Analysis of Speech Patterns in Children for Detecting Potential Depression.

Sanjana M Moodbagil
Information Science and Engg
BMS College of Engineering
Bangalore, India
msanjana9819@gmail.com

Rashmi V Bhat
Information Science and Engg
BMS College of Engineering
Bangalore, India
rashmivbhat1998@gmail.com

Supreeta Anand Byatnal
Information Science and Engg
BMS College of Engineering
Bangalore, India
supreeta.byatnal@gmail.com

Nalini M K, Asst Professor
Information Science and Engg
BMS College of Engineering
Bangalore, India
nalini.ise@bmsce.ac.in

***Abstract*—Anxiety and Depression in children can emerge at a very early stage and often go undiagnosed. Symptoms are frequently disregarded until the time children can communicate their discomfort easily. Whenever left untreated, it is regarded as internalizing disorders, which result in extended periods of adverse outcomes, including substance abuse and suicidal thoughts. These prolonged childhood depressions can quickly turn to teen depression. Hence, to avoid internalizing disorders, the detection of depression, and anxiety at an early stage is necessary. A design of dynamic relationship between characteristics of voice audio signals and diagnosis needs to be developed for identifying kids with internalizing illness. For this task, an approach which is driven by data, such as machine learning, is best performed by analyzing the participant 's reaction to a speech challenge, a behavioral activity designed to evoke anxiety. In a speech task, we use the recording of children's speech to analyze and extract features and ML to identify and diagnose clinically induced symptoms for anxiety and depression in children aged 3 to 7 years. In this paper, we have implemented Linear Regression, Random Forest, SVM-Linear, and SVM-Gaussian models with features extracted using MFCC and compared their accuracy. Random Forest gives the highest accuracy with 94%. We have also shown how transfer learning can be implemented by using WaveNet model with CNN model and its comparison to a CNN model implemented using MFCC features. The results point towards increased use of transfer learning with a much larger samples for very accurate prediction results.**

***Index Terms*—Machine Learning, Audio analysis, Transfer learning, Anxiety, Depression, Psychopathology, Voice Activity Detector (VAD), Mel-frequency Cepstral Coefficients(MFCCs)**

I. Introduction

Mental health is an essential aspect of the overall health of a child. It has a very tightly wound relationship with the physical health and ability to perform their best at school, work, and develop an overall healthy emotional outlook on life. It is easy for parents to identify the basic needs of childlike clothes that are warm when it is cold, food that is nutritious and good for their health, sleep to get energized. However, it may not be obvious to recognize the emotional needs of a child. Hence it becomes imperative to identify how a child is doing in terms of his/her mental health as good mental health is vital for the child to develop new skills and grow in the society. Depression is more than just typical blues or sadness that a child faces. It affects the ability to function normally in day to day life. Anxiety and Depression can be seen in a child of age as young as four years. Due to the inability to express their emotions clearly at this very young age, the symptoms often get overlooked. Studies show that about 2 percent of the children of preschool and school-age children are affected by depression. There are various types of symptoms that the child tends to exhibit, some of which include having an irritable mood, anger, loss of interest, physical body aches, and many more clinical signs. Just as in adults, various factors contribute to the risk of developing depression in children, some of which are as below.

Physical health – Physical health and mental health are directly related. Children who suffer from chronic diseases are most likely to be depressed. Surroundings – A stressful home life or a stressful peer group contribute to the risk of developing depression.

Family history – If a parent or any of the family members suffer from depression, it is most likely for the child to develop depression at a very young age.

Identifying that a child is developing the symptoms of depression is harder than it sounds. From the psychological viewpoint, it is said that one of the apparent signs of depression is the way a human expresses their emotions in their speech, assuming that depression often affects the acoustic qualities of the speech. We can perform a computer based, automated analysis of the speech to ease the process, help the clinicians to efficiently diagnose and hence treat the illness at a much earlier stage. Deep Learning methods are often seen to give fruitful results when it comes to the analysis of speech signals due to their unstructured nature. Therefore, in our proposed model, we present a machine learning tool that can objectively screen children for anxiety and depression, which is a valuable asset for early childhood intervention to prevent future problems of mental health issues.

A. Background Work

Children, like adults, experience depression and anxiety, which is known as internalization disorders. However, unlike adults, young children often have difficulty communicating their symptoms, making it challenging for health care professionals to diagnose and treat accurately. Medically or psychologically, there are no particular tests to show depression in children. But resources like interviews conducted by mental

health professionals and questionnaires for both parents and the child can help diagnose more accurately. The traditional diagnosis requires a meeting with a professional psychiatrist and their caregiver for 60-90 minutes. Studies have been going on by researches trying to use artificial intelligence and machine learning to make faster and reliable diagnosis. Earlier, they used mood induction technique called the Trier-Sociak Stress Task which triggers the child to feel anxious and stressed. As mentally ill children have trouble communicating, researchers have used speech patterns to diagnose conditions that are hard to spot and are often overlooked.

B. Motivation

Depression is a severe medical illness that affects the way we think, feel, and act. For this task, we analyze the performance of several statistical models, investigate how the low quality of data affects the performance of the model and hence compare the performances and use the best available model to test and diagnose depression in children . Parents are often unaware of symptoms of depression and anxiety at early stages. More than 20 percent of kids, however, tend to develop an internal disorder during childhood. This influences the growth and development of the child and predicts, if left unchecked, severe health problems later in life. Thus, there are social burdens which require the need for early detection of such disorders in the mental health of a child.

II. Related Work

Often, depression goes unnoticed in children. In our paper, we intend to diagnose depression in children by analyzing their speech patterns. In [1], France, D. J et al. has laid down that the acoustic properties of speech can be taken as possible cues of depression as there is enough evidence to establish the same. Thus, specific vocal parameters can be used to classify depressed speech.

In [2] Moore, E et al. explains how the vocal effect describes the mental and emotional state in which the speaker is and the overall condition. He enunciated the importance of including glottal features in voice analysis. The voice track produces a range of format frequencies (FMT) such as F0, F1, F2, and so on based on the speaker's enunciation of voice.For voice analysis, the spectral information is considered. The Voice activity detector (VAD) proves to be a great tool in detecting and extracting human voice signals from the input audio.

In [3], Cummins et al. talks about the Segment selection from the voice spectrum where the active audio regions that provide maximum discrimination between neutral and depressing speech. Feature extraction also plays a critical role in the diagnosis of depressed speech. Features such as cues, energy, and pitch would be customarily considered. The Mel Frequency Cepstral Coefficients (MFCCs), the Spectral Centroid Frequencies and Amplitudes (SCF / SCA), delay in the linear predictive group give us the detailed spectral information.

In [4] McGinnis, R. S. et al. explains about automating the process of depression diagnosis. Instead of the previous clinical process which involved structural interviews and standardized questionnaires, the paper proposes that data can be extracted with the help of wearable sensors that sense vocal reactions from children to various tasks. Signal features are derived from the voice signal which includes range, mean, root mean square (RMS), skew, peak to RMS amplitude, standard deviation, signal power within certain frequency bands, and features extracted from the autocorrelation of the signal. Various machine learning models such as k-nearest neighbor (kNN), decision tree, support vector machine with a linear kernel (SVM), and logistic regression (LR) models are proposed for the diagnosis.

In [5], a similar approach is followed by McGinnis, E. W. et al. for diagnosis in chlidren using a 3-minute speech task. Machine learning was employed with Trier Social Stress Task to diagnose depression in children of ages 3 to 7 years old. Audio data is processed via VAD that discriminates audio presence and absence in the human voice signal. Within each phase, parameters of audio signals were computed for every speech epoch. Features included mean frequency, speech epoch duration, zero-crossing rate of the audio signal, (MFCC), spectral flatness, dominant frequency, perceptual spectral centroid (PSC), kurtosis and skew of the power spectral density (PSD) etc. Classification models such models logistic regression (LR), support vector machine with a gaussian kernel (SG), and random forest (RF) are proposed to classify depressed and neutral speech.

In [6], Chu-Xiong Qin et al. proposes a model to build Multiple convoluted and recurrent layers that provide an integrated model for speech analysis for both acoustic features and also a language-based model. A high-level feature extraction approach using NMF (Non-Negative Matrix Factorization) is combined with multilingual deep neural network (DNN) training, and a CTC model is taken into consideration for speech analysis. A transfer learning approach is proposed.

In paper [8], Ying Yang et al. conduct an experiment with 57 participants as samples who were of different age groups ranging from 19 to 65 years and were suffering from depression at various levels. Through this experiment, it was seen that changes in the vocal prosody could determine how severe their depression was. SP mean and variability and mean and variability of F0 were considered the measures here. Hence it was proved through this experiment that change in fluctuation and rhythm of the voice/ voice prosody is a very common symptom of depression. Therefore, by using this information it is possible for researchers to come up with different techniques that use audio and voice clips to detect and diagnose depression in early stages.

In [10], Jinming Li,Xiaoyan Fu et al. proposed an approach where Depression AudioNet framework was used to extract

different features like MADN and MFCC. The authors also proposed that using joint tuning layers by making use of information related to contextual emotions will help increase the performance. The outcomes for MAE and RMSCE were seen to be 7.07 and 9.15 respectively.

In [12], a one dimensional CNN and Transformer methods are proposed by Genevieve Lam et al. to automate the detection of depression using data augmentation that is context aware and makes use of acoustic features. This method is based on topic modelling and it was observed that it was efficient for audio and text forms. One dimensional CNN and Transformer model separately achieved great F1 scores but when they were combined, they outperformed the existing performance and showed strong outcome of F1 Score 0.87

In [15] Sahar Harati et al. proposes a model that is predictive in nature on the features that are based on emotions. The comparision is provided between the models like Support Vector Machine, K- Nearest Neighbor, Hidden Markov Model with Gaussian Mixture Emissions and Grated Recurrent Unit that is based on RNN model for features extracted using different methods like Basic feature set, Switching Linear Dynamic System and the Emotional features. The architecture proposed in [15] uses the Long Short Term Memory with 2 hidden layers. The primary outcomes of this approach shows that the model that is proposed efficiently categorizes "depressed" and "improved" stages of Deep Brain Simulation Treatment with the Area Under the Curve being 0.8.

III. Methodology

A. Audio Data Preprocessing

Raw audio data is obtained by providing children with a speech task to perform. Participating children's approach to the task is recorded in .wav file format. This unstructured data is to be processed before analysis. We have used a data set containing 100 audio samples of children speech recording from Zenodo, an open access database.

The audio waveform if first low-pass filtered before they are digitized by sampling at the rate of 44.1kHz. Low-pass filtering is done to avoid aliasing, distortion or error. It is then processed using Voice Activity Detector (VAD) which is a speech detection technique to differentiate human voice from unwanted interference. It takes care of noise suppression, echo cancellation and mic beamforming. In reality, children with mental illness or with characteristics of introversion tend to avoid answering uncomfortable questions resulting in an irregular audio data with few blank parts. Thus the audio is further processed to retain only parts of data containing human voice. A median filter was used to smoothen very high noises and disturbances and remove long pauses between sentences while retaining the short pauses. Since children may take a many long pauses to think for responses, we have considered the median filter to have a permissible pause window of 15 sec. This is done to ensure the natural flow of speech. This step involved the extraction of acoustic features which will further undergo analysis.

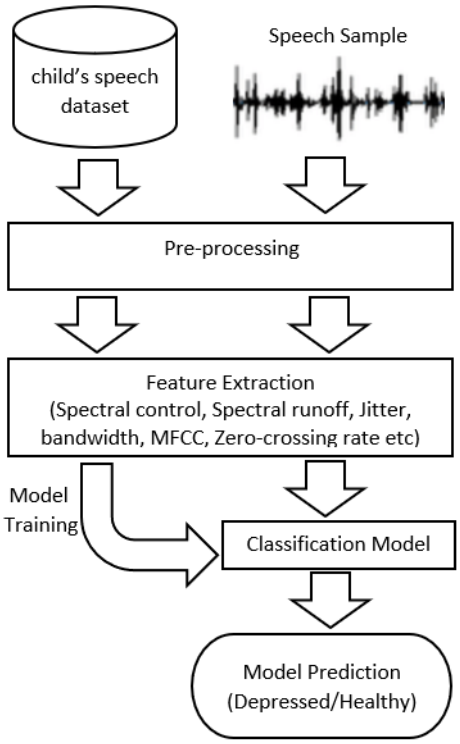

Fig. 1. System Architecture

B. Feature Extraction

Every audio signal consists of numerous features which are extracted and based on which, a classification model is developed to differentiate between the two classes - Normal child and Depressed child. We have incorporated a built-in python module called librosa for analysis annd extraction of basic audio functionalities. Some of the functionalities include Short-term Fourier Transform, Spectral Centroid, Spectral Runoff, Spectral Bandwidth, Zero-Crossing Rate, MFCCs and Chroma Feature.

Short-Term Fourier Transform (STFT) converts the signal to allow us to identify the amplitude of numerous frequencies at a particular time.

Spectral Centroid reveals where the spectral energy frequency is concentrated. Center of mass of the sound is identified. It returns us the number of frames present in each data (Figure 2(a) and 2(c)).

Spectral Runoff gives the runoff frequency for each frame returned by the spectral centroid, i.e, it gives a shape to the signal. It indicates the frequency very high frequencies drop to null. To achieve this, we ought to determine the portion of the bins in the power spectrum that is at lower frequency (Figure 2(b) and 2(d)).

Zero-Crossing Rate is used to measure signal's smoothness by measuring the Zero-crossing value at each segment of the signal. It helps differentiate between voices and noises. A

voice signal such as that of a human oscillates very slowly in contrast to voiceless fricatives.

Chorma Feature is a vector containing features that explain the amount of energy of each pitch type (C, D, E, C..) is in the signal. It offers an effective way of describing the measure of similarity between different adjacent segments of the signal.

These features are extracted into a csv file and used in decision making schemes like detection, knowledge fusion, speech recognition, classification etc. For depression detection, classification models that are proposed in this paper uses the above mentioned features to obtain information related to pitch, energy, various frequencies etc., that help identify children with mental illness.

*1) **MFCC**:* Mel-Frequency Cepstral Coefficient of any signal produces a set of features of small size (approximately 40 features for our audio data) to concisely explain the overall shape of the spectral. These features describe the characteristics of any audio signal, providing sufficient frequency channels to analyze the audio signal.

To extract features using MFCC, the signal is first split into very small sized frames. A set of 12 features are extracted at each frame. The advantage of using methods like MFCC and Wavenet is that they help capture detailed characteristic features of the signal from a frame as small as 20ms width. Therefore the signal is split into short frames.This is done because, the audio signal is constantly changing. In short time scales, it is assumed that the signal doesn;t change much and the features remain relatively stationary to make it easier for analysis.

MFCC obtains the power spectral for each frame which identifies the various frequencies present in the frame. These frequences are called formants (F0, F1, F2, etc.). F0 is related to pitch. F1, F2, and F3 are used to identify different vowels.

MFCC then carries out Periodogram spectral estimation to discern the minute differences between closely spaced frequencies. This effect increases with higher frequencies. Clusters of periodogram bins are summed up to to know the frequency at different frequency regions. This is carried out Mel Filterbank. As frequencies increase, it is less concerned about their variation and more concerned about the amount of energy occurring at every spot.

Filterbank energies are converted to logarithm scale because loudness of a sound cannot be heard on a linear scale. To increase loudness on a linear scale, energy will have to be amplified. This means that the large variations in the energy may not sound different and may result in incorrect features. Filterbank energies are then decorrelated by carrying out Discrete Cosine Transform (DCT) which produces a covarience matrix that can used to model the features (eg. CNN classifer).

*2) **WaveNet**:* WaveNet is a deep learning multilayered neural network used for producing raw audio waveforms and extract features. It carries out autoregressive learning with casual or dilated casual convolutions. It does predictive

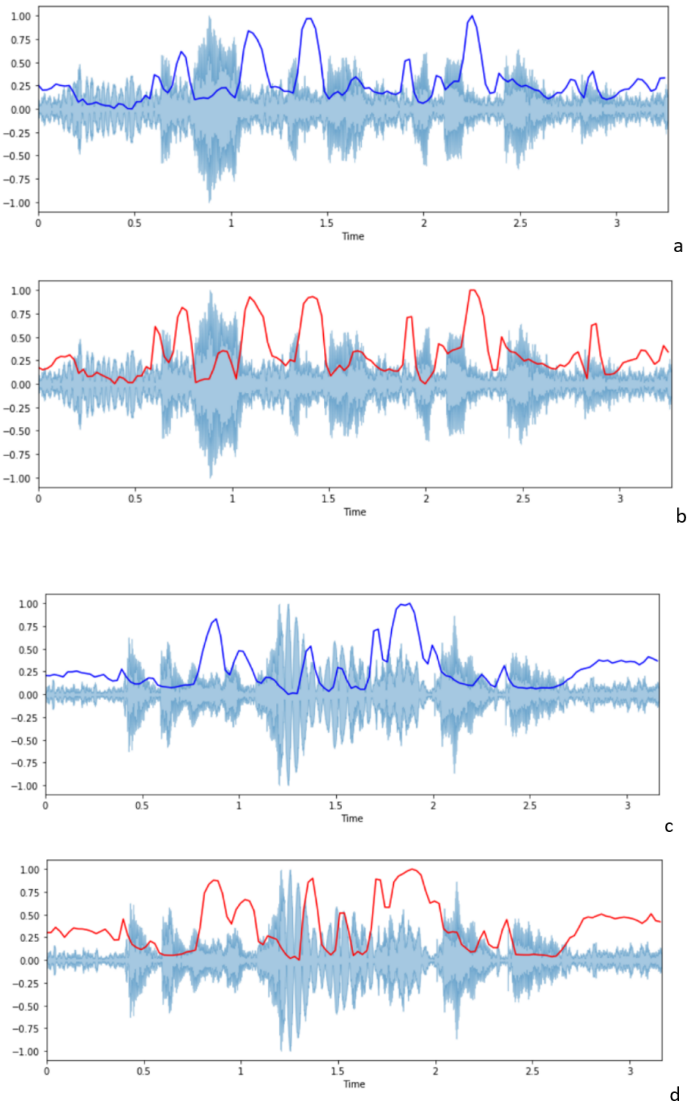

Fig. 2. 2(a) and 2(b) shows the Spectral Centroid and Spectral Runoff respectively for a audio signal of a child with no mental illness. 2(c) and 2(d) shows the Spectral Centroid and Spectral Runoff respectively for a child with depression and anxiety.

distribution for audio samples conditioned on the previously observed values. Which means that each time a sample is predicted, it is fed back to the model to predict the next sample. There is a convolutional sliding window that is used on the audio samples, which at every step tries to predict a sample value that it hasn't seen yet. Therefore it builds a network around these audio samples and learns the casual relationship by trying to predict a sample using a few previous steps of the neural network. WaveNet models with a casual convolution cannot violate the order in which the input data is modelled whereas in dilated casual convolutions, a filter is applied to the network to skip some input values with certain steps. Although the output is the same value, dilated casual convolution is often used as it's more efficient considering only the important input features to the next level and allows the to operate effectively

on a coarser scale.

In this paper, we have used WaveNet model for feature extraction. It produces a lot more features than MFCC and extracts every characteristic detail of the audio sample.

IV. Implementation

A. Model Development using MFCC Features

The proposed system involves the implementation of four binary classification models. They include Random Forest , Logistic Regression , Support Vector Machine (SVM) Gaussian Kernel, and SVM Linear Kernel. The clinically consent audio data set used has 42 observations of 8 features extracted using MFCC and is fed to the classifier model. It is then trained using the methods of supervised learning. After proper training and validation, the model predicts class values on the test set. A comparative study between the models is carried out to identify the most accurate model that results in least false positives or that which gives the best accuracy.

Pseudo code
Procedure: Classifer Model
input: Audio Dataset and Features
output: Classified results
Begin:
Step 1: Read dataset
Step 2: Split Dataset into train and test data
Step 3: Use train data to train the classifier models
Step 4: predict the result
Step5: Return the predicted class of each test audio.
Step 6: Calculate accuracy measures
End.

The resulting diagonsis of mental ilness predicted is used to calculate metric values such as Mean Square Error (MSE) or Mean Square Deviation (MSD) , Mean Absolute Error (MAE), R-Squared value, Root Mean Square Error (RMSE) and the Accuracy. The error rate (1 - Accuracy) is calculated to study the false predictions.

B. CNN Classifier modelling MFCC and WaveNet features

Convolutional Neural Network (CNN) is a multilayered deep learning neural network. Although best suited for image classification, it has shown a promise in audio classification, having the ability to predict with high accuracy when trained with a large dataset.

By using the concept of Transfer Learning, we have used WaveNet feature extraction model on top of CNN and compared the results with a CNN model built using features extracted from the inbuilt MFCC in Librosa.

42 observations in 8 features are extracted using MFCC and 848 observations among all features are extracted using WaveNet. These features are fed to the same CNN classifier model.

V. Result

The prediction diagnosis resulted from the four binary classification ML models shows that Random forest predicts with the highest accuracy of 94%, followed by Logistic Regression (84%), SVM-Linear (83%) and SVM-Gaussian (66%).

Model/Metrics	MSE	MAE	R-Squared	RMSE	Accuracy
Logistic Regression	0.16	0.16	0.30	0.41	0.84
Random Forest	0.05	0.05	0.76	0.23	0.94
SVM-Gaussian	0.33	0.33	-0.40	0.58	0.66
SVM-Linear	0.17	0.17	0.30	0.41	0.83

TABLE I
THE TABLE SHOWS THE PREDICTION VALUES OF VARIOUS METRICS.

Table 1 shows the values for 5 output metrics. MSE gives the approximate squared difference between actual and the estimate values. Random Forest has the least MSE of 0.05 which makes it the best estimator. Less value of MSE ensures that there is an agreement between the reality and the predicted value. MAE gives the mean prediction error which is the difference between the real value and the predicted value. MSE and MAE should be low while both testing and training. Low values during training and higher values during testing can result in model overfitting. R-squared shows how well the model fits the observed values around the fitted regression values. The model fits the data better when these two values have less difference and are unbiased. R-squared is shown using residual plots which reveal biased results and useless residual patterns. The table 1 shows that Random Forest explains the variability of the model around the mean value 76% of the time. The negative value in SVM-Gaussian shows that this model is predicting worse than the mean value. i.e., the observed values flow in the opposite direction of the fitted line on residual plots. Thus in SVM-Gaussian, the best fit of the model is very poor. RMSE for the prediction measures how far spread out the residuals or the prediction errors are from the observed values. Lower value of RMSE indicates that the model is fit better. By analyzing all the metrics, Random Forest is the best classifier. Logistic Regression model fits just as good as SVM-Linear as they both have similar values in Table 1.

The comparative analysis carried out between CNN using MFCC features and CNN using WaveNet features show that Wavenet produced more reliable and robust features where as due to simplicity of the MFCC structure, time taken it to compute the features was much faster than WaveNet with its complex neural network structure.

Despite these differences and WaveNet appearing more effective, they had almost equal accuracy. When MFCC features were used On CNN, it resulted in 76% prediction accuracy whereas the transfer learning model showed an accuracy of 78%. On future study, it was made clear that when small size data set is trained by using large number of features produced using WaveNet resulted in overfitting, thus reducing the predictive accuracy.

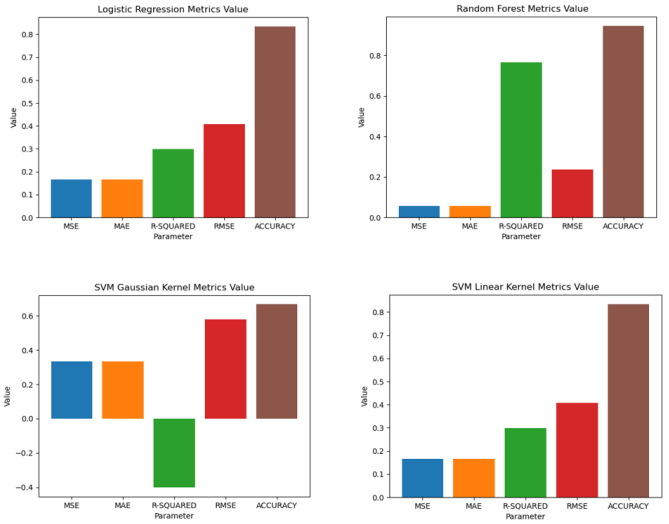

Fig. 3. Shows the output metrics for the four binary classification models using MFCC features.The metrics analyzed are Mean Square Error (MSE) or Mean Square Deviation (MSD) , Mean Absolute Error (MAE), R-Squared value, Root Mean Square Error (RMSE) and the Accuracy.

VI. Future Scope

The accuracy of Transfer Learning model can be increased by using large data sets. Since WaveNet and CNN are complex neural networks containing many layers, they require very large audio samples to produce accurate predictions without the problem of overfitting. By doing this, we can get the best fit of the model.

VII. Conclusion

The obtained results show that the analysis of child audio speaking task with machine learning classification models allows us to identify children facing psychopathology at a very early stage. This system beats the traditional clinical approaches in duration, complexity and accuracy yielding a predictive model that gives correct results 94% of the time. Time taken to extract features from MFCC is very less which helps reduce the complexity of the overall working system. help us It is very important that a child is diagnosed at a very early age to lead a quality and confident life in future.

References

[1] France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., Wilkes, M. (2000). “Acoustical properties of speech as indicators of depression and suicidal risk”. IEEE Transactions on Biomedical Engineering, 47(7), 829–837. doi:10.1109/10.846676

[2] Moore, E., Clements, M. A., Peifer, J. W., Weisser, L. (2008). “Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech”. IEEE Transactions on Biomedical Engineering, 55(1), 96–107.

[3] Cummins, Nicholas Epps, Julien Breakspear, Michael Goecke, Roland. (2011). “An Investigation of Depressed Speech Detection: Features and Normalization”. Proc. Interspeech. 2997-3000.

[4] McGinnis, R. S., McGinnis, E. W., Hruschak, J., Lopez-Duran, N. L., Fitzgerald, K., Rosenblum, K. L., Muzik, M. (2018). Rapid Anxiety and Depression Diagnosis in Young Children Enabled by Wearable Sensors and Machine Learning. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

[5] McGinnis, E. W., Anderau, S. P., Hruschak, J., Gurchiek, R. D., Lopez-Duran, N. L., Fitzgerald, K., … McGinnis, R. (2019). Giving Voice to Vulnerable Children: Machine Learning Analysis of Speech Detects Anxiety and Depression in Early Childhood. IEEE Journal of Biomedical and Health Informatics, 1–1.

[6] Chu-Xiong Qin, Dan Qu and Lian-Hai Zhang.”Towards end-to-end speech recognition with transfer learning”. Qin et al. EURASIP Journal on Audio, Speech, and Music Processing (2018) 2018:18

[7] Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo. “A Study of acoustic features for Depression Detection”, IEEE 2014.

[8] Ying Yang, Catherine Fairbairn, and Jeffrey F. Cohn, Associate Member, IEEE “Detecting Depression Severity from Vocal Prosody”, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 4, NO. 2

[9] Le Yang “Multi-Modal Depression Detection and Estimation”. 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)

[10] Jinming Li,Xiaoyan Fu ,d Zhuhong Shao, Yuanyuan Shang.“Improvement on Speech Depression Recognition Based on Deep Networks”, IEEE 2018.

[11] Huang, Z., Epps, J., Joachim, D., Sethu, V. (2019). Natural Language Processing Methods for Acoustic and Landmark Event-based Features in Speech-based Depression Detection. IEEE Journal of Selected Topics in Signal Processing, 1–1.

[12] Lam, G., Dongyan, H., Lin, W. (2019). Context-aware Deep Learning for Multi-modal Depression Detection. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[13] Zhao, Z., Bao, Z., Zhang, Z., Deng, J., Cummins, N., Wang, H., … Schuller, B. (2019). Automatic Assessment of Depression from Speech via a Hierarchical Attention Transfer Network and Attention Autoencoders. IEEE Journal of Selected Topics in Signal Processing, 1–1.

[14] Rani, B. (2016). Detecting depression: A comparison between I-Vector technique and fuzzy membership functions. 2016 International Conference on Inventive Computation Technologies (ICICT).

[15] Harati, S., Crowell, A., Mayberg, H., Nemati, S. (2018). Depression Severity Classification from Speech Emotion. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).