International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 03 | Mar 2021 www.irjet.net p-ISSN: 2395-0072

A DEEP LEARNING APPROACH FOR DEPRESSION CLASSIFICATION USING AUDIO FEATURES

Himanshu Churi[1], Parul Keshri[2], Simran Khamkar[3], Prof. Amruta Sankhe[4]

[1,2,3]Dept. of Information Technology, Atharva College of Engineering, Mumbai.
[4]Assistant Professor, Dept. of Information Technology, Atharva College of Engineering, Mumbai.

Abstract - *Depression is the foremost cause of mental-health disorder worldwide. Major depressive disorder (MDD) is the most common mental disorder that affects psychological as well as physical status which could lead to loss of lives. The lack of tests for diagnosis as well as the subjectivity involved in detecting it, has resulted in growing interest to automate depression detection using conventional cues. Extensive and tedious clinical interviews are currently the primary methods to detect depression. Although helpful, this method lacks subjectivity and efficiency. Here in this paper, a deep learning-based model is put forward to aid clinics by classifying patients as depressed or non-depressed. Audio files transformed into spectrograms are fed to the Convolutional Neural Network (CNN) model to find auditory characteristics of possibly depressed patients. The DAIC-WOZ dataset used, contains a sample imbalance occurring due to the inequitable number of depressed and non-depressed data samples, which is eradicated by introducing random sampling before training the model. The proposed framework achieved an overall accuracy of 64.8%.*

***Key Words*: Depression, Spectrogram, Deep Learning, Convolution Neural Networks, PHQ-8, DAIC-WOZ.**

1. INTRODUCTION

Depression is one of the most common psychological disorders worldwide, affecting more than 264 million people [1]. Depressed individuals are more inclined to sadness, tension, low mood, and loss of energy, along with other symptoms. They find it extremely hard to concentrate on their work and communicate with other people. In extreme cases, as reported by, half of all completed suicides are related to depressive and other mood disorders [2]. Because of that, many researchers have focused on developing systems to diagnose and prevent this mental disorder to help psychiatrists and psychologists to assist patients as soon as possible [2].

Depression is never caused by just one thing. In other words, many factors such as biological factors, psychological factors, and stressful life could be the cause of depression. It has been identified as an economic burden as it impacts justice and social systems. In addition, major depression can lead to suicide and substance abuse. Therefore, detecting and treating depression and its severity is a top priority.

The traditional approach to diagnosing depression relies on conducting clinical interviews to screen candidates for depression [3]. However, these assessments depend to a large extent on the questions asked by the clinician, the verbal reports of patients, the behaviors reported by relatives or friends, and the mental status examinations, such as the Scale for the Assessment of Negative Symptoms (SANS), the Hamilton Rating Scale for Depression (HRSD) and the Beck Depression Inventory (BDI-II). They all make use of subjective ratings, and due to the lack of objective and quantitative measurements, their results tend to be inconsistent at different times or in various environments. Therefore, it becomes a necessity to develop Depression Classification methods for supporting the diagnosis of depression. Recently with the emergence of machine learning and artificial neural networks, several methods have been developed in recent years for supporting clinicians during the diagnosis and monitoring of clinical depression [3].

This paper proposes the use of deep neural networks for classifying depression using the audio features in the DAIC-WOZ dataset. In this research depression classification using speech is carried out in three stages which are pre-processing, feature extraction and classification. In the pre-processing stage, unwanted silence and the interviewers are removed from the audio signal. In the feature extraction stage, the audio features are extracted by processing on the provided Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) dataset and further spectrogram is generated which is fed to the neural network model. The remaining part of the paper is organized as follows: Section 2 reviews some of the existing approaches in the field of depression detection and section 3 briefs the overview of the system methodology. Section 4 highlights the depression detection process and section 5 highlights CNN and the Model Architecture. Finally, section 6 discusses the results and conclusions of the model.

2. RELATED WORK

Although there exist a large number of studies that dedicate to model the correlation between the emotional states and the properties of visual or vocal clues, the effort on depression classification (or depression prediction) is not that extensive. In this section, we briefly analyze the methods developed for automatic depression detection

roughly categorized into audio based, video based, and multi-modal based, according to the information adopted in the recent years.

Studies have shown that speech contains many peculiar characteristics for detecting a person's mental condition. Chlasta et al. [4] proposes a novel method for detecting depression in speech using deep convolutional neural networks. Experiments were carried out on Distress Analysis Interview Corpus (DAIC) and used ResNet-34 for classification. The results suggest that audio spectrograms are a promising screening feature for people with depression. And also, the use of short voice samples reduced the effect of noise.

In another study by Williamson et al. [5] explored two vocal tract representations, i.e., formant-frequency tracks and Mel-cepstral features, to encode the vocal tract resonant frequencies and spectral shape dynamics. With the two feature sets reflecting the changes in coordination of vocal tract motion associated with MDD, a Gaussian mixture model (GMM)-based multivariate regression scheme was then designed to make the final prediction. Later, they enhanced the approach of generating high level features from low-level features by a multi-scale correlation structure and timing feature sets. Dham et al. [6] utilizes text, audio and visual data in DAIC-WOZ database for depression detection. Gaussian Mixture Model (GMM) clustering and Fisher vector approach was applied to the visual data, low level audio features and head pose and text features. SVM classifier was applied separately on the extracted features. Finally, decision-level fusion was used to combine the results of different modalities [6].

Cohn et al. [7], who is a guide in the affective computing area, performed research where he fused both the visual and audio characteristics to include behavioral considerations, which are heavily related to mental disorders. Their findings suggest that the construction of an automatic system to recognize depression is possible, which will benefit clinical theory and practice. They used Manual FACS coding, active appearance modelling (AAM) and pitch extraction to measure facial and vocal expression. Classifiers using leave-one-out validation were SVM for FACS and for AAM and logistic regression for voice. The accuracy for detecting depression was 88% for manual FACS and 79% for AAM. Accuracy of vocal prosody was 79% [7]. These findings suggest the feasibility of automatic depression detection, raise new questions in automated facial image analysis and machine learning, and have compelling implications for clinical theory and practice.

Chao et al. [8], investigated the recent dominant deep learning models on the improvements of the audio-based methods and the vision-based ones in depressive disorder analysis. The extracted audio video features were firstly fused in feature level as representation of the abnormal behavior, and then the Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) was exploited to describe dynamic temporal information. They used multi-task learning to boost the performance and reported competitive results on the AVEC 2014 dataset [8]. This study indicated the promising future of exploiting the temporal information and both modalities in automatic depression detection. Jan et al. [9] combined features extracted from facial expressions and vocal expressions using deep learning and regression techniques respectively from naturalistic video recordings.

3. METHODOLOGY

The framework developed to categorize between depressed and non-depressed patients uses spectral features of speech and then analyses the emotions in it. Audio recordings (wav) format are trained using a Convolution Neural Network model. The dataset contains some troubled recordings which are needed to be filtered out. These files if not deprecated could cause an issue with the feature extraction process. The audio files contain the interviewers voice along with the interviewee's; thus, it is necessary to carry out segmentation in order to remove the interviewers voice along with other grot like silence, noise, etc. After all of the cleaning has been performed on the audio files, a spectrogram is generated and fed as an input to the convolutional neural network for feature extraction.

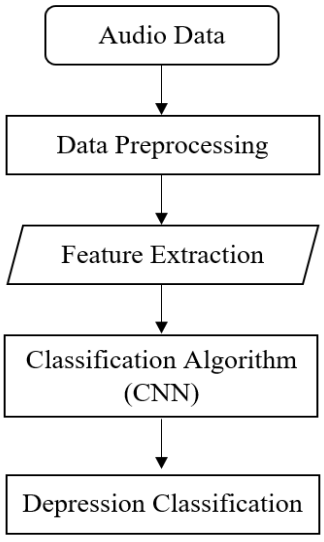


Fig-1: Flow Diagram of the Proposed Model

Fig. 1 shows the flow of the model where at first, audio data is pre-processed by accepting it as an input and then, detaching grot like unwanted voice, silence, noise, etc. with the help of pyAudioAnalysis library. Features are extracted from the audio files using the same library before converting the audio files into spectrograms. Finally, these images are fed into CNN, a classification algorithm. The proposed deep-based framework is presented with more details in the following contents.

4. DEPRESSION CLASSIFICATION MODEL

4.1 Dataset Description

The Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) dataset was used to conduct the experiments. It is a fraction of a large dataset called the Distress Analysis Interview Corpus (DAIC). DAIC contains clinical interviews to aid in the diagnosis of psychological disorders like anxiety, depression and post-traumatic stress disorder (PTSD) [10]. A computer agent assists the interview process by interacting with the people and identifying verbal and behavioral signs of mental-health issues. DAIC-WOZ contains audio, video files and extensive questionnaire responses of the interviews which have been transcribed and annotated. This part of the collection contains the Wizard-of-Oz interviews, conducted by an animated virtual interviewer named Ellie, who is controlled by a human interviewer remotely [10].

The database was published by the University of California for supporting the research purposes. The archive consists of 189 folders of participant's interview sessions. Each folder contains audio recordings, video notes, transcripts, etc. corresponding to each session of an interviewer's interview, one participant is only interviewed once. The audio was recorded at 16 kHz and each interview session lasted for 7-33 minutes, averaging about 16 minutes. Class labels have been allocated based on PHQ-8 score. If the PHQ-8 score is greater than 10, then the participant is depressed, else he or she is non-depressed.

The dataset is randomly separated into three parts namely training, testing and development parts. The distribution of the depressed to non-depressed participants is not well balanced, the ratio of non-depressed to depressed participants comes out to be 4, while on the other hand the gender distribution of the dataset is very well balanced.

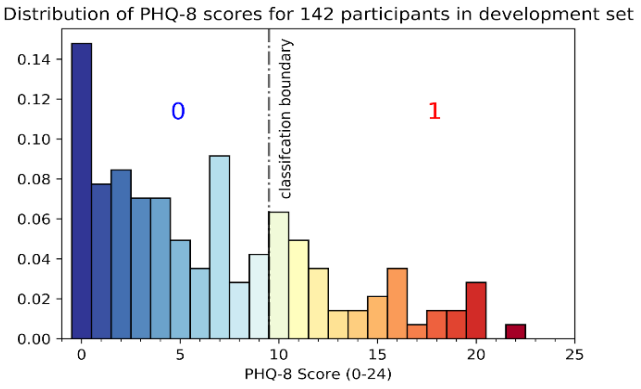


Fig-2: Distribution of PHQ-8 Scores

Distribution of PHQ-8 (24 level) scores as seen in Figure 2. Here, the thin, black, broken line indicates the boundary between the depressed and non-depressed scores of the participants/patients. This limit has been proposed by authors of questionnaires in [11] and they consider depressed patients with a greater or equal than 10 score.

In this paper, our main aim is to distinguish subjects if they are depressed or not using the audio recordings. According to the above paragraph and the Depression Classification Sub-Challenge (DCC) we have converted the level 24 PHQ-8 scores into binary 0 indicating not-depressed and 1 indicating depressed.

4.2 PHQ-8 Questionnaire

The Eight-Item Personal Health Questionnaire Depression Scale (PHQ-8) is in the spectrum of diagnostic tools used for measuring the severity of depressive disorders in large clinical studies [11]. As its name suggests, it is composed of eight multiple choice questions regarding a patient's level of distress, eating habits, and concentration. Answers are divided into four categories scoring from 0 to 3, respectively: not at all; several days; more than half days, and; nearly every day. The higher the value, the more likely the patient has the condition, and participants who achieve a PHQ-8 score equal to or greater than 10 are diagnosed as depressed [11].

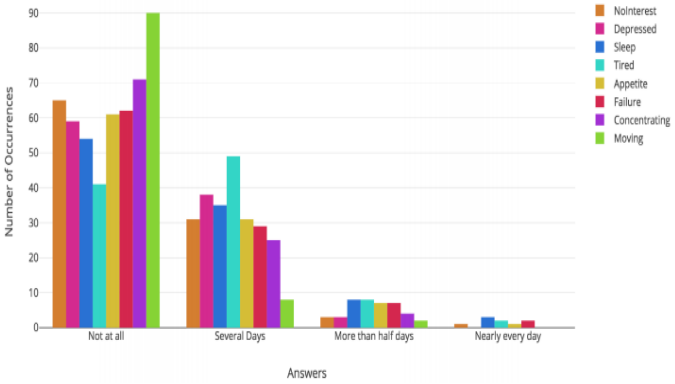


Fig-3: Distribution of questionnaire answers [12]

The distribution of all questionnaire answers is shown in Figure 3. Answers provide a broad scope of emotional and behavioral changes in the routine of patients. In the DAIC-WOZ dataset, interview participants have experienced more changes in sleep and tiredness [12]. In this paper, the final PHQ-8 score is used to identify when an individual is depressed or not.

4.3 Dataset Preprocessing

The data that has been received is raw and thus it needs to be pre-processed prior to feeding it as an input to the model. Two steps that are needed to be performed in order to clean data are Segmentation and Noise removal.

In the Segmentation at first, each and every audio file is taken and the interviewees voice is separated from other noises using the pyAudioAnalysis library available in python.

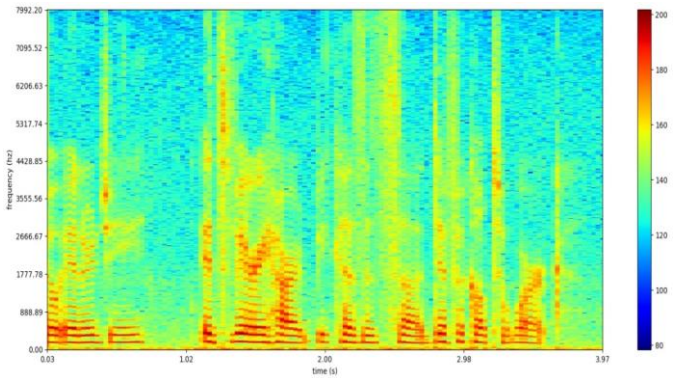
The interview recordings are separated into two categories: audio segments of the interviewers and segments of the interviewees. Further the audio segments of the interviewers are then ditched as it is computer generated, plain and emotionless, and we are only concerned about the interviewer's answers. Some of the audio files in which the interviewer had used separate microphones had very low background noise whereas the ones shot using cheap cameras or smartphones had a lot of it. In order to remove noise supplemental processing needed to be done. All of the noise removal was done using the pyAudioAnalysis library. After doing this, we have a clean and trimmed dataset with us which was used further in the feature extraction process.

4.3 Feature Extraction and Spectrogram Generation

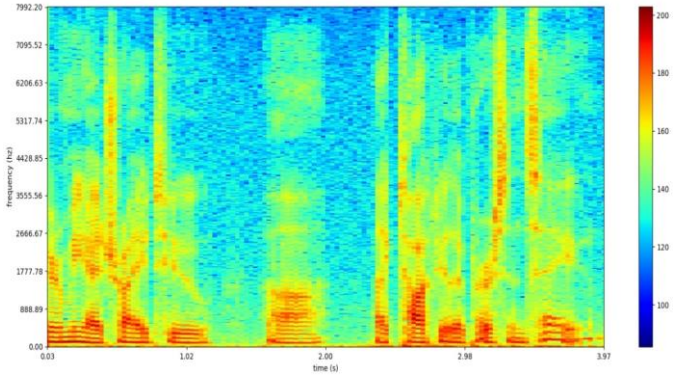
Extracting the correct features is crucial in order to obtain the desired result. There are numerous ways to extract auditory features, one of which include extracting short and midterm audio features such as Energy, zero crossing rate, chroma vectors, MFCCs, etc. After successful extraction these features serve as inputs to classification algorithms. These features mentioned above are low-level and have aroused concerns that they might go unnoticed in cases of depressed individuals.

Running a classification algorithm on the 34 short-term features resulted in an encouraging F1 score of 0.59. But a lot of research implementation has already occurred using such parameters, so we have just considered the results from these studies as a baseline. Instead, we use spectrograms, a visual representation of the audio files and feed it to a convolutional neural network (CNN) as they maintain high-level of detail as compared to others.

A spectrograph is a visual representation of the frequency spectrum of a signal that changes over time. The spectral features (frequency-based features), which are obtained by converting the time-based signal into the frequency domain using the Fourier Transform are used for analyzing the speech features. Spectral features play a significant role in analyzing the energy variations, amplitude etc. Spectrogram is generated for all the audio files in the enlarged dataset. An example of the generated spectrogram of the audio slice for depressed and non-depressed class is shown in Fig. 4. which our model uses as input.



a) Participant ID: 321 (Depressed class)



b) Participant ID: 482 (Non-Depressed class)

Fig-4: Example of Spectrogram input to the CNN of one audio slice for the (a) Depressed and (b) Non-depressed class respectively.

Spectrograms are created by performing the short-time Fourier transform (STFT) of each input audio signal. STFT is a short-term processing technique that breaks the signals, possibly overlapping frames using a movable window technique and calculates the Discrete Fourier Transform (DFT) at each frame. Each segmented audio clip sampled at 16kHz, is initially applied to a STFT using a hamming window length of 513 to generate a 2D matrix. The duration of each crop is S = 4 s and each spectrogram is represented as a matrix with dimensions F0 × T0, where the frequency F0 and the temporal T0 dimensions are respectively, 513 and 125.

5. CLASSIFICATION ALGORITHM

5.1 Convolution Neural Networks

CNN has proved to be dynamic in nature with its applications in various fields like image recognition, video analysis and recently successful applications in speech recognition and signal processing, the spectrograms which were made after the feature extraction of each audio file is fed to the 6-layer convolutional neural network.

CNNs use a filter also known as a kernel which is slid over the spectrogram image and the patterns are determined. It starts by learning features like vertical pixel lines, color layers, and starts picking up features like the shape of frequency-time curve. Such features provide powerful representation to help classify whether a person is depressed or not. Due to the highly detailed representation of speech in spectrograms, false noise signals can be inconveniently picked up by the network and can interfere with the learning process. This noise can be reduced by implementing different regularization parameters in the network, but since the training data is not abundant, it is quite difficult for the network to distinguish the true predictors of depression from the false signal.

5.2 Random Sampling

A major problem in learning shallow or deep depression models lies in uneven sample distribution. Many current benchmarks suffer from data imbalance between positive and negative samples or among different depression levels, which incurs a large bias in classification or regression. In the DAIC-WOZ database provided by AVEC 2016 for the DCC challenge, the number of non-depressed subjects is about four times bigger than that of depressed ones in both training and development parts. These samples if are embraced directly for learning, the result will have a strong bias to the non-depressed class, which will make the model untrustworthy.

Along with the problem of imbalance, the length of the sample can also create problems viz a longer signal of a particular person may tend the model to focus on particular characteristics that are only restricted to that person, this can make the situation worse, this can be solved by randomly sampling the training set.

To minimize the effect of the above discussed problems, each participant's segmented spectrogram was partitioned into 4-second slices after which participants were selected randomly from both classes (depressed and non-depressed) in equal numbers. Then, a fixed number of slices were sampled from each of the selected participants to ensure the CNN has an equal interview duration for each participant. The cropping also ensures that the CNN model gets the input of equal dimensions, which was not possible since the audio length was different for every patient. Thus, random sampling is an essential step to have a balanced input to CNN and also minimize the individual effect by random sampling.

5.3 Model Architecture

The Convolutional Neural Network (CNN) architecture consists of 6 layers in total with 2 convolutional layers, 2 max pooling layers alternately placed and 2 fully connected layers as shown in Fig. 5. Each input image to the CNN was 513x125 in dimension, representing 4 seconds of audio and frequencies ranging from 0 to 8kHz. Each input is normalized according to decibels relative to full scale (dBFS).

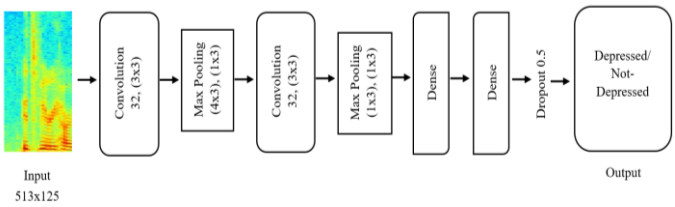


Fig-5: CNN Model Architecture

The input layer of CNN consisted of 32 filters and a kernel size of 3x3 followed by a ReLU activation function. After this a max pooling reduces the image map with a 4x3 filter and 1x3 stride. A second convolution with the same feature size and kernel size is employed which is then followed by a similar max-pooling layer to the previous one but the filter is reduced to 1x3. After these two dense layers are used which in turn is followed by a dropout layer of 0.5. Last but not the least, a softmax function is applied which returns the probability that the image is of depressed class or not. An Adadelta optimizer is used, which dynamically adapts the learning rate based on the gradient with a batch size of 32.

6. RESULTS

Depression changes the way one thinks, percepts things in a negative way but it's treatable with counseling and therapy. The major issue is regarding its detection, which can be solved using machine learning in an efficient manner. In this paper, we have presented a deep learning model CNN to recognize audio features and put forward a depression identification method. The F1-score of the model was observed to be 67.18% using Adadelta optimizer and Categorical Cross Entropy cost function. Table 1 below shows the report for performance characteristics of the model for DAIC-WOZ dataset.

Table-1: Performance Characteristics of Model

Sr. No	Model	Recall (%)	Precision (%)	F1 Score (%)	Overall Accuracy (%)
1.	Model1	67.94	66.44	67.18	64.8

The spectral features of the audio are exploited for identifying the patterns that distinguish depressed and normal groups. Since the dataset suffers from class imbalance problems, the experiments were done on the extended dataset which was generated by slicing the audio files into equal chunks with a duration of 4 seconds. The network is trained with the spectrogram images of both depressed and non-depressed classes respectively.

The last layer of the CNN, i.e., softmax layer is used for the classification purpose. The number of epochs is 7. An overall accuracy of 64.8% is achieved. Tests are conducted by capturing new audio files and analyzing the class labels predicted by the model. The proposed model has been successful in classifying between depressed and non-depressed samples using audio features.

7. CONCLUSION AND FUTURE WORK

We studied various stages involved in the classification of depression using audio signal: Signal Pre-processing which involves pre-processing of the audio data obtained from the database. There is a notable difference between the audio of a normal person and a depressed person in terms of amplitude, length and frequency of random pauses, in-frequent and sudden change of tone, etc. as they tend to be more sensitive and emotional. In order to preserve these precious features, training using spectrogram was decided.

In this paper, we proposed a Deep Neural Network model for Depression classification using the patient's speech. These deep learning models find common patterns in depressed patients and thereby assist in classifying depression in new patients. The trained CNN model has an accuracy of about 0.64. In addition, random sampling was used to balance the two classes, eliminating any bias in the model. The model was trained on a DAIC-WOZ dataset containing a PHQ8 score of patients along with their audio file.

Some further work is required to be done to optimize the model to achieve a greater percentage of efficiency. The dataset also needs to be expanded with a lot of samples of individuals, it also needs to be perfectly balanced as well as extra features like BDI reports could be added, and needs to be tried out on complex CNN architecture. This paper only demonstrates the use of audio characteristics for depression classification. A future implementation can be done using any one or multimodal approach using EEG (electroencephalogram), electrical signals emitted by the brain, which are far more accurate, video analysis of the interviewer while answering the questions, textual analysis of the spoken words etc. to achieve high accuracy and to generalize the model.

ACKNOWLEDGEMENT

These proposed work under the guidance of Prof. Amruta Sankhe Ma'am and supported by Atharva College of Engineering. We would like to thank USC Institute of Creative Technologies for providing the DAIC-WOZ dataset for this paper.

REFERENCES

[1] World Health Organization. Depression. (2020) [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression/

[2] A. Vázquez-Romero and A. Gallardo-Antolín, "Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks," *Entropy*, vol. 22, no. 6, p. 688, Jun. 2020.

[3] S. Krishna and Anju J, "Speech Based Depression Detection using Convolution Neural Networks," *Regular*, vol. 9, no. 9, pp. 405–408, 2020.

[4] K. Chlasta, K. Wołk and I. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks", *Procedia Computer Science*, vol. 164, pp. 618-628, 2019.

[5] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, Oct. 2013.

[6] S. Dham, A. Sharma, and A. Dhall, "Depression scale recognition from audio, visual and text analysis," *arXiv preprint,* arXiv:1709.05865, 2017.

[7] J. F. Cohn *et al.*, "Detecting depression from facial actions and vocal prosody," *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, Amsterdam, Netherlands, 2009, pp. 1-7, doi: 10.1109/ACII.2009.5349358.

[8] L. Chao, J. Tao, M. Yang and Y. Li, "Multi task sequence learning for depression scale prediction from video," *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, Xi'an, China, 2015, pp. 526-531, doi: 10.1109/ACII.2015.7344620.

[9] A. Jan, H. Meng, Y. F. B. A. Gaus and F. Zhang, "Artificial Intelligent System for Automatic Depression Level Analysis Through Visual and Vocal Expressions," in IEEE Transactions on Cognitive and Developmental Systems, vol. 10, no. 3, pp. 668-680, Sept. 2018, doi: 10.1109/TCDS.2017.2721552.

[10] J. Gratch et al., "The distress analysis interview corpus of human and computer interviews." in *LREC*. Citeseer, 2014, pp. 3123–3128.

[11] K. Kroenke, T. Strine, R. Spitzer, J. Williams, J. Berry and A. Mokdad, "The PHQ-8 as a measure of current depression in the general population", *Journal of Affective Disorders*, vol. 114, no. 1-3, pp. 163-173, 2009.

[12] Neves, Deângela Caroline Gomes, "Using Artificial Intelligence to Aid Depression Detection", 2019, http://monografias.ufrn.br/handle/123456789/9270.