# Using Deep Neural Networks for Detecting Depression from Speech

Mirela Gheorghe[1], Serban Mihalache[1,2], Dragos Burileanu[1]
[1]*Speech and Dialogue Research Laboratory (SpeeD), University "Politehnica" of Bucharest, Romania*
[2]*Research Institute for Artificial Intelligence "Mihai Draganescu", Romanian Academy*
mirela.gheorghe@upb.ro, serban.mihalache@upb.ro, dragos.burileanu@upb.ro

***Abstract*— Diagnosing depression has become a major concern in the last years that led the research community to try innovative and creative ways of recognizing it. This paper proposes a system that can identify depression from speech samples based on a classification process performed by deep neural networks. The system was tested on the Distress Analysis Interview Corpus of human and computer interviews (DAIC-WOZ) and the Multi-modal Open Dataset for Mental-disorder Analysis (MODMA), two spontaneous speech databases recorded in realistic conditions and in two different languages. The system is able to generalize well on previously unseen data. Improvements have been obtained over other results reported in literature, yielding an unweighted accuracy (UA) of 91.25% and a weighted accuracy (WA) of 92.10% for DAIC-WOZ.**

***Keywords—depression detection, speech technology, deep neural networks, DAIC-WOZ, MODMA***

## I. Introduction

Clinical depression is a psychological disorder caused by difficulty coping with stressful life events, manifested by persistent feelings of sadness and negativity, low self-esteem, loss of interest or satisfaction. For people with severe symptoms, the risk of suicide may increase. In recent years, the number of reported cases of depression has grown significantly and represents a major cause of concern for the World Health Organization [1].

In the process of diagnosing depression, traditional assessment tools are used (e.g., the Beck Depression Inventory and the Patient Health Questionnaire depression scale, PHQ-8), which are based on the patients' perception of their own symptoms that they identify or are aware of, and on the experience of the clinicians who examine them. Consequently, the diagnostic process is subjective and requires time and practice to produce a reliable result [2]. Given this aspect, there is great interest among researchers in the fields of psychology, medicine, and computer science for automatic depression recognition [3].

So far, the best results have been achieved by combining multiple techniques and using machine learning methods. Researchers have used various techniques to analyze speech patterns, including acoustic and prosodic analysis or natural language processing. Acoustic analysis has been employed to extract features such as the fundamental frequency and the intensity, which have been found to be related to depression [4]. Prosodic analysis has been leveraged to study the rhythm and intonation of speech, which can provide insight into the emotional state of an individual [5]. Natural language processing has also been used to analyze the content of speech, such as lexical aspects and the topics discussed; depressed individuals tend to favor specific words and discuss certain topics more frequently [6]. Among the best machine learning models that have been reported in recent studies for depression detection from speech are Support Vector Machines (SVMs), Random Forests (RFs) [3, 7], Deep Neural Networks (DNNs) [6], and Convolutional Neural Networks (CNNs) [8-10].

Previous research has emphasized the existence of speech characteristics correlated with depression (e.g., reduced speech rate, pitch differences, increased duration of pauses, reduced tonal variation). Additionally, the acquisition of speech signals can be done remotely, without intrusion, and is relatively inexpensive. Identifying depression using speech is a simpler approach compared to alternative methods such as analyzing facial expressions, body language, brain activity, etc. Our proposed system uses features extracted from speech to train a deep learning model that would be able to assist specialists in faster, more accurate, and objective detection of depression symptoms and, consequently, in the selection of better and more effective treatments for potential patients. There are many challenges in creating a robust deep learning model and one of them is the lack of large, diverse datasets that include speech samples from individuals with depression, making it difficult to train accurate models. Recognizing depression from speech also raises important privacy and ethical concerns in terms of data security and informed consent and, because of this, most institutions are not able to obtain sufficient samples. Additionally, data annotation is based on the annotators' perception and is not fully accurate or objective, and the distribution of depression scores will affect the performance of the constructed model.

The main contributions of this work include:

- the development of a system for depression detection from speech using deep neural networks and algorithmically extracted features; and
- validating its high performance on two benchmark datasets that include speakers of two different languages (English and Chinese).

The rest of the paper is organized as follows: section II describes the proposed system architecture; section III provides methodological details for the experimental setup, as well as the achieved results and their interpretation; conclusions and future work are outlined in section IV.

## II. System Architecture

The proposed system for identifying depression from speech consists of a deep neural network that receives as input a large collection of descriptors obtained by using statistical functions on algorithmically extracted (also referred to as hand-crafted) acoustic, prosodic, spectral, and cepstral features. The system's final output is a binary classification that assigns the samples used in the training and testing phases to one of two classes: *depressed* or *non-depressed*. The proposed system diagram is illustrated in Fig. 1.

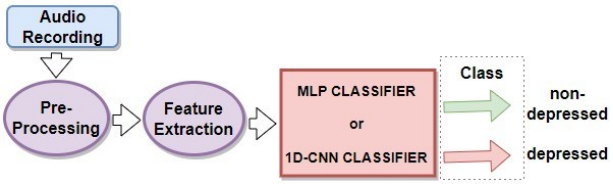


Fig. 1. The general diagram of the proposed system that solves a binary classification task to identify the class to which speech samples belong: *depressed* or *non-depressed*.

The pre-processing stage consists of resampling the audio recordings from the databases at 16 kHz and converting them to a mono PCM format. Additionally, the amplitude was normalized to the standard [−1, 1] range, the silent time intervals were removed, and a median filter was applied to reduce noise. The audio recordings were also divided into 25 ms frames with a 15 ms overlap using a Hamming window.

After pre-processing, several sets of features and their statistical measures (mean and standard deviation) were extracted for each audio recording. The first set, labeled 'MFCC', contains cepstral descriptors (the first 13 Mel-frequency cepstral coefficients and their delta and delta-delta coefficients) as they are among the most popular features used successfully in many speech processing tasks. The 'MBF' set comprises features obtained from the analysis of the instantaneous amplitude and frequency of speech signals considered as series of AM-FM micro-modulations. Another important set of features, 'APF', concerns acoustic and prosodic descriptors (loudness, rms energy, zero-crossing rate, pitch, jitter, shimmer, harmonic to noise ratio). The last set, 'SPF', includes a large number of spectral descriptors (spectral centroids, spectral flux, spectral spread, spectral skewness, spectral kurtosis, spectral slope, spectral entropy, spectral roll-off points, the low-frequency band energy, the high-frequency band energy, and the log-filterbank energies). A complete description of these features and the extraction process can be found in our previous work [11, 12].

Another aspect that required attention was the imbalance between the classes of the two databases, i.e., the existence of a significant difference between the number of observations belonging to each class (especially for the DAIC-WOZ database). The DAIC-WOZ dataset includes 43 *depressed* vs. 99 *non-depressed* speakers, while MODMA includes 22 *depressed* vs. 30 *non-depressed* speakers. The presence of class imbalance has important consequences for the learning process, typically producing classifiers that have poor predictive accuracy for the minority class and that tend to classify new samples as belonging to the majority class [13]. To overcome this problem, we applied SMOTE (Synthetic Minority Oversampling Technique) analysis to the features, an oversampling technique that generates synthetic samples from the minority class. Specifically, this involved duplicating samples from the minority class, despite these not adding new information to the model. The alternative, class weighting, did not produce improved results in preliminary experiments.

The last step in preparing the input data was applying z-score normalization to each descriptor, standardizing their distributions to achieve zero-mean and unity variance. The actual detection task is performed by the classifier, consisting of either a Multilayer Perceptron (MLP) model or a Convolutional Neural Network using one-dimensional filters (1D-CNN). As can be seen in Fig. 2, the MLP classifier uses $N_h$ hidden layers with $N_n$ neurons for each, and an output layer whose size is either equal to one (direct binary classification using the 'sigmoid' activation function) or is equal to the number of classes, i.e., two (indirect binary classification using the 'softmax' activation function). The 1D-CNN model is more complex and is presented in Fig. 3. The core structure is made up of three layers: 1D convolutional, followed by 1D max-pooling and batch normalization. This structure is then repeated $N_{layers}$ times, the classifier having a total number of $N_{layers} + 1$ convolutional layers. The output from the final block is then flattened and fed to a fully-connected layer (with the same configuration as in the case of the MLP model), with dropout being added to combat overfitting.

## III. Experimental Setup and Results

The experiments were implemented in Python using the Keras framework. Training and testing were done on a workstation with an Intel Xeon W-1290P CPU, 128 GB of RAM, and an Nvidia RTX A4000 GPU.

### A. Datasets

For the considered task, the only publicly accessible databases recorded in realistic conditions were DAIC-WOZ, the most often cited corpus, and MODMA, a new, promising dataset, released in 2020, that so far has not been used extensively by other researchers.

#### 1. DAIC-WOZ

The DAIC-WOZ (Distress Analysis Interview Corpus of human and computer interviews) database [14-15] was designed to simulate the standard protocols for identifying people at risk for post-traumatic stress disorder (PTSD) and major depression. This comprises the Wizard of Oz (WOZ) interviews that were conducted with the help of an animated virtual character named Ellie, who asked questions while being controlled by a human investigator from another room.

A total of 189 recordings are available, with durations between 5-10 min, which have been standardly split into 107 for the train set, 35 for the validation set and 47 for the test set. The participants (87 female, 102 male) were all fluent English speakers. For the train and validation sets, information regarding some PHQ8 scores (No interest, Depressed, Sleep, Tired, Appetite, Failure, Concentrating, Moving) was provided. The higher these scores, the more likely to encounter a depressed person. In the test sets, only the gender information of the interviewees is given, but the label on their depressive status is not provided. As a result, only 142 recordings from the total available could be used, namely those from the training and validation sets.

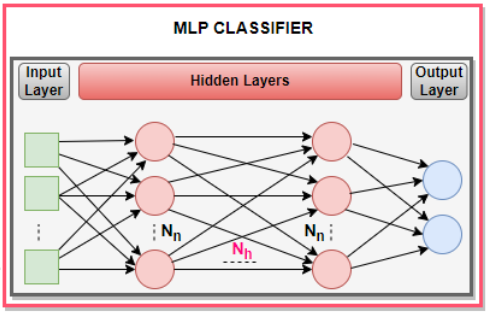


Fig. 2. The structure of the MLP model used to classify the input data as coming from a *depressed* or a *non-depressed* person.

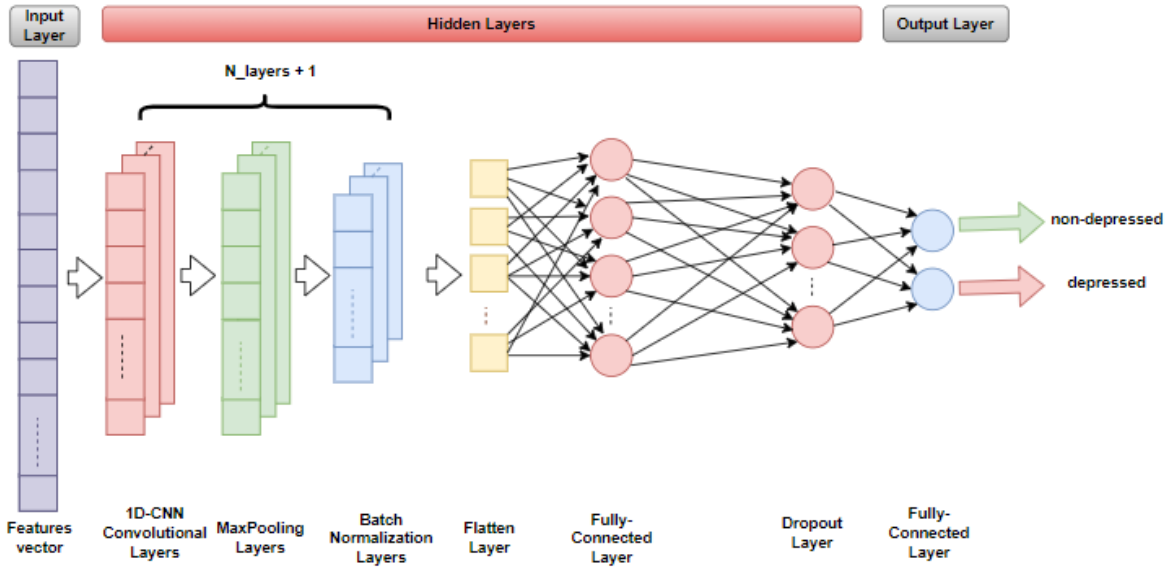


Fig. 3. Detailed overview of the 1D-CNN architecture containing the input layer, one or more groups of convolutional layers, pooling layers, and batch normalization layers, the flattening layer, one fully-connected layer with dropout, and the output layer (comprising fully-connected neurons). The input layer receives the feature vectors extracted from the speech signal. The convolutional layers apply 1D filters to the data, producing sets of feature maps. The pooling layers then perform down-sampling on the feature maps, reducing their spatial dimensions and increasing their invariance to small translations. The flattening layer converts the final multi-dimensional feature maps into a 1D vector. Lastly, the fully-connected layers output a prediction based on the learned features.

### 2. *MODMA*

The MODMA (Multi-modal Open Dataset for Mental-disorder Analysis) dataset [16] includes EEG data and speech recordings from clinically depressed patients and from a control group. To ensure the reliability of the data concerning depression, the patients were diagnosed and recommended by at least one clinical psychiatrist, with the score of the Patient Health Questionnaire (PHQ-9) being greater than or equal to 5, and the Mini-International Neuropsychiatric Interview (MINI) reaching the criteria for depression.

Audio recordings of 52 Chinese speakers were created during interviewing, reading, and picture description. The subjects were 22 patients (11 female, 11 male) diagnosed with depression and 30 healthy controls (9 female, 21 male). For every participant there are 29 different audio files distributed as follows: 1-18 are interviews, 19 is text reading, 20-25 are vocabulary reading, 26-28 are picture description, and 29 is a thematic apperception test. Each file has a duration of approximately 10 s. In our experiments, only the interview recordings were considered, as they are more relevant.

## B. *Setup and details*

The MLP classifier consists of fully-connected or dense layers. Several configurations for the hyperparameters were tested: the number of hidden layers (between 1 and 4), the number of neurons per each hidden layer (32, 64, 128, 256, 512, and 1024), the number of neurons in the output layer (a single neuron, using the 'binary cross-entropy' loss function, or 2 neurons, using the 'categorical cross-entropy' loss function), the activation function for the neurons in the hidden layers ('tanh', 'sigmoid', and 'relu'), using different dropout rates (none, 10%, 20%, 30%, 40%, and 50%), the optimization algorithm ('sgd', 'rmsprop', 'adagrad', 'adadelta', and 'adam'), and the rate used for L1 regularization (0.1, 0.01, and 0.001).

The 1D-CNN model has 2, 3, or 4 ($N_{layers}$ + 1) groups of 1D convolutional layers, 1D max-pooling layers, and batch normalization layers. Max-pooling was used to reduce the spatial dimensions of the feature maps produced by the convolutional layers, which helps to decrease the computational cost of the network. Batch normalization was included to help stabilize the distribution of the activations across different layers in the network. The hyperparameters chosen include: the number of filters (32 and 64), the kernel size (1 and 3), the pooling size (none and 2), and the number of neurons in the final fully-connected layer (20, 32, 64, 128, 256, 512, and 1024). Additionally, the same optimization algorithms, activation functions, dropout rates, and output layer configurations were tested as for the MLP model.

To establish the robustness and performance of the tested models as objectively as possible, the input data was divided into training (60%), validation (20%), and testing (20%) sets, while also ensuring the same proportional distribution over genders and classes. Furthermore, 10-fold cross-validation was employed to make sure the splitting is unbiased, and the performance metrics for each fold were calculated and then averaged, obtaining the final experiment-level values.

The metrics used were: unweighted accuracy (UA), weighted accuracy (WA), precision (P), recall (R), and the F1-measure. For both types of classifiers, training was done using a batch size of 32, over 100 epochs and employing the early stopping technique by monitoring the validation error with a latency of 3 epochs.

## C. *Results and discussions*

As can be seen in Table I and Table II, the cepstral ('MFCC') and the modulation-based feature sets ('MBF') proved to be the most effective, in contrast to the acoustic, prosodic, and spectral ones ('APF', 'SPF'). From the same tables, we can observe that the proposed models are robust and achieve similar results on the two databases, despite the fact that the audio recordings come from speakers of different languages (English and Chinese), which suggests that the models exhibit language independence.

TABLE I. PERFORMANCE METRICS VS. FEATURE SETS OBTAINED FOR THE DAIC-WOZ AND MODMA DATASETS USING THE MLP CLASSIFIER

| MLP classifier | | | |
|---|---|---|---|
| **Database** | **Feature set** | **Hyperparameters** | **UA / WA [%]** |
| DAIC-WOZ | 'MFCC' | - hidden layers: 2<br>- neurons per hidden layer: 1024<br>- neurons in the output layer: 1<br>- activation function: 'relu'<br>- dropout rate: none<br>- optimizer: 'adam' | 86.25 / 70.59 |
| | 'APF' | - hidden layers: 3<br>- neurons per hidden layer: 256<br>- neurons in the output layer: 2<br>- activation function: 'relu'<br>- dropout rate: 20%<br>- optimizer: 'rmsprop' | 75.73 / 87.33 |
| | 'MBF' | - hidden layers: 2<br>- neurons per hidden layer: 512<br>- neurons in the output layer: 1<br>- activation function: 'relu'<br>- dropout rate: none<br>- optimizer: 'adam' | 80.99 / 87.81 |
| | 'SPF' | - hidden layers: 2<br>- neurons per hidden layer: 512<br>- neurons in the output layer: 1<br>- activation function: 'relu'<br>- dropout rate: 20%<br>- optimizer: 'rmsprop' | 64.3 / 92.8 |
| MODMA | 'MFCC' | - hidden layers: 2<br>- neurons per hidden layer: 1024<br>- neurons in the output layer: 1<br>- activation function: 'relu'<br>- dropout rate: none<br>- optimizer: 'rmsprop' | 78.33 / 80.24 |
| | 'APF' | - hidden layers: 4<br>- neurons per hidden layer: 256<br>- neurons in the output layer: 2<br>- activation function: 'relu'<br>- dropout rate: 10%<br>- optimizer: 'adam' | 66.07 / 78.70 |
| | 'MBF' | - hidden layers: 2<br>- neurons per hidden layer: 1024<br>- neurons in the output layer: 2<br>- activation function: 'relu'<br>- dropout rate: 10%<br>- optimizer: 'rmsprop' | 77.22 / 81.79 |
| | 'SPF' | - hidden layers: 2<br>- neurons per hidden layer: 64<br>- neurons in the output layer: 1<br>- activation function: 'relu'<br>- dropout rate: 10%<br>- optimizer: 'adam' | 61.01 / 90.39 |

The best results for both datasets and both classifiers are summarized in Table III. The highest unweighted accuracy (UA) was achieved by the 1D-CNN model using the cepstral feature set, with slightly better results for the DAIC-WOZ database. The model used 4 convolutional layers with 64 filters, kernel size of 1, pooling size of 1 (i.e., no pooling), 1024 neurons for the fully-connected (FC) layer, a single neuron for the output layer, the 'adam' optimizer, the 'relu' activation function, and a dropout rate of 40%. Overall, for the DAIC-WOZ database, both classifiers performed well, with the 1D-CNN classifier having slightly better performance (91.25% vs. 83.2% F1-score). For the MODMA database, the 1D-CNN classifier again performed better than the MLP classifier (79.45% vs. 75.4% F1-score).

TABLE II. PERFORMANCE METRICS VS. FEATURE SETS OBTAINED FOR THE DAIC-WOZ AND MODMA DATASETS USING THE 1D-CNN CLASSIFIER

| 1D-CNN classifier | | | |
|---|---|---|---|
| **Database** | **Feature set** | **Hyperparameters** | **UA / WA [%]** |
| DAIC-WOZ | 'MFCC' | - convolutional layers: 4<br>- filters: 64<br>- kernel size: 1<br>- pooling size: 1 (no pooling)<br>- neurons in FC layer: 1024<br>- neurons in the output layer: 1<br>- dropout rate: 40%<br>- optimizer: 'adam' | 91.25 / 92.10 |
| | 'APF' | - convolutional layers: 3<br>- filters: 64<br>- kernel size: 1<br>- pooling size: 1 (no pooling)<br>- neurons in FC layer: 256<br>- neurons in the output layer: 2<br>- dropout rate: 30%<br>- optimizer: 'sgd' | 81.57 / 86.23 |
| | 'MBF' | - convolutional layers: 4<br>- filters: 128<br>- kernel size: 1<br>- pooling size: 1 (no pooling)<br>- neurons in FC layer: 20<br>- neurons in the output layer: 1<br>- dropout rate: 20%<br>- optimizer: 'adam' | 83.26 / 91.91 |
| | 'SPF' | - convolutional layers: 4<br>- filters: 32<br>- kernel size: 1<br>- pooling size: 1 (no pooling)<br>- neurons in FC layer: 64<br>- neurons in the output layer: 2<br>- dropout rate: 30%<br>- optimizer: 'sgd' | 78.07 / 81.27 |
| MODMA | 'MFCC' | - convolutional layers: 3<br>- filters: 32<br>- kernel size: 1<br>- pooling size: 1 (no pooling)<br>- neurons in FC layer: 512<br>- neurons in the output layer: 1<br>- dropout rate: 30%<br>- optimizer: 'rmsprop' | 79.16 / 92.53 |
| | 'APF' | - convolutional layers: 2<br>- filters: 32<br>- kernel size: 1<br>- pooling size: 1 (no pooling)<br>- neurons in FC layer: 20<br>- neurons in the output layer: 1<br>- dropout rate: 10%<br>- optimizer: 'sgd' | 66.36 / 81.36 |
| | 'MBF' | - convolutional layers: 4<br>- filters: 64<br>- kernel size: 1<br>- pooling size: 1 (no pooling)<br>- neurons in FC layer: 256<br>- neurons in the output layer: 2<br>- dropout rate: 40%<br>- optimizer: 'adam' | 84.16 / 86.74 |
| | 'SPF' | - convolutional layers: 2<br>- filters: 32<br>- kernel size: 1<br>- pooling size: 1 (no pooling)<br>- neurons in FC layer: 20<br>- neurons in the output layer: 1<br>- dropout rate: 10%<br>- optimizer: 'adam' | 54.16 / 94.38 |

Table IV compares the performance metrics of our proposed model to those of other systems reported recently in literature. For the DAIC-WOZ database, the proposed 1D-CNN model outperforms other CNN models, with our

TABLE III. DETAILED RESULTS CONCERNING THE BEST PERFORMANCE METRICS OBTAINED FOR THE DAIC-WOZ AND MODMA DATASETS.

| | | MLP classifier | | | | | 1D-CNN classifier | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Performance | | | | | Performance | | | |
| **Database** | **Feature set** | **Class** | **P [%]** | **R [%]** | **F1 [%]** | **UA / WA [%]** | **Class** | **P [%]** | **R [%]** | **F1 [%]** | **UA / WA [%]** |
| DAIC-WOZ | 'MFCC' | *depressed* | 70.0 | 87.5 | 77.0 | **86.25** | *depressed* | 87.5 | 87.5 | 87.5 | **91.25** |
| | | *non-depressed* | 94.0 | 85.0 | 89.4 | / | *non-depressed* | 95.0 | 95.0 | 95.0 | / |
| | | **Avg.** | **82.0** | **86.3** | **83.2** | **70.59** | **Avg.** | **91.3** | **91.3** | **91.3** | **92.10** |
| MODMA | 'MFCC' | *depressed* | 74.0 | 76.0 | 75.0 | **77.22** | *depressed* | 86.9 | 66.0 | 75.4 | **79.16** |
| | | *non-depressed* | 80.0 | 77.0 | 82.1 | / | *non-depressed* | 76.7 | 91.6 | 83.5 | / |
| | | **Avg.** | **77.0** | **76.5** | **78.6** | **81.79** | **Avg.** | **81.8** | **78.8** | **79.5** | **92.53** |
| | 'MBF' | *depressed* | 88.0 | 76.0 | 82.1 | **84.16** | *depressed* | 88.0 | 76.0 | 82.0 | **84.16** |
| | | *non-depressed* | 82.5 | 91.6 | 86.0 | / | *non-depressed* | 82.5 | 91.6 | 86.0 | / |
| | | **Avg.** | **85.3** | **83.8** | **84.1** | **86.74** | **Avg.** | **85.3** | **83.8** | **84.0** | **71.24** |

TABLE IV. PERFORMANCE COMPARISON BETWEEN THE BEST RESULTS ACHIEVED IN THIS WORK AND OTHER LITERATURE.

| **Database** | **Proposed model** | **UA / WA [%]** | **Other works** | **WA [%]** |
|---|---|---|---|---|
| DAIC-WOZ | 1D-CNN | **91.25 / 92.10** | [8] | 85.0 |
| | | | [9] | 71.0 |
| | | | [10] | 82.9 |
| MODMA | MLP | **84.16 / 86.74** | [3] | 83.4 |
| | 1D-CNN | 84.16 / 71.24 | | |

system achieving a weighted accuracy (WA) of 92.10%, the largest reported so far (to the best of our knowledge). For the MODMA database, the proposed MLP model performed the best, with the WA reaching 86.74%, surpassing other models based on decision trees (DT) [3] that only achieved a WA of 83.4%.

## IV. CONCLUSION AND FUTURE WORK

In this paper, we proposed and implemented a system for detecting depression from speech samples based on MLP and 1D-CNN deep learning models. The system was successfully tested on the DAIC-WOZ and MODMA databases, achieving significantly better results compared to other works recently reported in literature. Testing on datasets comprising recordings from speakers of two different languages achieved similar results, indicating robustness and good generalization capability for the system. The 1D-CNN classifier reached an unweighted accuracy (UA) of 91.25% and a weighted accuracy (WA) of 92.1% on the DAIC-WOZ dataset. The MLP classifier on the MODMA dataset achieved a WA of 86.74% and a UA of 84.16%. The results suggest a huge potential for a speech-based depression screening tool that could be used to assist healthcare professionals in the diagnosing and monitoring of patients, and to provide a scalable depression screening method enabling individuals to recognize their illnesses and seek professional help.

Our future work will focus on leveraging other types of features and exploring cross-lingual transfer learning, where knowledge learned from one language can be used to improve performance in another language.

## REFERENCES


[1] L. Albuquerque, A. R. S. Valente, A. Teixeira, D. Figueiredo, P. Sa-Couto, and C. Oliveira, "Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan," in PloS ONE, vol. 16, no. 4, Apr. 2021, pp. 1–20.

[2] L. Yang, D. Jiang, and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures," in IEEE Trans. on Affective Computing, vol. 12, no. 1, Mar. 2021, pp. 239–253.

[3] P. Wu, R. Wang, H. Lin, F. Zhang, J. Tu, and M. Sun, "Automatic depression recognition by intelligent speech signal processing: a systematic survey," in CAAI Transactions on Intelligence Technology, Jun. 2022, pp. 1–11.

[4] P. R. Parekh and M. M. Patil, "Clinical depression detection for adolescent by speech features", Proc. of the 2017 Int. Conf. on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp. 3453–3457, 2017.

[5] S. Prabhudesai, M. Parmar, A. Mhaske, and S. Bhagwat, "Depression detection and analysis using deep learning: study and comparative analysis," Proc. of the 10th IEEE Int. Conf. on Communication Systems and Network Technologies, pp. 570–574, Jul. 2021.

[6] Z. Huang, J. Epps, D. Joachim, and V. Sethu, "Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection," in IEEE Journal of selected topics in Signal Processing, vol. 14, no. 2, Feb. 2020, pp. 435–448.

[7] B. Yalamanchili, N. S. Kota, M. S. Abbaraju, V. S. S. Nadella, and S. V. Alluri, "Real-time acoustic based depression detection using machine learning techniques," Proc. of the Int. Conf. on Emerging Trends in Information Technology and Engineering, pp. 1–6, 2020.

[8] X. Miao et al., "Fusing features of speech for depression classification based on higher-order spectral analysis," in Speech Communication, vol. 143, no. 1, Aug. 2022, pp. 46–56.

[9] A. Vasquez-Romero and A. Gallardo-Antolin, "Automatic detection of depression in speech using ensemble convolutional neural networks," in Entropy, vol. 22, no. 6, Jun. 2020, pp. 688–705.

[10] Z. Huang, J. Epps, and D. Joachim, "Exploiting vocal tract coordination using dilated CNNs for depression detection in naturalistic environments," Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2020.

[11] S. Mihalache, D. Burileanu, and C. Burileanu, "Detecting psychological stress from speech using deep neural networks and ensemble classifiers," Proc. of the 11th Int. Conf. on Speech Technology and Human-Computer Dialogue (SpeD), pp. 74–79, Oct. 2021.

[12] S. Mihalache and D. Burileanu, "Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection," in Sensors, vol. 22, no. 3, Feb. 2022, pp. 1228–1249.

[13] J. Brownlee, "SMOTE for imbalanced classification with Python," MachineLearningMastery.com, 16-Mar-2021. [Online]. Available: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/. [Accessed: 18-Jan-2023].

[14] J. Gratch et al., "The distress analysis interview corpus of human and computer interviews," Proc. of the 9th Int. Conf. on Language Resources and Evaluation (LREC), pp. 3123–3128, May 2014.

[15] D. DeVault et al., "SimSensei kiosk: a virtual human interviewer for healthcare decision support," Proc. of the 13th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS), pp. 1–8, Jan. 2014.

[16] H. Cai et al., "MODMA dataset: a multi-modal open dataset for mental-disorder analysis," in Scientific Data Journal, vol. 9, no. 1, Feb. 2020, pp. 1–15.