on the DAIC-WOZ dataset.

In this section we will introduce the dataset, experimental protocols, and discuss the obtained results.

A. Dataset

DAIC-WOZ database is a part of the corpus Distress Analysis Interview Corpus (DAIC). It was compiled by USC Institute of Creative Technologies and provided as part of the 2016 Audio/Visual Emotional Challenge and Workshop (AVEC 2016) [14]. It was designed as a support for the diagnostic of psychological distress conditions, such as anxiety, depression and post-traumatic stress disorder.

This database provides audio, video recordings and psychiatric questionnaire responses (PHQ-8) in text format.

It contains audio recordings of 189 participants. These participants undergo clinical interviews conducted by a virtual interviewer named Ellie. The range of each interview is between 7-33 minutes.

The dataset was split into 3 sets: training set (107 subjects, 57%), validation set (35 subjects, 19%) and test set (47 subjects ,25%). However, the provided test set is not labelled. Thus, we used the validation set for the test.

Some interview clips of participants are troubled and cannot be segmentable because of volume levels, proximity to the interviewer, etc. This result in a total of 122 subjects (38 depressed and 84 non depressed).

B. Results

We carried our experimental studies so that we could explore the efficiency of our classification technique based on hybrid CNN-SVM for depression detection.
All experiments in this study were conducted on a laptop computer with Intel Core (TM) i5-8250U CPU @ 1.60GHz, and NVIDIA GeForce MX 150.
The dataset used in this work introduce a classification bias resulted from data imbalance between depressed and non-depressed samples. The number of non-depressed subjects is about three times larger than that of depressed ones.

To overcome this issue, the spectrograms of each participant were cropped into 4 second slices. Then, participants were randomly sampled in equal proportion from each class (depressed, not depressed).

Fig. 6 exhibits an example of spectrogram of a depressed and non-depressed subject.

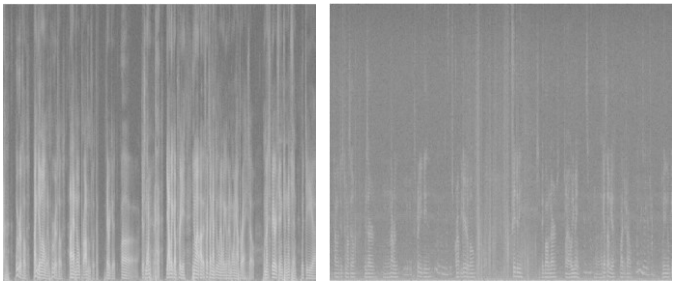

Fig. 6. Spectrograms (in grayscale) of a depressed subject (left) and non-depressed subject (right)

CNN model was trained on 40 randomly selected audio segments from 31 participants in each category (depressed / not depressed). The resulting data used for the training and test sets is presented in table 1.

TABLE I. Total data in the training and test sets

	Total data
Training set	2480
Test set	560

After the CNN training, a feature vector with 512 values was obtained. This feature vector was fed as input to SVM classifier for the training and testing phase.
Table 2 displays the test accuracies of CNN and hybrid CNN-SVM on depression detection using DAIC-WOZ dataset. Experimental results showed that our proposed model based on hybrid CNN-SVM for depression detection achieved an accuracy of 68% which outperforms the baseline CNN classifier (58.57 %).
The combination of CNN and SVM's advantages allow to achieve good results on depression detection while saving both time and effort in the feature extraction step.

TABLE II. Performance results

	CNN	**CNN-SVM**
Accuracy (%)	58.57	68

Table 3 and table 4 provide the prediction performance of both the baseline CNN model and our hybrid model through the confusion matrix table. This latter identifies the number of subjects (depressive and not depressive) correctly and wrongly classified.
For the baseline CNN model, 241 depressed subjects are correctly classified as depressed and 87 not depressed subjects are correctly classified as not depressed.
On the other hand, with the hybrid CNN-SVM model, 199 depressed subjects are correctly classified as depressed and 182 not depressed subjects are correctly classified as not depressed.

TABLE III. CONFUSION MATRIX OF BASELINE CNN MODEL

		Actual	
		Depressed	***Not depressed***
Predicted	***Depressed***	241 (TP)	193 (FP)
	Not depressed	39 (FN)	87 (TN)

TP: true positive, FP: false positive, TN: true negative, FN: false negative

TABLE IV. CONFUSION MATRIX OF BASELINE HYBRID CNN-SVM MODEL

		Actual	
		Depressed	**Not depressed**
Predicted	**Depressed**	199 (TP)	98 (FP)
	Not depressed	81 (FN)	182 (TN)

The performance of our proposed model is further evaluated by a comparison with other previous works. All these previous works are audio-based approaches that have used the same dataset (DAIC-WOZ database). In this evaluation, we employ other metrics than accuracy such as precision, recall and F1-score.

The comparison of our approach for depression detection with the previous related works is illustrated in table 5.

In terms of accuracy, our hybrid CNN-SVM model as well as our CNN model outperforms results reported in [5] with SVM, Random Forest and Gradient Boosting Tree (GBT) classifiers. However, in this work the accuracy obtained using Deep Neural Network (DNN) classifier is better than ours.

There are other metrics to evaluate the performance of the built model more accurately especially in the case of unbalanced dataset. Among these metrics, the precision, recall and F1-score [15]. These metrics takes into consideration the false positives (FP) and false negatives (FN) which are type I error and type II error respectively.

Regarding the precision, which measure the correctly classified depressed subjects from all the predicted depressed cases, our CNN model is better than [16], [5] and [8] respectively with 1.25×, 1.1 and 1.6×. Moreover, our hybrid model is better than these works respectively with 1.5×, 1.2× and 1.9×. This indicates that a significant part of depressed subjects has been correctly classified as depressed. Thus, proving the effectiveness of our approach.

Likewise, in terms of recall, which measures the correctly classified depressed cases from all the actual depressed cases, our CNN-SVM model is better than SVM, Random Forest, GBT and DNN classifiers used in [5] respectively with 1.2×, 1.1×, 1.2× and 1.3×.

Overall, considering the F1-score , which is the harmonic mean of the precision and recall, it can be seen that the F1-score of both the CNN and our proposed model achieved high values compared to the works in [16] and [8].

TABLE V. PERFORMANCE COMPARISON TO THE PREVIOUS WORKS

	Model	Accuracy	Precision	Recall	F1-score
[16]	1D CNN	--	0.44	0.78	0.56
[5]	• SVM • Random Forest • GBT • DNN	• 57.92 • 60.19 • 58.40 • 74.65	• 0.50 • 0.52 • 0.54 • 0.49	• 0.58 • 0.60 • 0.58 • 0.56	-- -- -- --
[8]	CNN-LSTM	--	0.35	1	0.52
Our work	CNN	58.57	0.55	0.86	0.67
	CNN-SVM	**68**	**0.67**	**0.71**	**0.69**

V. CONCLUSION

In this work, we have proved the applicability and the efficiency of the combined CNN-SVM model on depression detection using DAIC-WOZ database. In this hybrid model, the CNN was used to automatically extracting features and the SVM for depression detection (depressed/ not depressed). Comparing to the baseline CNN model, our proposed approach gives a better accuracy result with 68 %. This hybrid model offers not only the automatic feature extraction using the CNN, but also improved the classification accuracy through the combination of the SVM classifier.

Compared to the previous audio-based approaches, our proposed model achieved better results.

The performance of our proposed approach can be further improved by increasing the training samples or some fine tuning of the model's parameters.

In the future, we intend to improve the accuracy of our model by testing other architectures and to develop implementation approach based on embedded systems.

REFERENCES

[1] E. A. S. e. al., "Depression severity estimation from multiple modalities," in *IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2018, pp. 1-6.

[2] B. Li, J. Zhu, and C. Wang, "Depression severity prediction by multi-model fusion," in *HEALTHINFO 2018 : The Third International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing*, Nice, France, 2018, pp. 19- 24.

[3] L. S. a. M. V. S. Song, "Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features," in *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 158-165.

[4] L. A Nanni, S. Ghidoni, and S. Brahnam, "Handcrafted vs. non-handcrafted features for computer vision classification," *Pattern Recognition,* vol. 71, pp. 158-172, 2017.

[5] M. Tasnim and E. Stroulia, "Detecting depression from voice," Cham, 2019, pp. 472-478: Springer International Publishing.

[6] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, "Automated EEG-based screening of depression using deep convolutional neural network," *Computer Methods and Programs in Biomedicine,* vol. 161, pp. 103-113, 2018/07/01/ 2018.

[7] L. Yang *et al.*, "Hybrid depression classification and estimation from audio video and text information," presented at the Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, California, USA, 2017, pp. 45-51.

[8] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "DepAudioNet: An efficient deep model for audio based depression classification," presented at the Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, 2016, pp. 35-42.

[9] J. R. Williamson *et al.*, "Detecting depression using vocal, facial and semantic communication cues," presented at the Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, 2016, pp. 11-18.

[10] S. A. Alim and N. K. A. Rashid, "Some commonly used speech feature extraction algorithms," in *From Natural to Artificial Intelligence - Algorithms and Applications*, R. Lopez-Ruiz, Ed.: IntechOpen, 2018.

[11] H. Hu, B. Liu, and P. Zhang, "Several models and applications for deep learning," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, 2017, pp. 524-530.

[12] K. Kiefer. (2017). *depression-detect*. Available: https://github.com/kykiefer/depression-detect

[13] H. Wu, Q. Huang, D. Wang, and L. Gao, "A CNN-SVM combined model for pattern recognition of knee motion using mechanomyography signals," *Journal of Electromyography and Kinesiology,* vol. 42, pp. 136-142, 2018/10/01/ 2018.

[14] J. Gratch *et al.*, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014, pp. 3123-3128.

[15] M. Hossin, M. N. Sulaiman "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process (IJDKP),* vol. 5, no. 2, 2015, pp. 1-11.

[16] H. D. a. W. L. G. Lam, "Context-aware deep learning for multi-modal depression etection," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 3946-3950.