IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SPEECH EMOTION ANALYSIS USING MACHINE LEARNING FOR DEPRESSION RECOGNITION: A Review

Ashish Nayak [1], Royson Clausit Dmello[2], Sakshi S Bangera[3],Manjunath[4], Bhavya S[5]

[1,2,3,4]Student, [5] Senior Assistant Professor

[1,2,3,4,5]Electronics and Communication Engineering

[1,2,3,4,5]Mangalore Institute of Technology & Engineering, Mangalore, India

Abstract: Depression is a mental disorder that can affect a person's physical health and well-being. Untreated depression can affect a person's quality of life and cause a number of other symptoms. Current diagnostic methods are limited to clinical interventions. Therefore, this program is proposed to identify early depression and provide assistance with clinical management decisions during treatment.

Communication is important in communicating our thoughts and ideas to others. Machine Learning is rapidly developing in its ability to bring the most complex systems to everyday use. Intelligent systems work together and work with minimal user effort, relying heavily on voice input. The purpose of this article is to introduce various speech algorithms to detect depression using machine learning.

Index Terms - **DAIC-WOZ, LSTM, RAVEDESS, SVM**

I. INTRODUCTION

Depression is one of the major mental health problems facing people of all ages and genders in recent years. Work ethic, stressful life, emotional inequality, family disruption, and community health lead to depression. Depression is a very common and serious illness and can have a detrimental effect on a person's daily life. This attitude often leads to feelings of sadness, loss of interest in activities and activities and rarely leads to suicide. It affects the natural ability to work at work and at home. Depression affects one in 15 adults a year and the risk in women is twice as high as in men [1].

Behavioral symptoms can be seen in early detection of depression. In particular, speech analysis is considered to be an important factor in discriminating between depressed patients and healthy individuals, due to the influence of emotional status on voice quality. Depressed patients, in fact, show patterns of prosodic features of different expressions from healthy studies: the number of stops is high over time speech quality is low. Such acoustic features as loud noise, tone and low tones depressed patients rather than healthy controls [2].

Fortunately, depression is a treatable disease. Physicians perform clinical assessments based on patient reporting of their symptoms as well as general mental health questions such as a list of stress seizures. Depression assessment is a list of self-explanatory questions taken by the patient In response, points are given automatically. Patient Healt Questionnaire is a commonly used test and consists of nine clinical questions. Even when patients report their symptoms on their own, physicians accurately detect only half the time [3].

II. LITERATURE SURVEY

The Table 1 represents the literature review of various papers that compares the different algorithms, list of Data sets used and the accuracy of the results.

Table 1: Performance of different methods on dataset

Paper	Dataset	Algorithms	%Accuracy
[1]	DAIC-WOZ	SVM	70.2
		Random Forest	59.7
		Logistic Regression	63.8
[3]	DAIC-WOZ	LSTM	76.27
[5]	DAIC-WOZ	MFCC-AU LSTM	95.38
[14]	DAIC-WOZ	SVM	70.58
		Gaussian Mixture Model	88.23

The main purpose of the paper is to detect stress through speech. Acoustic features are used to train the separation algorithm to determine whether a person is depressed or not in this topic. Class dividers are trained using DAIC-WOZ database. Prosodic, Spectral, and Voice control features are extracted using the COVAREP tool kit. The method of making multiple small samples is used to eliminate the phase inequality produced by the data sets used. Results from classification strategies such as Logistic Regression, Random Forest, and SVM are compared. CureD is an Android self-test app. By using the app, a psychiatrist can perform an analysis in a remote area where his physical appearance does not occur. Under the guidance of a specialist psychiatrist, the application is evaluated and 90 percent accuracy is performed. After the SMOTE analysis, SVM had the best accuracy of all the separators used. The model is trained in external databases based on external pronunciation and the database can be expanded by adding audio samples of various native words and languages to the existing database [1].

A machine learning approach is used in paper [2] to suggest an automatic system. The SVM classifier is utilized in this approach to discern between healthy and sad states by examining voice data. Speech signals from psychiatrists were collected for training purposes, totaling 149 samples, 62 of which are healthy and 87 of which are problematic. All speech signals were recorded at a sample rate of 44.1 kHz with a resolution of 32 bits. Jitter, MFCC, Derivatives of cepstral coefficients, and spectral centroid are some of the acoustics properties covered. Polynomial and radial basis functions are the most commonly utilized kernel forms (RBF). Performance of this approach is evaluated on the basis of sensitivity, accuracy, precision and ROC area. The proposed simplified approach's key advantage is that it may be simply implemented in the clinical setting, even by individuals without advanced computer abilities. This technique has an overall accuracy of roughly 85 percent.

A deep Recurrent neural network-based framework is proposed in Paper [3] to detect depression and estimate its severity level from speech. The effectiveness of this method is tested in multimodal and multifeatured tests. This system makes use of the DAIC-WOZ dataset. The recordings are split into 2 groups in this method. The audio segment of the interviewer and audio segment of participants. Audio segment of interviewer is discarded and only participants audio segment is used. Low-level features are extracted from preprocessed audio recordings and are defined as MFCC coefficients. The spoken signal is initially separated into frames using a 2.5s windowing method with 500ms intervals. The suggested model's baseline is based on LSTM. Different augmentation techniques such as noise injection, pitch augmentation, shift augmentation, and speed augmentation are used to create new audio segments. The overall accuracy of this model is 76.27%.

Based on the Dempster-Shafer evidence-based approach, paper [4] proposes a multimodal fusion emotional awareness program. To use this model the author has selected a data set of 25 video clips containing five types of emotions such as happiness, sadness, relief, irritability and disgust. The author proposes to use the DS theory to combine an ECG-based emotional awareness model with an EEG-based emotional awareness model to integrate decision-making. HR and HRV features are derived from ECG data and are distributed using the Bi-LSTM network. P-wave, Q-wave, R-wave, S-wave, and T-wave mathematical properties, including definition, median, Sd, limit, minimum, and the difference between high and low values. Four types of features are used as inputs in the SVM separator to separate EEG signals. Based on time information from the previous minute, the LSTM network predicts the outcome of the next minute. Experimental results suggest that for more emotional information the combination of EEG and ECG signal information should be used and emotional accuracy can be improved by multimodal integration.

Convolutional neural networks and long-term memory modalities are described in the paper [5]. The DAIC-WOZ database is divided into three sections: 80 percent for training, 10 percent for certification, and 10% for testing. For binary stress classes, LSTM-based audio features perform slightly better than CNN, with 66.25 percent accuracy and 65.6 percent, respectively. In this paper, audio, visual, and textual data are analyzed, and are used in a wide range of in-depth learning methods for unimodal and multimodal expression. Only patient response is used for pre-noise processing. Activation function, step by step,

normal stop, and optimization are all used in the LSTM layer. The results of this model reveal that the proposed LSTM-based architecture exceeds the CNN-based architecture in terms of learning the dynamic representation of multimodal data.

Paper [6] recommends that de-identification be utilized to protect patients' privacy. This research focuses on a comprehensive investigation of depression detection utilizing voice conversion ad other de-identification techniques. In this research, various aspects are examined in order to recognize depression. A few features include assessing how this system performs when the gender is changed, comparing the performance of speaker dependent and speaker independent source settings in parallel using the data given, and comparing the performance of the two alternative voice conversion methods. This study's depression analysis technique is based on the i- vector. It allows for the compression of all low-dimensional features of a voice recording, such as gender, age message, and so on. For speech synthesis and voice conversion, a Generative Adversarial Network (GAN) is used. It combines two neural networks that are in competition: a discriminator and a generator. Speaker de-identification alters a source's vocal characteristics to make it sound like a different person.

The paper [7] investigates various methods for predicting depression. Decision trees, SVM, Logistic Regression, and KNN classifiers are among examples. Wearable devices and mobile phones are used to collect behavioral data. Twint is a tool that the author uses to determine whether or not a person is depressed. Following the database extraction, a model is created and trained to achieve the desired outcome. Finally, it detects the disease via the social media platform Twitter and determines whether or not the individual is depressed. The keywords are retrieved from Twitter using the application Twint to detect depression. In this study, the author offers the Deep Learning Mechanism for detecting depression. This method, according to the author, has the potential to improve accuracy. The study also includes an example in which the author observed 46 people and identified 85 distinct traits.

The speech emotion recognition system is defined by Paper [8] as a collection of approaches for classifying speech and detecting emotions from it. The Acted Speech Emotion Database, Elicited Speech Emotion Database, and Natural Speech Emotion Database are the three elements of the database for speech emotion recognition. Preprocessing, Framing, Windowing, and Voice Activity Detection are the four processes in speech emotion recognition. After extracting the database, the first stage will be preprocessing, which will be utilized to train the model in a SER system. Speech is continually segmented into fixed length segments during signal framing, which is also known as speech segmentation. The window function is applied to the frames in the following phase. This procedure is mostly utilized during fast Fourier transforms (FFT) to minimize the effect of leakage. So, in the final stage, the created voice speech will be generated using vocal speech vibrations, which will be detected. The speech is normalized in the following steps, and any noise is recognized. Feature selection and dimension reduction are critical steps in speech emotion recognition (SER).

A real-time emotion identification system that recognizes live speech is examined in paper [9]. RAVDESS and SAVEE databases were used in this study. Basically, this work extracts and analyses 34 audio characteristics. Gradient Boosting is used to classify emotions using the trained models. Other classifiers include Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). Four emotions are examined in this book. Anger, sadness, neutrality, and happiness are all emotions. There are three key steps in this system. Features extraction, feature normalization, and categorization are examples. pyAudio Analysis extracts 34 audio attributes from each audio file. The feature normalization stage is critical because the accuracy outcome is dependent on it. Normalization aids in the weighting or normalization of all features. Gradient boosting, SVM, and KNN are the three primary features of the third stage. In this stage, the emotions are classified using the three classifiers.

A study on long short term memory for the diagnosis of mood disorders is presented in paper [10], which is based primarily on evoked speech replies. The MHMC emotion database was used for both evaluation and training the model. The algorithm that was used to adapt the database id HSC. The CNN technique is utilized to train the model using the altered MHMC sequences. A toolbox called openSMILE is used to extract datasets, mostly for low-level feature extraction. The emotion profile evolutions were characterized using LSTM (LP). LSTMs were created to solve the problem of long-term data storage and reliance. Because the response may contain some irrelevant information in this study, an addition method called an attention mechanism is utilized to highlight the crucial or only required content in the response.

The main objective of the paper [11] is to evaluate the speech and behavioral factors associated with clinical depression in clinically depressed and non-depressed speakers for automatic analysis. Brno Phoneme Recognizer is used by the author to automatically separate phonemes from audio samples. The OpenSMILE Open Speech Toolbox is used to extract the performance values of 88 acoustic speech acoustic feature. Both the AViD and DIAC-WOZ informational data showed similar functional patterns of differentiating language-length stress, speech sign features, and development according to a set of middle and front vowels. Depressed speakers have a short vowel duration and a slight variation of the lower, back, and circular vowel areas, depending on the compression features of the language on both sets of data. According to the given speech qualities, the median length of the middle vowels was much shorter between the parameters in both the DAIC-WOZ and Avid data sets, while the background, circular, thick, low, and diphthong sets were longer.

In study [12], an algorithm was developed to distinguish between depressed and non-depressed people based on their twitter status and tweets. R studio was used to do a qualitative analysis. R is a programming language designed to improve quantifiable analysis. The proposed approach is used to evaluate the twitter dataset. Data from Twitter is extracted and shown in an excel sheet. Different scores are assigned to the various emotions. Positive, negative, and natural sentiments are assessed. A score is gained for pleasant feelings, but none is obtained for negative emotions such as disgust, wrath, and so on. The sentiment values of the identical tweets are retrieved and kept. Rows and columns separate the emotions and scores. The dataset could provide detailed information about the tweets. The study is conducted using text messages, but it is also possible to do analysis using photographs and sounds.

In order to effectively measure the severity of the stress arising from sound samples, paper [13] proposes a combination of learned manuscripts that are hand-painted and deep. The author introduces a process based on Deep Convolutional Neural Networks (DCNN). DCNNs are designed to extract in-depth readings from spectrograms and raw speech waves. The text adjectives known as the extended local binary patterns of median robust (MRELBP) are then manually extracted from

spectrograms. The author proposes a well-integrated tuning layer to combine raw speech with a DCNN-based spectrogram to improve model performance in stress perception. Assists in capturing additional data within hand-crafted and in-depth features. The studies were conducted on the depression sites AVEC2013 and AVEC2014. Compared with other methods based on sound signals, the results suggest that the adopted strategy is more tolerant and effective in diagnosing depression.

A speech-based depression detection system is described in Paper [14], which can be used as a screening tool to help depressed adolescents. The software MATLAB R2010a was used to create this system. The speech signal was first pre-processed, and the retrieved characteristics produced statistically significant results. Second, the prosodic, spectral, glottal, cepstral, and Teager energy operator (TEO) categories, as well as their combinations, were studied. For training, the DAIC-WOZ database, which is part of a bigger corpus, is used. The database included 85 sessions of associations lasting between 7 and 33 minutes. When employed alone, the TEO-based features outscored all other features and feature combinations. Glottal-based feature categories were shown to have a higher accuracy of 89.41% in the Gaussian mixture model and 41.17 percent in the Support Vector Machine classifier. The author also points out that a sad adolescent may not show visible indicators of depression. As a result, there are no set criteria for detecting depression, especially when the child is establishing new roles within the family, dealing with independence, and making academic and career decisions.

Paper [15] uses a person's speech signal to assess their depression condition. When patients vocalized three types of long vowels, speech signals were gathered. The openSMILE software was then used to extract the acoustic features. Weka software was used to select the features. Finally, a 4-fold cross-validation strategy was utilized to build an algorithm that accurately measured the severity of HAM-D score from speech signals for each long vowel /Ah/, /Eh/, and /Uh/, with accuracy of 75.5 percent, 78.7%, and 68.9%, respectively.

The purpose of paper [16] is to recognize emotions. DNN (Deep Neural Network) is employed. The Convolutional Neural Network model used in this research retrieves features from a raw signal. The raw data is divided into a 20-second sequence and used as an input after processing. To extract information from the raw signal, the kernel size is 8. The maximum pooling size is determined based on the kernel size in order to reduce the signal's frame rate. To capture the information in the data, two layers of LSTM are used.

The purpose of the paper [17] is to construct and apply a method of extracting a novel element to identify different emotions. The RAVDESS, which includes eight sensors, the Berlin (Emo-DB) database, which contains seven senses, and the SAVEE website, contains. seven senses, used in the study. The proposed method has a time limit as determining the deviation of most feature vectors with a few degrees takes longer than dealing with only one degree of deviation. The results of a large number of studies indicate that the most important factors have a direct impact on the accuracy of the categories. Working on RAVDESS is difficult because the two recordings were almost identical. Compared with the standard SER methods, the accuracy of the categories achieved in this work was approximately 86.1 percent, 96.3 percent and 91.7 percent on the RAVDESS, Emo-DB and SAVEE sites, respectively.

By aligning speech signals with visual sequences, paper [18] represents a way of perceiving emotions. Speech signals are first converted to acoustic features, which are then used to synchronize image sequences by transmitting them to another network. To improve the performance of emotional awareness, these three networks are integrated using a unique integration strategy. Several strategies based on the features of FER systems have been investigated. These methods use the image to locate the surface and extract geometric features or appearance from it. The link between the parts of the face is one of the geometric features. Speech symbols contain language content and clear paralinguistic information about speakers, such as emotion Many speech recognition algorithms, unlike FER, produce acoustic features because learning from the end (i.e., one-sided CNN) cannot produce features that work automatically. Speech recognition systems contain language content as well as vague paralinguistic information about speakers, such as emotions. In contrast to FER, many forms of communication produce acoustic features because, compared to acoustic features, end-to-end learning cannot automatically produce useful feature Summary and Observations.

According to the examined literature, multiple machine learning algorithms were used to analyze emotions from voice signals. The extraction of relevant features that may be utilized to train the machine learning model is the first step in the process. Acoustic aspects such as prosodic, spectral, glottal, and voice control were given priority. A toolkit called open SMILE is used for feature extraction, mostly for low-level feature extraction. It was also attempted to leverage multi-modal elements such as audio-visual to increase the machine learning algorithms' capacity to recognize depression.

DAIC-WOZ [20], RECOLA, TESS, and RAVDESS, all of which are freely available for research purposes, were heavily used. Some of the studies make use of datasets that incorporate data acquired from volunteers. The optimal dataset selection has a significant impact on the final result. As can be seen, dataset biases such as gender bias or age bias might have an impact on the effectiveness of the final machine learning model trained on such datasets to analyze emotions or depression. It's also crucial to keep a healthy ratio between sad and nondepressed data, as well as across samples related to other emotions, so that the model isn't distorted by majority sample bias.

It should also be mentioned that appropriate models for predicting emotional information from speech must be created using suitably large emotional speech corpuses. The creation of accurate prediction models and the construction of an adequate emotional speech corpus are the main concerns. The scarcity of data for the study makes it challenging to develop improved models. It's also tough to share information on depression recognition because it's so private.

The investigation employed machine learning algorithms such as SVM, Decision Trees, Logistic Regression, and KNN classifiers. The SMOTE approach was employed to improve the model's accuracy. Deep learning models such as CNN, ANN, and LSTM, on the other hand, outperformed traditional machine learning approaches.The models mentioned were trained on a dataset based on foreign accents, however the dataset might be expanded by adding audio samples of other native accents and languages to the existing datasets, extending the research's scope beyond linguistic barriers.

III. Conclusion

Depression is a significant medical condition. The proposed system is designed to reduce human intervention in the process of diagnosing depression in an individual. Speech is regarded as essential. The lack of publicly available speech databases made developing a well-trained model difficult. Keras and Scikit Learn are used to create the depression recognition system. The Distress Analysis Interview Corpus (DAIC) database was used to detect clinical depression in speech. To increase system performance, future developments in the proposed system would include implementing the model utilizing a large audio dataset, as well as merging face expression information recorded via video and adding languages for diagnoses, namely the local languages spoken in India. We also hope to integrate our proposed technology into a real-time depression detection program to provide emotional support.

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

[2] L. Verde et al., "A Lightweiht Machine Learning Approach to Detect Depression from Speech Analysis," 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), 2021, pp. 330-335, doi: 10.1109/ICTAI52525.2021.00054.

[3] Emna Rejaibi,Ali komaty,Fabrice Meriaudeau,Said Agrebi,Alice Othmani, "MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech", Biomedical Signal Processing and Control, Volume 71, Part A, 2021, 103107, ISSN 1746-8094, doi.org/10.1016/j.bspc.2021.103107.

[4] Tian Chen, Hongfang Yin, Xiaohui Yuan , "Emotion recognition based on fusion of long short-term memory network and SVMs" , Digital signal processing , School of Computer Science and Information Engineering 230009 (2021).

[5] Muhammad Muzammel, Hanan Salam, Alice Othmani, "End -to -end multimodal clinical depression recognition using deep neural networks: A comparative analysis.", Computer Methods and Programs in Biomedicine (UPEC), LISSI, Vitry Sur Seine 94400 -2021

[6] Paula Lopez-Otero, Laura Docio-Fernandez, "Analysis of gender and identity issues in depression detection on de-identification speech", Computer Speech & Language, E.E Telecommunication campus – 2021 dio.org/10.1016/j.csl.2020.101118.

[7] P. V. Narayanrao and P. Lalitha Surya Kumari, "Analysis of Machine Learning Algorithms for Predicting Depression," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020, pp. 1-4.

[8] Mehmet Berkehan Akcay, Kaya Oguz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers" , Speech Communication, Department of Computer Science Engineering doi.org10.1016/j.specom.2019.12.001 (2020).

[9] Iqbal, A. and Barua, K. "A real-time emotion recognition from speech using gradient boosting" In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (2019), IEEE, pp. 1–5

[10] Kun-Yi Huang Chung-Hsien Wu, Ming -Hsiang Su, " Attention based convolutional neural network and long short-term memory for short -term detection of mood disorders based on elicited speech responses." ,Pattern recognition , Department of Computer Science and Information Science Engineering. 0031-3203(2019).

[11] Brain Stasak , Julien Epps, Roland Goecke , " An investigation of linguistic stress and articulatory vowel characteristics for automatic depression classification" , Computer speech and Language 53(2019) 140-155 .

[12] A. Sood, M. Hooda, S. Dhir and M. Bhatia., "An Initiative to Identify Depression using Sentiment Analysis: A Machine Learning Approach," Indian Journal of Science and Technology, 2018, Vol 11(4).

[13] Lang He, Cui Cao., "Automated depression analysis using convolutional neural networks from speech" , Journal of Biomedical Informatics , NPU-VUB joint AVSP Research Labs, dio.org/10.1016/j.jbi.2018.05.007 -2018.

[14] P. R. Parekh and M. M. Patil, "Clinical depression detection for adolescent by speech features," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), 2017, pp. 3453-3457.

[15] Y. Omiya et al., "Estimng depressive status from voice," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 2795-2796

[16] Panagiotis Tzirakis ,Jiehao Zhang , Bjorn W. Schuller, " End-to-end Speech Emotion Recognition using Deep Neural Networks". Department of computing Imperial College, 978-1-5386-4658-8, 2018 IEEE

[17] Husam Ali Abdulmohsina,1, Hala Bahjat Abdul wahabb, Abdul Mohssen Jaber Abdul hossen, "A new proposed statistical feature extraction method in speech emotion recognition" Computer Science Department, Faculty of Science, University of Baghdad, 2021. 0045-7906/©2021PublishedbyElsevierLtd.

[18] Sung-Woo Byun, Seok-Pil Lee, "Human emotion recognition based on the weighted integration method using image sequences and acoustic features", https://doi.org/10.1007/s11042-020-09842-1 - 2021

[19] Livingstone SR, Russo FA (2018) 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 13(5): e0196391.2018, https://doi.org/10.1371/journal.pone.0196391,201

[20] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, Louis-Philippe Morency, "The Distress Analysis Interview Corpus of human and computer interviews", Proceedings of Language Resources and Evaluation Conference (LREC), 2014.