Speech Signal Analysis to Predict Depression

Mogeeb A. Saeed
Ass. Professor Computer network and cybersecurity department, Faculty of Engineering & IT, Taiz University, Taiz, Yemen
Email: mogeeb1982@gmail.com
Vladimir Komashinsky
Ass. Professor Solomenko Institute of Transport Problems of the Russian Academy of Sciences, Russia
Email: kama54@rambler.ru
Saleem A. Mohammed
Department of Information Technology, Faculty of Engineering & IT, Taiz University, Taiz, Yemen
Email: saleemaliquid@gmail.com
Noha N. Abdulqader
Department of Information Technology, Faculty of Engineering & IT, Taiz University, Taiz, Yemen
Email: nohan.shamiri@gmail.com
Laila Q. Saif
Department of Information Technology, Faculty of Engineering & IT, Taiz University, Taiz, Yemen
Email: lailaalsofy2244@gmail.com

---ABSTRACT---

Depression is a widespread global disease of increasing global concern. Early recognition of signs of depression is crucial to evaluating and treating or preventing mental illness. With advances in machine learning, it has become possible to develop intelligent systems capable of recognizing depression and its signs in speech by analyzing speech and processing audio signals. This study presents an AI model for detecting and predicting mental illnesses through speech analysis of medical datasets related to depression. We used Distress Analysis Interview Corpus-Wizard of Oz (DAIC-WOZ) database where 60% of the data was reserved for training, while 20% was allocated for testing the model and another 20% for model validation. The model includes a convolutional neural network (CNN) to detect and predict mental illnesses. The proposed CNN model achieved an accuracy of 82% in the training and testing phases. Ultimately, the results are useful for classifying depression during English speaking and will be useful to psychiatrists and psychologists in contributing to early detection of depression at an affordable cost.

Keywords **– AI , CNN , Deep learning , Depression Detection , Speech Analysis , Speech Features Extraction.**

Date of Submission: November 08, 2024 Date of Acceptance: December 19,2024

I. INTRODUCTION

The advancement of the global health agenda is significantly impeded by mental health disorders (World Health Organization, 2020). WHO data shows that 23 million children and approximately 280 million adults experienced depression in 2019 [1], The COVID-19 pandemic has escalated into a major international health crisis. Numerous detrimental consequences of the epidemic have been reported, such as a major decline in income, a stop to children's education, a major loss of employment, and a barrier to economic advancement. During the COVID-19 pandemic, several lockdowns were imposed forcing people to spend a lot of time indoors. Orders to stay at home, isolation, and quarantine were among the policies that increased the risk of mental health conditions, the most common of which was depression, as well as emotional and financial distress.

Depression is among the most widely acknowledged severe health conditions worldwide. Depression is the primary factor that can lead to other mental health conditions, if not appropriately attended. It even results in psychosomatic ailments, the diagnostic of which would reveal real physiological symptoms. Therefore, early identification and treatment of depression are essential to preserving health. The goal of the study is to detect depression more accurately utilizing audio signals and subject voice elicitation.

There are typically two ways to identify depression. They are based on two things: conventional surveys and interactions with medical specialists. These approaches each have drawbacks of their own. When giving the conventional questionnaire approach, careful consideration must be given to it. Health care providers such as psychologists, counselors, and psychiatrists are involved based on their availability and level of skill in their respective fields. Researchers are becoming more interested in creating Machine Learning (ML) models for illness detection in the healthcare industry due to the rapidly expanding and improving Information Technology sector. Individual biomarkers provide the relevant patterns for ML algorithms to learn [2], [3].

The "Gold standard" in machine learning research is the automatic audio-based depression detection method. This kind of approach may be used as a pre-assessment exam conducted at home on a client who may have depression. Users can gain from this by saving time, reducing expenses, and avoiding travel costs, all of which contribute to an accurate and timely diagnosis. It therefore leads to prompt medical attention that may be provided to identify clients, who are now referred to as patients [4].

The main contributions of our work are as follows:

1. A CNN model was built that is capable of predicting depression with an 82% accuracy by analyzing English speech.
2. The DAIC-WOZ dataset, prepared for studying mental illnesses such as anxiety and depression, was utilized.
3. Voice features such as MFCCs and mel spectrogram were used due to their correlation and ability to capture subtle speech differences between depressed and non-depressed individuals.

This study is divided into the following sections: Section II lists relevant literature, and Section III discusses the suggested technique. Section IV concentrated on the discussion and simulation findings, while Section V provides the paper's conclusion.

I. RELATED WORKS

Chlasta et al [5] the use of Recurrent Neural Network for depression detection was proposed and the DAIC-WOZ dataset was employed. The results indicate that acoustic spectral features are promising for detecting individuals with depression. The study also confirmed that using short audio samples reduces the impact of noise.

Suresh Mamidisetti et al [6] this study focused on automatic depression detection using audio signals through a stacking-based ensemble framework. It aimed to improve the accuracy of depression detection by combining the predictive capabilities of diverse classifiers using a stacking method. The study utilized an Arabic dataset for vocal emotions, BAVED. Additionally, the publicly available dataset, DAIC-WOZ, was used for benchmarking. Features were extracted using the openSMILE toolkit, specifically the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). Various classifiers, including K-Nearest Neighbours (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Trees (DT), were utilized for intermediate predictions, with Logistic Regression (LR) used as a meta-classifier to predict the final outcome.

Tanzila Saba et al [7] in this study, a combined CNN and RNN model was used to classify depression using Arabic speech. The data set used in this study is BAVED and contains seven Arabic sentences that are classified according to their emotional intensity. This combined model achieved better accuracy than the individual models. 70% of the data was used for training and 30% for testing. The proposed model can be used as a diagnostic tool to detect depression during Arabic speech.

Takaya Taguchi et al [8] this study aimed to differentiate between the diagnosis of depression and bipolar disorder using MFCCs acoustic features for comparative analysis. A dataset was obtained by recording 38 patients who met the criteria for major depressive disorder from the psychiatric department of Tsukuba University and the Medical Center for Psychiatry in Ibaraki Prefecture, and the Kourita Hospital. The study revealed that the second dimension of MFCC significantly differed between the groups and enabled the distinction between patients. The study showed that MFCC2 was relatively higher in cases of depression, however, there was no correlation between the severity of depression and MFCC, and it was not affected by gender. The study was very limited and had very small sample size, and psychological factors had a confusing impact on the voice.

Hongbo Wang et al [9] this work uses a bidirectional Gated Recurrent Unit (GRU) with an attention mechanism in conjunction with highway networks and a 3D filter bank (3D-CBHGA). Enhancing speech-based depression detection accuracy is the goal of this project. The dataset utilized in the study was the DAIC-WOZ ENGLISH DATASET, which is employed in investigations pertaining to mental and psychiatric problems. The experiments were conducted using four algorithms: ID-CBBG, 3D-CBHGA, SVM, and RF with accuracies of 71.43%, 74.29%, 62.68%, and 68.57% respectively. The results demonstrated that the model used has the ability to detect depression from speech signals. Furthermore, the results also indicated that the algorithms performed better on the data with removed silent parts compared to the data containing silent parts.

II. METHODOLOGY OF STUDY

This research aims to predict depression by analyzing English speech. In this section, we will explain the step-by-step methodology used in our research, starting from selecting data , processing and handling, extracting voice features, preparing the model, training the model, testing the model, and measuring its performance.

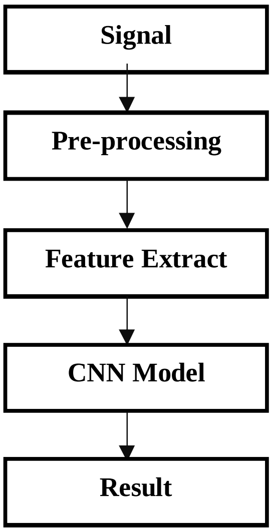

Fig.1. Flow Diagram of the Methodology

Selecting data

In this study, we used DAIC-WOZ dataset, which provided by University of Southern California and contains 189 audio recording samples from 154 individual subjects, four sessions were excluded due to artistic interventions during recording. The session duration ranges from 7-33 minutes, with an average of 16 minutes. These interviews were conducted by an AI interviewer named Eli controlled by a human interviewer in another room. The dataset was created as part of a broader project aiming to develop automatic diagnostic systems for mental illnesses. The dataset consists of audio recordings, video recordings, and PHQ scores. The primary objective is to indicate whether the participant is suffering from depression.

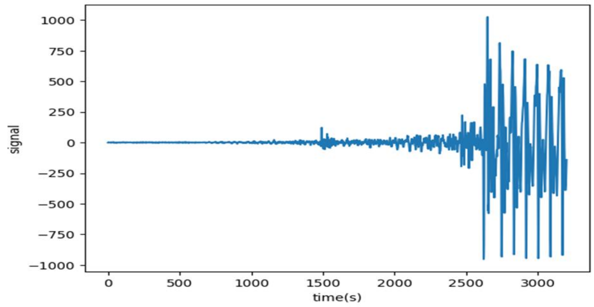

Fig.2. The first 0.2 seconds of the non-depressed audio signal

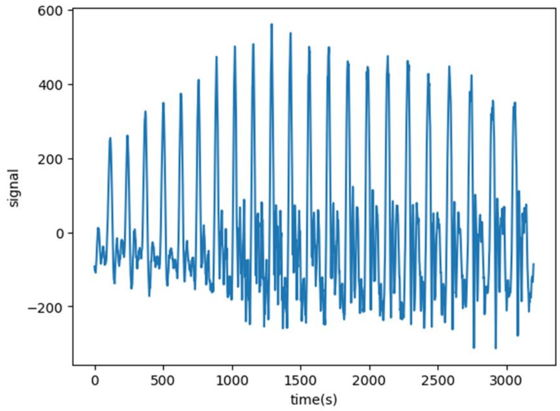

Fig.3. The first 0.2 seconds of the depressed audio Signal

Pre-processing data

The audio contains many periods of silence, as well as the voice of the virtual host, which is considered noise for the deep learning model because the voice of the host Eli is not related to depression, there are also some environmental noises like background noise that need to be removed. Fortunately, the database provides texts for all the audio recordings within the dataset, which have greatly facilitated data processing operations.

Audio segmentation

Sound segmentation is a fundamental task in audio signal processing. It involves dividing a continuous audio stream into smaller and more manageable units, often called frames or slices. In this study, transcript-based segmentation technique was used, where the audio signal was divided into a series of meaningful and purposeful small segments or frames. This segmentation was done based on specific criteria such as pauses, stops, speaker turn-taking, relying on textual files that define the start and end of the recording period of the audio signal for the mental patient. Segmentation played a crucial role in the research by providing a large amount of high-quality data, helping in noise detection and handling, as well as in identifying the start and end of the recording period for virtual axes and eliminating them. It also aided in organizing and facilitating the analysis of the audio signal and achieving precise data access more effectively.

Noise removal

The goal of this process is to reduce the level of noise without affecting the quality of the speech signal. In this study, we utilized the Spectral Subtraction technique, a simple and effective method for noise reduction. This technique involves estimating an average signal spectrum and an average noise spectrum in segments of the recording and subtracting the two spectra from each other [10]. The average noise spectrum is estimated in segments where the speech signal is absent for at least one second, ensuring that the average signal-to-noise ratio is improved. As a result, we significantly reduced the noise using this technique, ensuring that the enhanced signal contains minimal speech distortion.

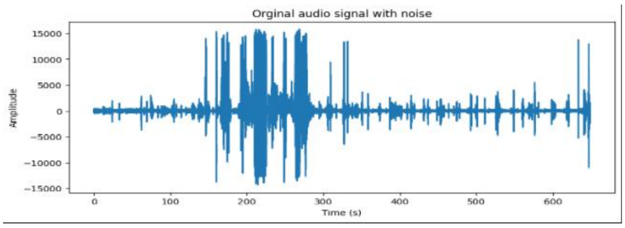

Fig.4. Original audio signal with noise

Spectral subtraction is based on the assumption that the background noise is uncorrelated with the speech signal, leading to the following relationship between the power spectra of the noisy signal $|X(w)|^2$, clean signal $|Y(w)|^2$,and noise $|B(w)|^2$:

$$|X(w)|^2 = |Y(w)|^2 + |B(w)|^2 \qquad (1)$$

The noise-reduced signal spectrum is obtained by subtracting the noise spectrum from the noisy signal spectrum:

$$|Y(w)|^2 = |X(w)|^2 - |B(w)|^2 \qquad (2).$$

The noise spectrum is averaged over multiple frames where only noise is present:

$$|B(w)|^2 = 1/M \sum_{j=0}^{M-1} |YsPj(w)|^2 \qquad (3).$$

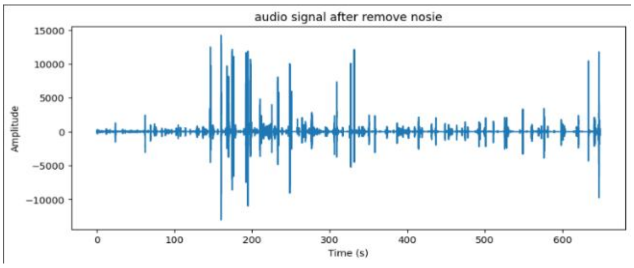

Fig.5. Audio signal after remove noise

Features extraction

In this study, two important features of the famous audio features MFCCs and mel spectrogram were used [11]. These are powerful features capable of capturing the unique characteristics of the audio signal and providing a strong representation of the audio data [12].

Mel-frequency Cepstral Coefficients (MFCCs): MFCCs are a powerful tool for capturing unique characteristics of audio signals [13]. Their effectiveness in audio classification stems from their ability to provide a compact yet robust representation of audio data, emphasizing frequency components most relevant to human auditory perception.

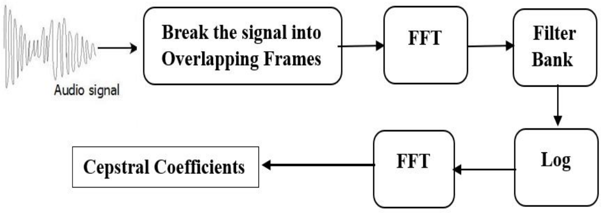

Fig.6. MFCC extraction diagram

Mel Spectrogram: A mel spectrogram is a spectrogram where the frequency axis has been warped to the mel scale, approximating the human ear's perception of sound. This transformation emphasizes the representation of lower frequencies while reducing the contrast in higher frequencies.

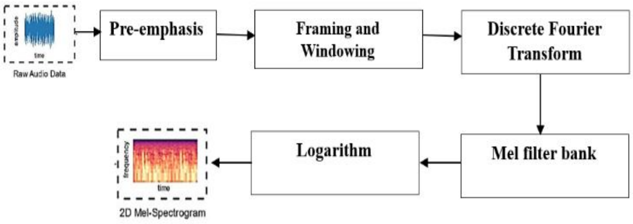

Fig.7. Mel spectrogram illustration diagram

We extracted the acoustic features from depressed and non-depressed individuals. 13 coefficients were extracted from the MFCCs and 128 coefficients from the mel spectrogram. It has been shown in this study that using more coefficients leads to an improvement in the accuracy of representing the audio signal. As a result, it enables learning models to identify more complex patterns in the audio data and to have a better understanding of the audio signal.

Proposed model

In this study, we used the CNN model due to its ability to learn complex patterns and deal with audio data. This network works on self-improvement through learning as in the traditional artificial neural network. The model we proposed deals with the features of MFCCs and mel spectrogram extracted from audio signals after converting them to two-dimensional matrices of depth and then entering them into the model and then classifying them into depressed or non-depressed. The figure.8 shows the structure of the used model.

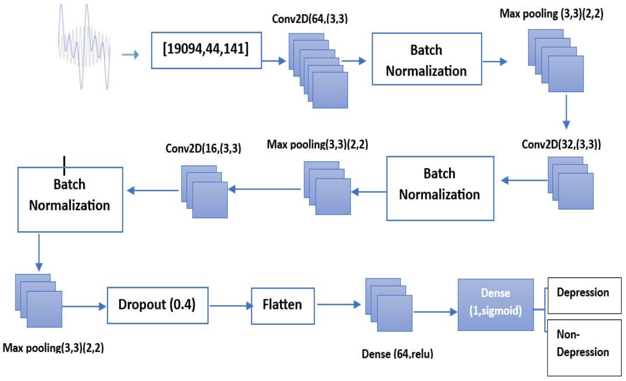

Fig.8. CNN Model Architecture

The model processes audio data represented as a 2D feature map, derived from Mel spectrograms and Mel-frequency cepstral coefficients (MFCCs), with dimensions [19094, 44, 141]. The architecture consists of several convolutional blocks, each comprising convolutional layers for feature extraction, followed by batch normalization and max pooling layers for dimensionality reduction. Specifically, the model utilizes multiple Conv2D layers with varying filter sizes to capture both local and complex features of the audio signals. Batch normalization is applied to enhance training stability, while max pooling layers effectively down sample the feature maps, preserving essential information.

To mitigate overfitting, a dropout layer is included, randomly deactivating a portion of neurons during training. The output from the convolutional layers is flattened into a one-dimensional vector, which is then fed into fully connected dense layers. The final layer employs a sigmoid activation function to produce a binary classification output, indicating the likelihood of depression.

Training and testing

In this study, the data was divided into 60% for training, 20% for testing, and 20% for model validation. Due to our use of the deep learning approach, we divided the data into three sections for training, testing, and validation, while if we used the machine learning approach, the data is only divided into two sections, training and testing only. This was the reason for dividing our data in this way. We monitored the model's performance during training and implemented early stopping functions based on the model's validation accuracy. Early stopping was used if there was no improvement in the performance metric.

III. RESULTS

The performance of proposed model, especially in terms of specificity, appears to be superior, suggesting that proposed approach may provide improved diagnostic accuracy.

This study used the accuracy measure as the primary measure to evaluate the effectiveness of the depression detection model. The accuracy of this study model is 82%, and the high performance of this study can be attributed to the powerful feature extraction techniques used, namely vertical frequency slope coefficients (MFCCs) and spectrograms. These features effectively capture the nuances of speech that indicate depression, facilitating accurate classification.

Binary classification

In this study, we used two training models, SVC and CNN, with different audio features extracted from the data. Such as MFCC, Chroma Features, ZCR, Energy and Mel Spectrogram.

Table 1 Results of SVC model with different features use Binary classification

Feature	Accuracy	Precision	Recall	F1 Score
MFCC	44%	25%	39%	30%
Chroma feature	42%	18%	50%	27%
ZCR	46%	32%	67%	44%
MFCC + ZCR	56%	35%	44%	39%

Table 2 Results of CNN model with different features use Binary classification

Feature	Accuracy
Energy	67%
Energy “with data resampling”	56%
Energy” with class weighting“	73%
MFCC	74%
Mel spectrum	79%

Table 3 Result of CNN proposed model with MFCC + Mel spectrum features use Binary classification

Feature	Accuracy	Precision	Recall	F1 Score
MFCC + Mel spectrum	82%	82.90%	98.97 %	91.43 %

Multi classification

This study tried predict depression level Since the degree of depression is consistent across many depression evaluations. Stated differently, there are five categories of depression: mild (1), moderate (2), moderately severe (3), severe (4), and minimal (0). The model may now accept audio data from additional depression corpuses built with various assessment techniques thanks to this methodology.

Table 4 Representation of the PHQ-8 score of multi classification

Depression Level	Total Severity Score (PHQ-8)
Minimal Depression	0 ~ 4
Mild Depression	5 ~ 9
Mild Depression	10 ~ 14
Moderately Severe Depression	15 ~ 19
Severe Depression	20 ~ 24

That can help to accurate diagnosis determine the extent and severity of depression, enabling doctors and therapists to design an appropriate and effective treatment plan.

Table 5 Result of CNN with MFCCs and Mel spectrum features use Multi classification

Model	Feature	Accuracy
CNN	Mel spectrum	48%
	MFCCs	50%
	Mel spectrum and MFCCs	53%

When we compared the performance of our model against several baseline models commonly employed in depression detection research. Specifically, the model exhibited higher accuracy compared to the baselines. This superiority underscores the effectiveness of our approach in accurately identifying depression from audio recordings.

IV. CONCLUSION

Early detection of depression is crucial to maintaining individuals' quality of life and maintaining overall health. In this paper, we focused primarily on speech and its relationship to depression as a basic means of identifying depression, as our model in this paper achieved an accuracy of up to 82% using the CNN Model to classify depressed individuals from healthy individuals.

One of the most important advantages of this approach is the ease of access to speech, which is a source of information that helps in detecting depression.

Despite the strengths of the DAIC dataset, it is not without limitations. The dataset primarily includes recordings from clinical interviews conducted by Ellie, an AI-based virtual interviewer. As such, the generalizability of findings to real-world scenarios involving diverse interview settings and interview techniques may be limited.

While this study achieved promising results, there are inherent limitations that deserve consideration. For example, relying on audio features alone may not be enough. Adding facial expressions and gestures may contribute to detecting depression well.

REFERENCES

[1] Mental health atlas 2020. Geneva: World Health Organization; 2021.

[2] N. K. Iyortsuun, S. H Kim, M. Jhon, H. J. Yang & S. Pant, A review of machine learning and deep learning approaches on mental health diagnosis, Healthcare, 11(3), 2023, 285. 10.3390/healthcare11030285

[3] H. Penuganti & G. Satyanarayana, Design and development of a deep learning model for classification of Alzheimer's disease using magnetic resonance images. International Journal of Advanced Networking and Applications, 15(6), 2024, 6174–6181.

[4] V. Ravi, J. Wang, J. Flint & A. Alwan, A step towards preserving speakers’ identity while

detecting depression via speaker disentanglement, Interspeech, 2022, 3338-3342. https://doi.org/10.21437/Interspeech.2022-10798

[5] K. Chlasta, K. Wołk & I. Krejtz, Automated speech-based screening of depression using deep convolutional neural networks. Procedia Computer Science, 164(3,4), 2019, 618-628.

[6] S. Mamidisetti & A. M. Reddy, A stacking-based ensemble framework for automatic depression detection using audio signals, International Journal of Advanced Computer Science and Applications, 14(7), 603-612.

[7] T. Saba, A. R. Khan, I. Abunadi, S. A. Bahaj, H. Ali & M. Alruwaythi, Arabic Speech Analysis for Classification and Prediction of Mental Illness due to Depression Using Deep Learning, Computational Intelligence and Neuroscience, 1, 2022, 8622022.

[8] T. Taguchi, H. Tachikawa, K. Nemoto, M. Suzuki, T. Nagano, R. Tachibana & T. Arai, Major depressive disorder discrimination using vocal acoustic features, J Affect Disord. 2018 Jan 1;225:214-220. doi: 10.1016/j.jad.2017.08.038. Epub 2017 Aug 16. PMID: 28841483.

[9] H. Wang, Y. Liu, X. Zhen & X. Tu, Depression speech recognition with a three-dimensional convolutional network. Frontiers in human neuroscience, 15, 2021, 713823.

[10] N. Upadhyay & A. Karmakar, Speech enhancement using spectral subtraction-type algorithms: A comparison and simulation study, Procedia Computer Science, 54(2), 2015, 574-584.

[11] J. Gratch, R.Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo & L. P. Morency, The distress analysis interview corpus of human and computer interviews. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC' 14), May 2014, 3123- 3128.

[12] N. W. Hashim, M. Wilkes, R. Salomon, J. Meggs & D. J. France, Evaluation of voice acoustics as predictors of clinical depression scores. Journal of Voice, 31(2), 2017, 256.e1-256.e6.

[13] A. Thenmozhi &P. Kannan ,Performance Analysis of Audio and Video Synchronization using Spreaded Code Delay Measurement Technique, International Journal of Advanced Networking and Applications, 10(01), 2018, 3728-3734