Original Article

Speech based Depression Analysis using Web Services and Convolutional Neural Networks

Dr. Adlin Sheeba[1], Dr. S. Nikkath Bushra[2], Dr. Dahlia Sam[3], E. Ajitha[4]

[1]Department of Computer Science and Engineering, St. Joseph's Institute of Technology, India, E-mail: adlinsheeba78@gmail.com

[2]Department of Information Technology, St. Joseph's Institute of Technology, India, E-mail: ferozbushra@gmail.com

[3]Department of Information Technology, Dr. M.G.R. Educational and Research Institute, India, E-mail: dahliasam@drmgrdu.ac.in

[4]Department of Computer Science and Engineering, St. Joseph's Institute of Technology, India, E-mail: ajithaeliza@gmail.com

DOI: 10.47750/pnr.2023.14.03.060

Abstract

People can experience depression, a mood disorder with wide-ranging effects. Depression can happen for many causes in daily life, which has an impact on our health. Stress, a lack of a work-life balance, and other factors make it a serious issue in the current era. It is a key factor in the overall global burden of disease, a leading cause of disability, and a suicide cause. Therapy is now the only effective treatment for depression. However, given the high demand for psychological therapy, it is challenging to find therapists for all potential depression situations. Additionally, there is no reliable biomarker to detect depression in an individual. The most reliable method for identifying depression is sound. Previous research has demonstrated that sad people exhibit fewer prosodic qualities of sound than a normal person. The paper employs deep learning neural networks to teach the computer how to distinguish between a healthy person's sound and a depressed person's sound using speech emotion identification to identify depression. Utilizing a convolutional neural network and the Mel Frequency Cepstral Coefficient, the relevant sound waves may be retrieved. The result is a trained model that can be used to make predictions. This model is then made available as a web service for service requesters to use.

Keywords: Web Services, Deep Learning, Depression Analysis, Neural Network, Speech.

I. INTRODUCTION

In the area of artificial intelligence, emotion recognition is a new application. It is the capacity to identify a person's emotions from speech, together with the key affective states that influence those feelings. This is because one's speech can convey their underlying emotions through their tone and pitch. The way one feels reflects their mental condition. Body language and facial expression are two ways that people naturally express their feelings to one another.

The selection of a speech multimodal database, the extraction of pertinent characteristics, and a suitable classification algorithm are all necessary for a SER system to be successful. The goal is to programme the machine to recognise human emotions and act in accordance with those emotions. In the real world, people used to reduce their stress levels by playing games, viewing funny movies and comic books, and listening to music. Since these solutions need a lot of time and are not totally automated, an automated system is preferred. Therefore, SER is essential in all areas to ensure that people can live without depression on a daily basis.

Over 50% of depression patients go undiagnosed, which results in inadequate medical care. The lives of individuals are continually impacted by these unreported situations. The enhancement of society and an increase in the level of living will result from identifying cases of depression and assisting in providing them with the necessary medical care.

There are numerous methods for identifying a person's emotions through speech. They entail pre-processing, which involves formatting the data before usage. Framing, feature selection, and noise reduction are examples of frequent pre-processing operations. The features chosen may be prosodic, spectral, or determined by the voice quality. Then, any auxiliary modalities, such as verbal elements or visual cues, can be included. Finally, the emotions are categorised using the proper classifier. The current algorithms identify them using a convolution neural network and combine prosodic data with spectral features to determine the emotion. However, there is no system specifically designed to identify users who are depressed with the intention of offering psychology support to the user.

The suggested approach seeks to identify depression in individuals based on the emotion shown in their speech. The six categories of emotions identified by the Discrete Emotional Model are sadness, happiness, fear, anger, disgust, and surprise. The sadness factor will be our main concern. The suggested system employs a certain set of characteristics to determine the degree of grief in a speaker's voice. The amount of sadness is then determined using a Convolutional Neural Network that only

considers the feeling of sadness and is accessible as a web service. This is used to suggest that a person receive more psychological care.

II. LITERATURE REVIEW

We looked at few research that used CNN to find SER.

The paper in [1] discusses the most recent advancements in classifier technology, including the application of convolutional and recurrent neural networks. There is also discussion of a list of current difficulties. Spectrograms, which are images generated automatically from audio recordings, were used in the experiment to apply data to residual CNNs [2].

In a straightforward model for speech emotion identification, convolutional neural networks are employed as a classifier. It makes use of 15 prosodic and spectral features together. Eight emotions are among them. The emotions present in this system are surprise, neutral, disgust, fear, joy, teasing, sadness, disgust, fury, and anger [3].

The authors describe depression and discuss testing and evaluation techniques used today and then uses data from earlier studies to establish speech as a sign of depression. Then, it groups the different characteristics found in well-known depression databases under different scores using the HAMD scale. Compared to traits with lower scores, those with higher scores have an impact on depression. Finally, it is discovered that speech characteristics have a strong association to depression [4].

The ability of speech investigation and arrangement to identify signs of depression in youngsters before they manifest as complete signs has been studied by KuanEe Brian Ooi et al. (2014) [5]. The labelled dataset that is used for depression estimate is the key bottleneck particularly when deep learning is employed [6].

Speech data from users must be analysed for depression. However, this poses privacy issues regarding the receiver end's potential use of the speech data. Deidentification is therefore made available. Deidentification is the process of hiding speech data that can include personal information while maintaining the data required for speech analysis of depression [7].

Numerous studies have examined the topic of cross-language emotion recognition from voice using machine learning algorithms.

By conducting some basic empirical analyses on cross-corpus emotions learning, these studies have mostly highlighted the necessity for in-depth investigation [8]. An attention-based CNN was constructed and tested using the IEMOCAP data set [9].

LSSED is a difficult large-scale English database that simulates real distribution and is used for speech emotion recognition. It has been discovered that current algorithms have a tendency to overfit limited databases, making it difficult to generalise them to actual settings [10].

In order to train future SER models for any language with few training instances, the authors have examined the viability of few-shot learning for SER models and suggested an effective approach (FMAML) that is superior than MAML. Three baselines were used as a comparison point for the authors' strategy. Unsurprisingly, in a few shots learning scenario, MAML based techniques outperform traditional supervised learning by a wide margin. [11].

The following elements are incorporated into system architecture to enable SER as effectively as possible: raw datasets, environments, and features. A new developed framework known as DeepResLFLB was suggested [12] based on these elements. Additionally, several activation functions can be used in each neural network segment to enhance DeepResLFLB performance [13][14][15][16][17].

III. METHODOLOGY

The speech-based depression analysis system's goal is to determine if a person is depressed or not and to detect any early warning signs of impending depression. Fig. 1 shows the architecture used for the proposed work known as SER-WS (Speech Emotion Recognition based on Web Services).

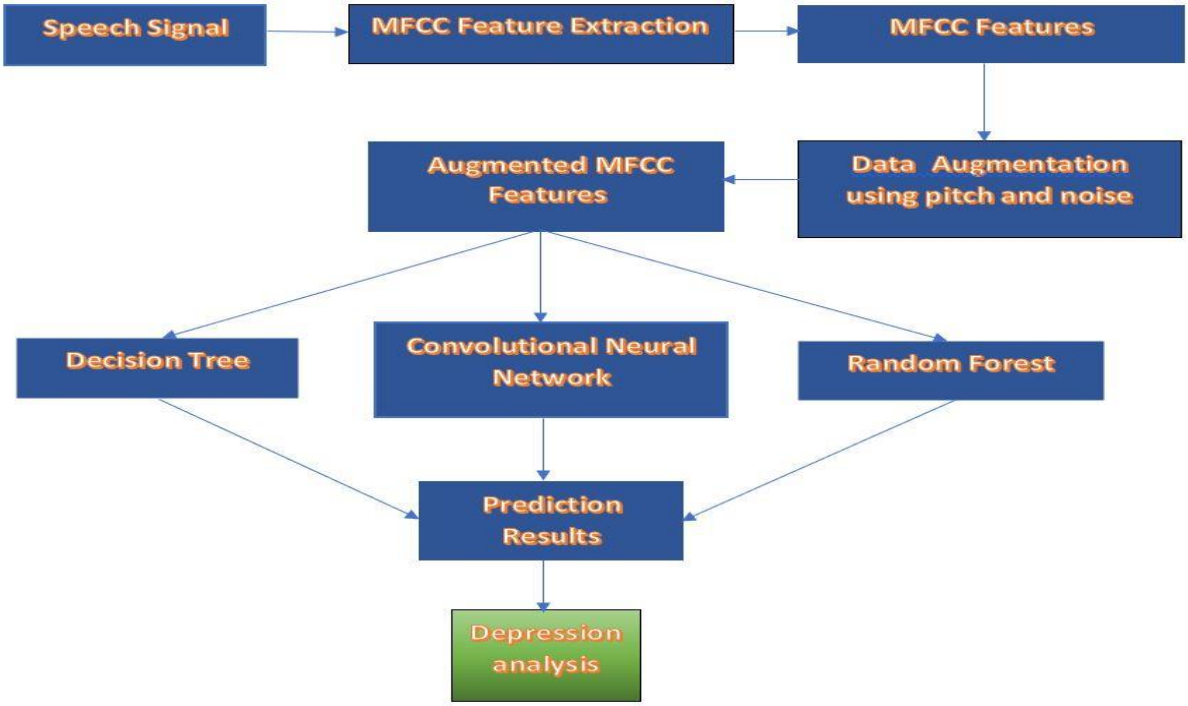

Fig. 1 Architecture of SER-WS

The system consists of 4 major components

A. Speech input

B. Feature extraction

C. Data augmentation

D. Convolutional neural networks

A. Speech input

Sound files are the input for depression analysis. Users' voices are recorded and then transferred to the other modules to create these sound files. However, it must go through a number of pre-processing stages before it can be used, like being converted to an appropriate sound file format and having the background noise removed. The sound file cannot be used until these pre-processing steps have been completed.

B. Feature Extraction

For feature extraction, MFCC is employed. The MFCC accepts human perception while accounting for speech frequency and takes in linguistic data while ignoring any background noise. Speech signals are used to create short frames. For each frame, computations for the Fourier transform and power spectrum are made, and the results are then scaled to the Mel frequency.

The Mel Filter bank processes the power spectrum to calculate the energy total for each frame. After that, the Filter bank energy log is found. The discrete cosine transform of the log Filter bank energy is applied, as seen in Fig. 2.

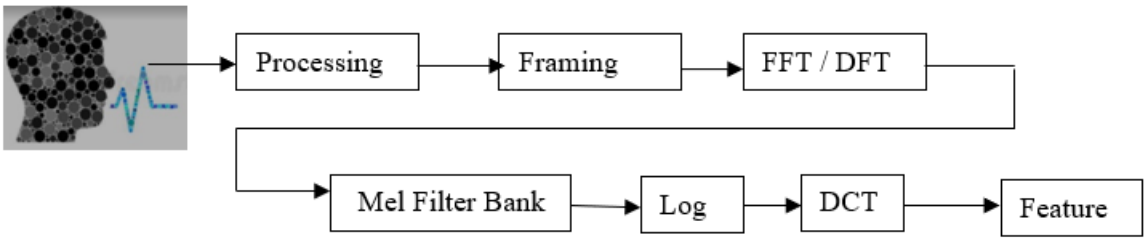

Fig. 2 Feature Extraction of SER-WS

C. Data Augmentation

The dataset can be expanded with the aid of data augmentation. This is advantageous because there aren't many high-quality datasets for speech. This contributes to expanding the training dataset. As a result, a sizable dataset of high quality is obtained. By altering the noise or pitch, augmentation is accomplished. In order to create a new, finished dataset, the original dataset is concatenated with the supplemented dataset.

D. Convolutional Neural Network

To train a model to recognise depression, convolutional neural networks are utilised. The RAVDESS dataset, a labelled dataset with 1440 files, serves as the input dataset. These files provide labels that describe the current emotion the user is feeling. After the training is completed then the model can be used to predict the user's depression.

Pseudocode

1. To obtain the labelled values for constructing the dataset, read the file names of the Ravdess dataset.

2. Label the emotions to identify the positive and negative ones. These are used to determine whether or not a person has depression.

3. Add Mel Frequency Cepstral Coefficients to the dataset using Librosa.

4. Use noise and pitch to enhance the data. Concatenate the result with the initial dataset.

5. Create a Convolutional Neural Network model, then train it using the dataset.

6. Access the user's voice file to determine whether or not he has depression using the trained model.

We define the emotion distribution as illustrated in Fig. 3. A dataset with our own labels is created utilising the Ravdess emotions and the emotions are read from the audio file names in which they are categorised. Negative feelings like fear, sadness, and anger are viewed as contributing to depression.

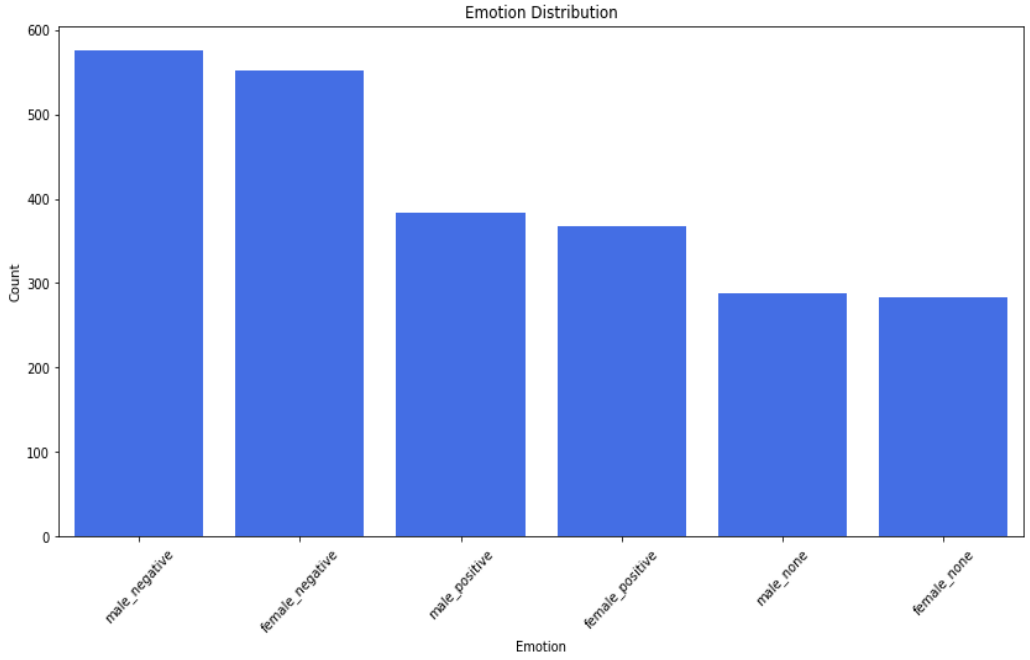

Fig. 3 Emotion distribution for SER-WS

The training of the CNN is finished, as seen in Fig. 4. Creating the training model and specifying the different CNN layers, like convolutional layer, pooling layer, and dropout layer, constitute the training process. The next step is to set activation functions and optimizers. Next, the model is fitted using training data.

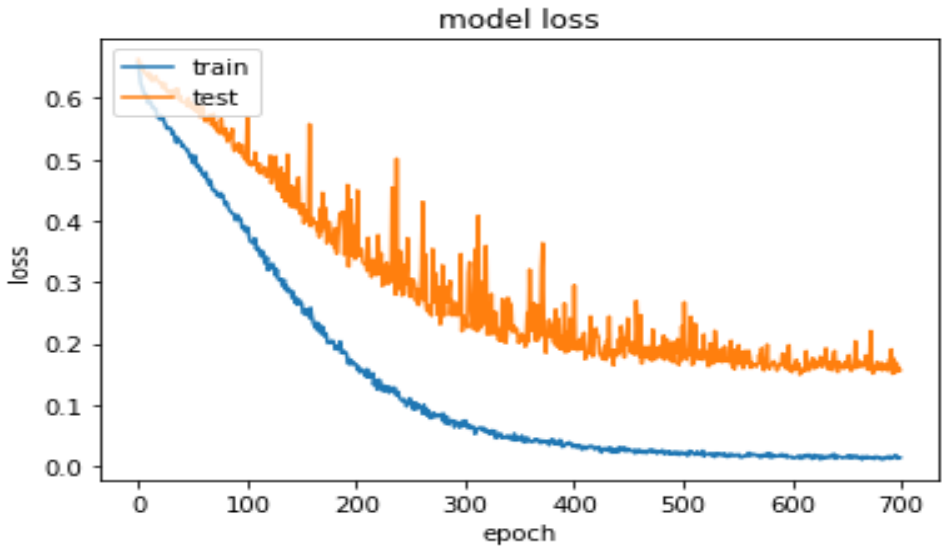

Fig. 4 Training the neural network

IV. PERFORMANCE ANALYSIS

Fig. 5 displays the classifiers' performance. 78% accuracy was achieved by the decision trees, which is reasonable. An accuracy of 80% was achieved using the Random Forest with a depth of 10. Only convolutional neural networks can achieve 85% accuracy, which is the highest possible level. Additionally, the model was able to receive greater training because to the usage of Data Augmentation techniques.

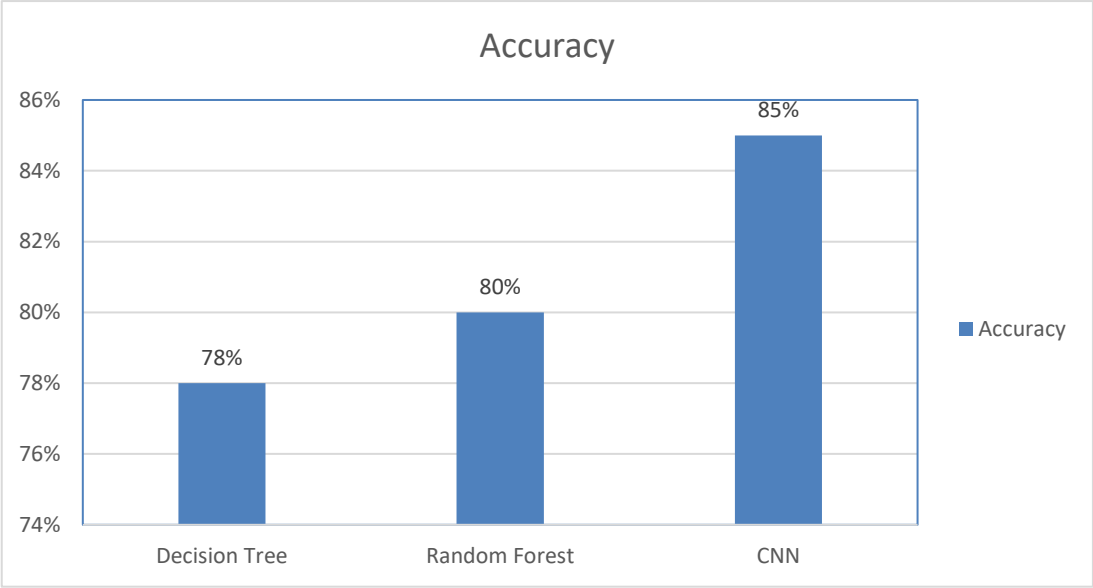

Fig. 5 Performance Graph for Accuracy Comparison

This makes it very evident that CNN is the ideal algorithm to use for this task because it consistently outperforms all other algorithms when given the Ravdess Dataset.

V. CONCLUSION AND FUTURE WORK

The SER based on CNN, Random Forest, and Decision Tree classifiers is shown. The signal processing unit, which extracts pertinent elements from speech signals and categorises them to highlight the emotion for each class, is a key component of SER. When compared to other machine learning techniques, CNN is demonstrated to be the best classifier. Due to its enhanced capability for Human Computer Interaction, research on autonomous SER is gaining traction. Doctors and service requesters can use the model to identify depression at an early stage because it is available as a web service. Mixed models of the methodologies can be used for better outcomes. Future demand for this work in the medical and psychiatric fields will be considerable due to the rise in incidences of depression among people worldwide. With the use of artificial intelligence, it may be simpler to diagnose and treat people with depression in the future.

REFERENCES

[1] Mehmet Berkehan Akçay and Kaya Oğuz,"Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers", Speech Communication, Vol. 116, pp. 56-76, 2020

[2] Karol Chlastaa, Krzysztof Wołka , Izabela Krejtzb,"Automated speech-based screening of depression using deep convolutional neural networks", Procedia Computer Science, Vol. 164, pp. 618-628, 2019.

[3] J. Nicholson, K. Takahashi and R. Nakatsu, "Emotion Recognition in Speech Using Neural Networks", Neural Computing & Applications, Vol. 9, pp. 290-296, 2000.

[4] Nicholas Cummins, Stefan Scherer, JarekKrajewski, Sebastian Schnieder, Julien Epps, Thomas F. Quatieri," A review of depression and suicide risk assessment using speech analysis",Speech Communication,Vol. 71, pp.10-49, 2015.

[5] KuanEe Brian Ooi, Margaret Lech, Nicholas Brian Allen, "Prediction of major depression in adolescents using an optimized multi-channel weighted speech classification system", Biomedical Signal Processing and Control, Vol. 14, pp. 228-239, 2014.

[6] Le Yang, Dongmei Jiang and Hichem Sahli, "Feature Augmenting Networks for Improving Depression Severity Estimation From Speech Signals", in IEEE Access, Vol. 8, pp. 24033-24045, 2020.

[7] Paula Lopez-Otero, Laura Docio-Fernandez, "Analysis of gender and identity issues in depression detection on de-identified speech", Computer Speech & Language, Vol. 65, pp. 101-118, 2021.

[8] Siddique Latif1 , Adnan Qayyum1 , Muhammad Usman2 , and Junaid Qadir, "Cross Lingual Speech Emotion Recognition: Urdu vs. Western Languages", 2018 International Conference on Frontiers of Information Technology (FIT), DOI: 10.1109/FIT.2018.00023, 2018.

[9] MingkeXu , Fan Zhang , Xiaodong Cui , Wei Zhang,"Speech Emotion Recognition With Multiscale Area Attention And Data Augmentation", https://arxiv.org/abs/2102.01813, 2021.

[10] Weiquan Fan,Xiangmin Xu, Xiaofen Xing, Weidong Chen, Dongyan Huang, "LSSED: A Large-Scale Dataset And Benchmark For Speech Emotion Recognition", arXiv:2102.01754, 2021.

[11] Anugunj Naman, Liliana Mancini," Fixed-MAML for Few Shot Classification in Multilingual Speech Emotion Recognition", https://arxiv.org/abs/2101.01356, 2021

[12] Sattaya Singkul, Thakorn Chatchaisathaporn, Boontawee Suntisrivaraporn, Kuntpong Woraratpanya, "Deep Residual Local Feature Learning for Speech Emotion Recognition", https://arxiv.org/abs/2011.09767, pp. 1-12, 2020

[13] Livingstone SR, Russo FA, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English", PLOS ONE, Vol. 13, pp.1-35, 2018

[14] Yue, J., Zang, X., Le, Y., "Anxiety, depression and PTSD among children and their parent during 2019 novel coronavirus disease (COVID-19) outbreak in China". Curr Psychol, Vol. 41, No. 8, pp.5723-5730, 2022.

[15] Silva, W.A.D., de Sampaio Brito, T.R. & Pereira, C.R. "COVID-19 anxiety scale (CAS): Development and psychometric properties", Curr Psychol, Vol. 41, No. 8, pp. 5693-5702, 2022.

[16] Chen, B., Li, Qx., Zhang, H. et al. "The psychological impact of COVID-19 outbreak on medical staff and the general public", Curr Psychol, Vol. 41, No. 8, pp. 5631-5639, 2022.

[17] Robles-Bello, M.A., Sánchez-Teruel, D. and Valencia Naranjo, N. "Variables protecting mental health in the Spanish population affected by the COVID-19 pandemic", Curr Psychol, Vol. 41, No. 8, pp.5640-5651, 2022.