International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 07 | July 2023 www.irjet.net p-ISSN: 2395-0072

Mental Illness Prediction based on Speech using Supervised Learning

Nandini B M[1], Ankita Vishwanatha Hegde[2], Bhumika Shetty[2], N Aditya Bhat [2], Rakesh R[2]

[1]Assistant Professor, Dept. of Information Science & Engineering, National Institute of Engineering, Mysuru, Karnataka, India
[2]Student, Dept. of Information Science & Engineering, National Institute of Engineering, Mysuru, Karnataka, India

---

Abstract - *This research presents a novel approach to accurately detect depression using voice-based analysis and machine learning algorithms. The main objective is to create a model that predicts whether an individual is experiencing depression using their audio inputs. The platform allows users to provide voice recordings, which are analysed using various acoustic features. By employing a supervised learning algorithm, the model effectively classifies the recordings as indicative of mental illness or not, thereby achieving high accuracy in predicting depression from voice recordings. The findings of this research demonstrate the potential of voice-based analysis and machine learning algorithms in early diagnosis and treatment of depression. By accurately identifying individuals who may be experiencing depression, the model enables the recommendation of necessary support and intervention. The study offers valuable perspectives and a practical tool for clinicians and researchers in the domain of mental health, offering a new avenue for improved mental health assessment and intervention. Harnessing the strength of voice analysis, this approach contributes to enhancing the well-being of individuals affected by depression.*

Key Words: **Machine Learning; Speech Based Detection; Depression; Convolutional neural network; Mental Illness Prediction**

1. INTRODUCTION

Mental health conditions affect a substantial portion of the global population, with approximately 450 million individuals experiencing such challenges worldwide. Among these conditions, depressive disorders being a major contributor to the worldwide burden of diseases and are projected to become the second leading cause. Depression, a prevalent and severe mental illness, manifests as persistent feelings of sadness and a lack of enthusiasm in daily activities. Detecting and predicting the presence of mental illness, particularly depression, poses a formidable challenge.

Untreated depression can have profound consequences, including diminished motivation, persistent sadness, and low self-esteem. Physical health issues such as weight fluctuations, fatigue, and bodily pains can also arise. Additionally, depression can impair an individual's ability to function effectively, impacting work, education, and the ability to derive pleasure from previously enjoyed activities. Furthermore, individuals with depression may be at an increased risk of developing co-occurring mental health conditions like anxiety and substance abuse. In severe cases, depression can lead to suicidal ideation or attempts. Timely intervention and support are very important when dealing with depression. As the adage goes, "Prevention is better than cure," emphasizing the need and importance of early detection can have far-reaching implications, including escalation of the disorder.

The complexity of mental illness makes prediction and identification challenging. Over the past few years, there has been growing interest in utilizing specifically in the realm of supervised machine learning techniques, to discern and forecast mental health conditions based on speech patterns. Supervised learning entails training a model using labelled data to learn parameters that accurately classify audio samples as indicative of a mental illness or not. Subsequently, the model can be deployed to effectively predict mental illness in new audio samples. Such an approach holds promise for delivering more accurate and timely diagnoses, enabling improved treatment and ultimately enhancing the quality of life for individuals grappling with mental health disorders.

Supervised learning models based on speech analysis offer an avenue to identify psychological conditions by scrutinizing speech patterns. Individuals with depression exhibit distinctive characteristics, including reduced speaking speed, diminished intonation, lower voice intensity, diminished variations in speech features, and increased pauses. Additionally, alterations in voice bandwidth, amplitude, energy, and other vocal attributes can be observed. Leveraging these features, the model is trained and evaluated, yielding precise and reliable outcomes.

In this paper, we delve into the realm of mental illness prediction using supervised learning algorithms and voice analysis. We explore the potential of speech-based models to discern and predict depression, aiming to influence the

advancement of early detection and intervention strategies. This voice-based analysis will help in improving mental health assessment thereby enhancing overall well-being of the individuals affected by depression.

The following section of this paper will explore the technical elements of this model such as data preparation, Feature Extraction, model architecture used to predict depression from the speech. It also discusses about the suggestions provided by the platform for nearby hospitals based on the prediction. By examining the intricate relationship between speech patterns and mental health, we aim to illuminate the advancements achieved in the development of practical tools that have the potential to assist clinicians, researchers, and individuals in their pursuit of enhanced mental well-being.

2. BACKGROUND STUDY

The current landscape of mental illness prediction revolves around the development of questionnaire-based screening systems. These systems aim to incorporate a range of risk factors related to mental health, encompassing demographic information, psychosocial stressors, and clinical history, to develop all-encompassing questionnaires that can accurately predict the existence of mental disorders. To enhance the accuracy of these questionnaires, researchers have explored the integration of machine learning techniques.

The process of mental illness prediction through questionnaires involves the careful design of an appropriate questionnaire that covers essential aspects of the person's mental health, including emotional well-being, coping strategies, and lifestyle factors. Once the questionnaire is completed by the user, the responses are subjected to evaluation and analysis. This assessment phase aims to decipher whether the individual exhibits signs of mental illness or distress based on their provided answers. Subsequently, the findings are analyzed, considering the responses to determine the likelihood of the person having a mental illness. In the depression detection model [2] built by Mariyam Begom, Anamika Ahmed, Md. Muzahidul Islam Rahi, Raihan Sultana, Md Tahmidur Rahman Ullas, Md. Ashraful Alam, PhD, they have put forward a model that incorporates a conventional psychological assessment alongside the utilization of machine learning techniques to diagnose and assess various levels of mental disorders using questionaries. This accomplishment was attained through the utilization of algorithms such as Linear discriminant analysis, Convolutional Neural Network [5], K Nearest Neighbor Classifier [9], Support vector machine [6] and Linear Regression on the two datasets of anxiety and depression. Their model has achieved the highest accuracy of 96.8% for depression and 96% for anxiety using the CNN [5] algorithm.

However, while questionnaire-based systems may appear to offer results that show great potential, they are not without their limitations. Traditional approaches, such as relying solely on questionnaires, can be time-consuming and resource intensive. The subjective interpretation of responses introduces a degree of uncertainty, potentially leading to inaccuracies in predicting mental illness. Furthermore, questionnaires may have inherent limitations, encompassing a predetermined set of answers that may not capture the complexity of an individual's mental state. Biases can also arise if the questions are poorly formulated or if respondents provide incomplete or inaccurate information. Moreover, questionnaires may not be easily accessible to individuals with certain disabilities or specific mental health conditions. Additionally, language barriers can hinder accurate understanding and response provision, potentially undermining the reliability of the predictions. Considering these challenges, there is a pressing need to explore innovative methodologies that can overcome the limitations of questionnaire-based systems.

This paper aims to tackle these issues by proposing an alternative approach that utilizes acoustic features in speech analysis to enhance the efficiency and accuracy of mental illness prediction. By leveraging state-of-the-art machine learning algorithms, we strive to develop a more reliable and accessible system that can help in early detection and intervention, ultimately leading to improved psychological well-being results for individuals. The model proposed employs machine learning algorithms and offers a user-friendly platform that allows audio input from individuals seeking assessment. Through the examination of these inputs, the algorithms ascertain whether the person is experiencing depression or not. In instances where depression is detected, appropriate recommendations are made, including referrals to nearby psychologists or self-help groups.

In the depression detection model [1] built by B. Yalamanchili, N. S. Kota, M. S. Abbaraju, V. S. S. Nadella and S. V. Alluri, uses potential of acoustic cues, such as speech patterns, prosody, and voice quality to indicate depression. The methodology section describes the data collection process, preprocessing steps performed on the audio data to prepare it for extraction of the features. The authors discuss the performance metrics, including specificity, sensitivity, and overall accuracy, to test the models' effectiveness. The authors of this paper have used SMOTE [7] analysis to remove class imbalance and successfully achieved 93% accuracy using SVM [6] algorithm. The extraction acoustic features from the audio is done using COVAREP toolbox and then the features are fused. Thus, this study has helped to obtain Depression Classification Model (DCM).

3. METHODOLOGY

This paper proposes a "Mental Illness Prediction Model" using a supervised learning methodology. Fig. 1 represents the design flow of the proposed model. Large amount of data collected from depressed and non-depressed people is segmented to remove unwanted disturbances and noises present. The segmented data will then be given to an algorithm to extract the features. After extracting the feature, the data should be classified to get accurate results. For the classification, convolutional neural networks (CNNs [5]) are used. Therefore, the data is visually represented as spectrograms. The classification process involves class balancing, modelling the architecture, and training the model.

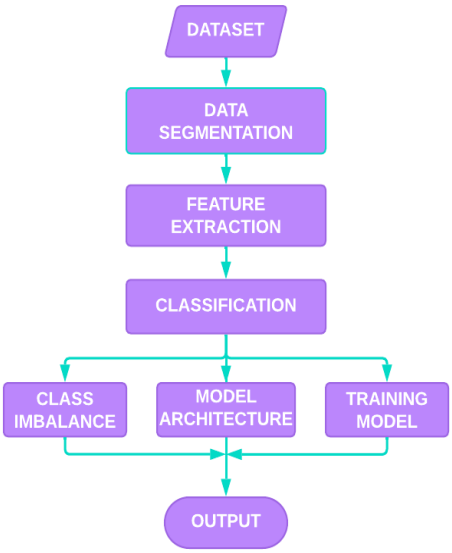

Fig. 1 Design Flow

3.1 Dataset Collection

The Man-machine conversation dataset serves as an important resource for studying and analyzing interactions between humans and various machine interfaces, such as chatbots and virtual assistants. This extensive dataset encompasses a broad spectrum of conversational data, incorporating both natural language dialogues and structured information, such as user intent labels. One prominent dataset within this collection is the DAIC-WOZ [10], which is specifically designed for training models in the domain of man-machine conversations. It comprises recordings of audio conversations between individuals and a virtual interviewer named Ellie, with a total of 189 sessions. Fig.2. shows the virtual interview with Ellie. These sessions have an average duration of approximately 16 minutes, offering substantial data for analysis and investigation.

Fig. 2. Virtual interview with Ellie

To ensure a thorough understanding of the dataset's content, the USC Institute of Creative Technologies played a crucial role in compiling the DAIC-WOZ dataset. The Institute released this dataset as part of the 2016 Audio/Visual Emotional Challenge and Workshop (AVEC 2016). The dataset is unique in that it includes individuals representing both depressed and non-depressed populations. Prior to every interview session, participants were required to complete a psychiatric questionnaire known as the PHQ-8. Based on the responses provided, a binary classification was derived to indicate whether

the participant was classified as depressed or not depressed. This binary "truth" classification adds significant value to the dataset, enabling researchers and professionals to explore and develop various depression prediction models using the collected audio recordings.

3.2 Data Preparation

For speech segmentation, the required acoustic features were extracted from the speech signals using pyAudioAnalysis. These features, such as energy, pitch, duration, and spectral content, facilitated the division of speech signals into distinct segments. The library's robust algorithms for pitch detection, energy detection, spectral analysis, and temporal analysis were employed for accurate segmentation of the speech data. Furthermore, speaker diarization was performed using the speaker identification capabilities offered by pyAudioAnalysis. This aided to differentiate participant speech input with that of the interviewer. The integration of pyAudioAnalysis into the project provided reliable speech segmentation and speaker diarization. The library's comprehensive set of features and algorithms empowered the extraction of meaningful speech segments and the accurate identification of distinct speakers within the recordings.

TABLE 1

AUDIO FILES BEFORE SEGMENTATION

Name Of Audio Files	Duration
301_AUDIO.wav	10:42 minutes
302_AUDIO.wav	13:42 minutes
303_AUDIO.wav	12:38 minutes

TABLE 2

AUDIO FILES AFTER SEGMENTATION

Name Of Audio Files	Duration
P301_no_silence.wav	07:50 minutes
P302_no_silence.wav	05:16 minutes
P303_no_silence.wav	10:03 minutes

The duration of a few audio files in the audio dataset before segmentation is shown in TABLE 1. The time duration post removal of silence and interviewer voice has subsequently reduced which is shown in the TABLE 2.

3.3 Feature Extraction

Speech feature extraction involves retrieving meaningful information from speech signals to classify speech content or understand spoken words. Acoustic feature extraction from speech includes extracting various features like pitch, energy, spectral features, formants, and prosodic features. Pitch represents the frequency of sound produced and helps identify the speaker's vocal range. Energy indicates the speaker's loudness. Spectral features describe energy distribution across the frequency spectrum, enabling voice quality identification. There are various methods for acoustic feature extraction. This research has employed Mel Frequency Cepstral Coefficient [8]s (MFCC) which transforms speech signals into coefficients representing short-term power spectrum, widely used in speech processing tasks. Further, convolutional neural networks (CNNs [5]) with spectrograms have been employed, which give a visual representation of frequency spectrum variations over time. Spectrograms, created using a spectrograph, are widely used in sound engineering, speech recognition, and linguistics. They help analyze signal frequencies, detect patterns and changes, and study language structures and musical compositions.

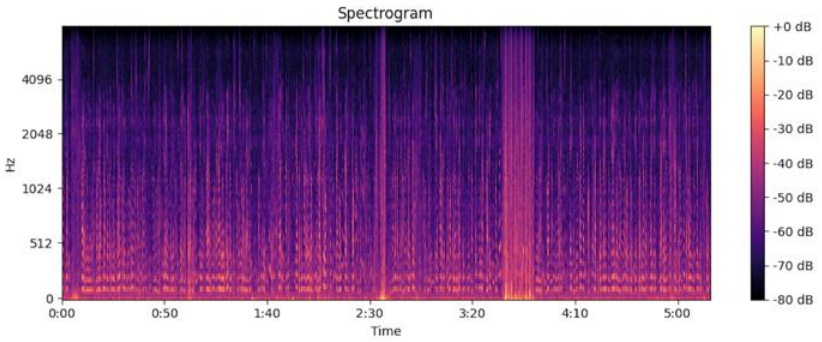

Fig. 2. Sample Spectrogram

This research utilizes the Short-time Fourier Transform (STFT) technique to compute the spectrogram. The STFT is a well-known signal processing technique that analyzes a signal's frequency content over short, overlapping time windows. The STFT uses appropriate parameters to compute the spectrogram matrix. This spectrogram matrix is then reshaped and flipped to correspond the desired orientation for visualization. The resulting spectrogram matrix is subsequently employed for additional processing purposes. Since the convolutional neural network (CNN [5]) models cannot process audio, the spectrogram matrices are used as input to the model. Thus, Spectrograms are specifically useful in scenarios where temporal and frequency information is crucial, as they provide a compact and visually intuitive representation of the audio data. The logarithmic scaling applied to the frequency axis improves the perceptual depiction of the spectrogram, aligning it with human auditory perception. The script's flexibility allows for experimentation with different window sizes, overlap factors, and scaling factors, enabling users to adapt the spectrogram generation process to their specific requirements.

3.4 Classification of Spectrograms

Classification of spectrograms focuses on building dictionaries for the depressed and non-depressed classes using the corresponding segmented matrix spectrogram representation. This classification uses PHQ-8 values to create depressed and non-depressed dictionaries. The PHQ-8 values are derived from the Truth Dataset provided by The University of California. This dataset encompasses the Patient Health Questionnaire-8 (PHQ-8) binary score values, coupled with predefined values indicating the depression status of individuals.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Participant_ID	PHQ8_Binary	PHQ8_Score	Gender	PHQ8_NoInterest	PHQ8_Depressed	PHQ8_Sleep	PHQ8_Tired	PHQ8_Appetite	PHQ8_Failure	PHQ8_Concentrating	PHQ8_Moving
2	302	0	4	1	1	1	0	1	0	1	0	0
3	307	0	4	0	0	1	0	1	0	2	0	0
4	331	0	8	1	1	1	1	1	1	1	1	1
5	335	1	12	0	1	1	3	2	3	1	1	0
6	346	1	23	0	2	3	3	3	3	3	3	3
7	367	1	19	1	3	3	2	2	2	3	3	1
8	377	1	16	0	2	2	1	2	3	3	2	1
9	381	1	16	1	2	3	3	3	1	3	0	1
10	382	0	0	1	0	0	0	0	0	0	0	0

Fig. 3. Truth Dataset

In the Fig.3, the PHQ-8 values are indicated in binary where 0 indicates non-depressed and 1 indicates depressed. Various other representation like PHQ-8 score is also provided. The PHQ-8 score less than 10 indicates non-depressed and greater than 10 indicates depressed. Based on this, the two distinct dictionaries are created in which one specifically customized for individuals who exhibited signs of depression and another for those who did not, are then employed for the purpose of classification tasks, such as training a Convolutional Neural Network (CNN [5]) to classify spectrograms as either depressed or non-depressed.

3.5 Training and Testing data

The pre-processing of our dataset is done by employing diverse procedures. We split our dataset into training and testing sets through the utilization of random sampling. This technique involves randomly dividing the dataset into

representative subsets. This ensures unbiased split of the dataset and evaluating the performance of our model. Thus, it helps in obtaining reliable results. The training and testing dataset is ensured to have balanced number of both depressed and non-depressed samples. The test sample size is set to 0.2, while train sample size is 0.8. The training sample is allocated with larger portion of the data to ensure the model learns the patterns, relationships, and features from the data input.

3.6 Model Architecture

This proposed system uses a deep learning neural network called convolutional neural network, or CNN [5]. The model uses 6-layer CNN as shown in Fig.4. The 6-layer convolutional neural network model consists of two convolutional layers, two max-pooling layers and two fully connected layers. The first layer in the architecture is a convolutional layer that is used to detect patterns in images and videos. The first layer is followed by a pooling layer where max pooling technique is used. It is used to reduce the size of the input while preserving the important features of the input. Convolutional layer and pooling layer are repeated once again as third and fourth layer. The subsequent pair of layers comprise fully connected layers, where each neuron is connected to every neuron in the preceding layer. In a fully connected layer, the output of a neuron is determined by the weighted sum of inputs received from the previous layer. These weights are adjusted during the training phase of the neural network to minimize the discrepancy between the predicted output and the actual output.

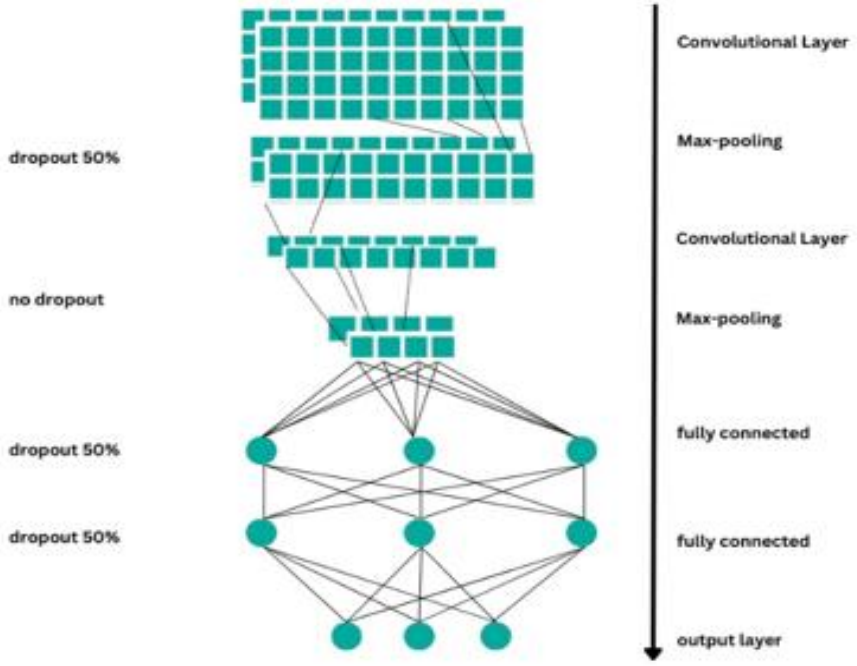

Fig. 4. Layers of architecture

During the pre-process stage, normalization is performed to prepare the spectrogram samples for input into the CNN [5] model. By normalizing the data, model can learn from consistent and standardized features, which enhances its ability to classify normal and depressed participants accurately. This normalized data is then reshaped as per expected input dimensions, the spectrogram samples are prepared in a format suitable for feeding into the CNN [5] model. This ensures compatibility between the model architecture and the data, enabling effective training and evaluation of the model.

3.7 Results and Accuracy

The effectiveness of the trained model has been evaluated using various assessment criteria like accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the model's performance in classifying normal and depressed participants. With the number of epochs 21 and batch size of 7, the trained model has achieved to provide accuracy of 67%. While with the number of epochs 36 and batch size of 6, the trained model has achieved to provide accuracy of 71%. Below figures Fig. 5.1, 5.2 and 5.3 show model accuracy, model loss and ROC curve of model trained with batch size of 6 and number of epochs 36 respectively.

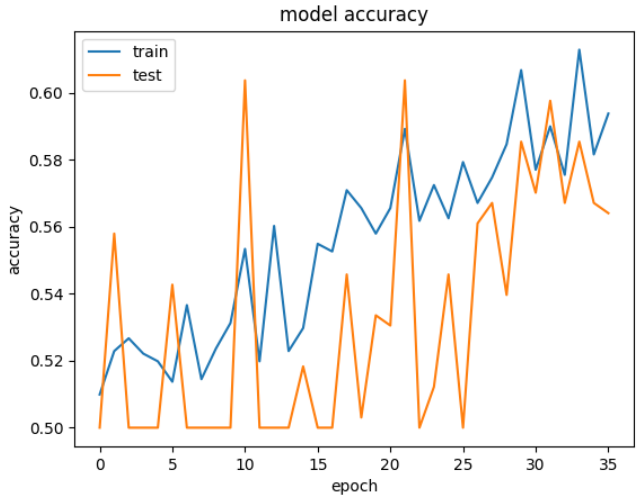

Fig 5.1 Model Accuracy of Trained Model with 71% accuracy

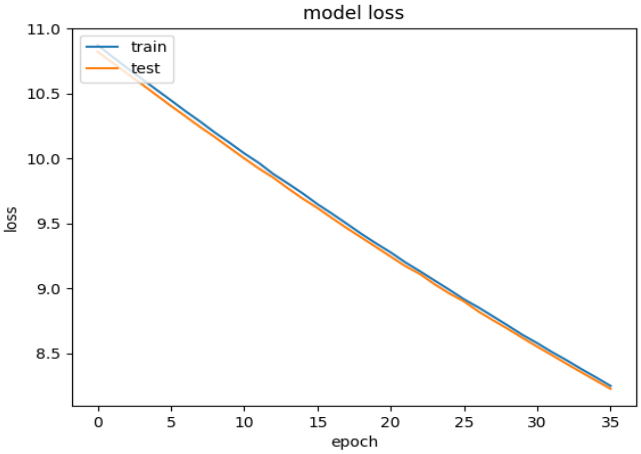

Fig 5.2 Model Loss of Trained Model with 71% accuracy

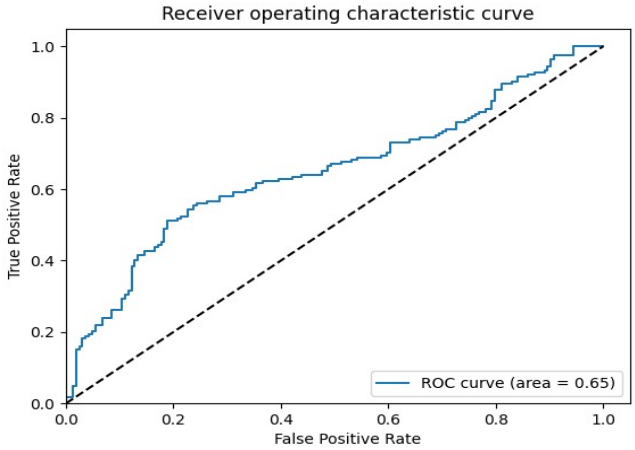

Fig 5.3 ROC curve of Trained Model with 71% accuracy

3.8 Graphical User Interface

The system is designed to assess the mental status of users who provide audio recordings in the .wav format files and their corresponding location information. When a user wants to check their mental status, they upload the audio file along with their entered location. The audio file undergoes pre-processing to ensure optimal input for the model. The pre-processed audio is then passed through a trained model to determine the presence of depression.

If the model detects signs of depression, the system proceeds to the next step. It utilizes an open-source and freely available API to retrieve nearby hospitals based on the latitude and longitude calculated from the user's entered location. This API helps identify suitable healthcare facilities in the user's vicinity that specialize in mental health treatment. By leveraging this API, the system guarantees that users are provided with relevant and accessible resources for their condition.

In the case where the user is deemed healthy based on the model's analysis, the system delivers a message of congratulations for their mental well-being. No further actions are taken about hospital recommendations since the user does not require immediate mental health care. This outcome acts as a positive reinforcement for the user, motivating them to continue prioritizing their mental health and well-being.

Overall, the system's design encompasses audio pre-processing, model analysis for depression detection, and integration with an open-source API to obtain nearby hospital recommendations. By combining these components, the system efficiently evaluates a user's mental status, offers appropriate support if depression is detected, and provides encouragement for those who are mentally healthy.

4. CONCLUSION

This study suggests that speech analysis holds significant promise as a tool for predicting mental illness. While the current level of precision in speech analysis may not be sufficient for making definitive diagnostic decisions, it can still provide valuable insights into an individual's mental state. By analyzing various aspects of speech, such as tone, pitch, and linguistic patterns, researchers and clinicians can identify potential indicators of mental health issues. This early detection can be crucial in allowing healthcare professionals to intervene sooner and provide more effective care. Therefore, further exploration of speech analysis in the context of mental illness prediction is a compelling avenue for research.

5. ACKNOWLEDGEMENT

We would like to use this opportunity to show our sincere gratitude to everyone who helped us, directly or indirectly, to finish this project. We appreciate The National Institute of Engineering, Mysuru, Karnataka, India, for giving us the chance and the tools we needed to do this research successfully. Additionally, it gives us great pleasure to thank Nandini B M, Assistant professor in the Department of Information Science & Engineering at NIE Mysuru who served as our project guide throughout. And we appreciate her consistent enthusiasm and support.

6. REFERENCES

[1] 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," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-6, doi: 10.1109/ic-ETITE47903.2020.394

[2] A. Ahmed, R. Sultana, M. T. R. Ullas, M. Begom, M. M. I. Rahi and M. A. Alam, "A Machine Learning Approach to detect Depression and Anxiety using Supervised Learning," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411642.

[3] D. Shi, X. Lu, Y. Liu, J. Yuan, T. Pan and Y. Li, "Research on Depression Recognition Using Machine Learning from Speech," 2021 International Conference on Asian Language Processing (IALP), 2021, pp. 52-56, doi: 10.1109/IALP54817.2021.9675271.

[4] R. Katarya and S. Maan, "Predicting Mental health disorders using Machine Learning for employees in technical and non-technical companies," 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE), 2020, pp. 1-5, doi: 10.1109/ICADEE51157.2020.9368923.

[5] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[6] S. Ghosh, A. Dasgupta and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," 2019 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2019, pp. 24-28, doi: 10.1109/ISS1.2019.8908018.

[7] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya and M. Ismail, "SMOTE [7] for Handling Imbalanced Data Problem: A Review," 2021 Sixth International Conference on Informatics and Computing (ICIC), Jakarta, Indonesia, 2021, pp. 1-8, doi: 10.1109/ICIC54025.2021.9632912.

[8] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," in IEEE Access, vol. 10, pp. 122136-122158, 2022, doi: 10.1109/ACCESS.2022.3223444.

[9] K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.

[10] Dataset Title: The Distress Analysis Interview Corpus (DAIC) WOZ Dataset Link: http://dcapswoz.ict.usc.edu/ Accessed: 18th November 2022.