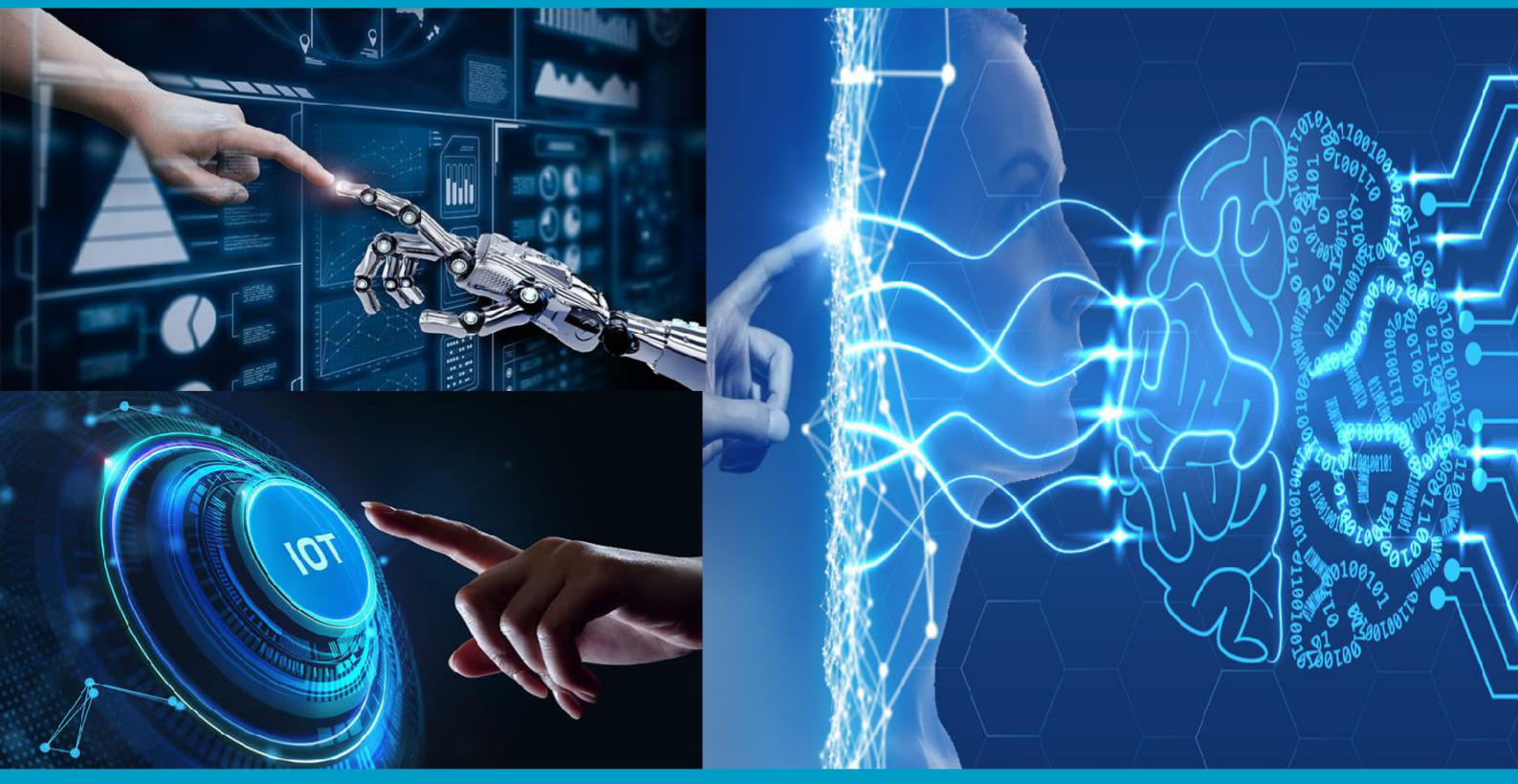
IJMRSETM

e-ISSN: 2395 - 7639



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

Volume 11, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.580

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 3, March 2024 |

Depression Detection using CNN Model

Sk.Wasim Akram[1] , Inkolu Alekya[2] , Kalavakolanu Vanaja[3] , Mahammad Irfan[4] , Katta Rahul[5]

Assistant Professor, Department of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh, India[1]

UG Students, Department of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh, India[2345]

ABSTRACT: Millions of people experience depression because mental disease is not treated and cared for in a timely manner. Traditional methods have a very low diagnostic accuracy rate; PHQ scores and patient interviews were used to detect depression. Depression, bipolar illness, anxiety disorders, and sleep disorders are all primarily brought on by it. In some instances, it might even lead to suicide and self-harm. It is therefore a challenging task to identify those who are depressed and to provide therapy as soon as feasible. In order to accurately determine whether a person has depression or not, a deep learning architecture is presented in this work. We trained the convolutional neural network model on an audio dataset in order to determine an individual's level of depression. The depression can be precisely identified using the suggested method.

KEYWORDS: Depression, audio , CNN , Deep Learning

I. INTRODUCTION

A mental illness called depression is characterized by enduring emotions of melancholy, emptiness, interest loss, and joylessness. Globally, depression has been acknowledged as a contributing factor to disability. It causes bad thoughts and has an effect on one's physical health. Millions of people experience depression because mental disease is not treated and cared for in a timely manner. Suicide is said to have depression as one of its causes. Traditional methods have a very low diagnostic accuracy rate; PHQ scores and patient interviews were used to detect depression. Depression, bipolar illness, anxiety disorders, and sleep disorders are all caused mainly by it. A person with mild depression may think harmful thoughts about harming themselves and other people. A person suffering from depression will eventually stop engaging in daily activities. Despite the fact that psychotherapy and medicine are useful therapies for depression.By seeking early therapy from a psychotherapist, depression can be healed in its early phases. Psychotherapists interview patients to learn about their mental health and use the Hamilton Rating Scale to measure each patient's degree of depression. This work focuses on methods for reliably identifying sadness from speech recordings found in audio files. It is difficult to design a system that can encode discriminant traits for depression identification since different recording settings, such as noise and disturbance, can significantly affect how a voice seems. It is difficult to gather a big dataset with labels that contain some noise and uncertainty. The best performance for identifying depression is offered by deep learning systems. In order to reliably identify depression, a deep learning architecture is presented in this research.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 3, March 2024 |

II. LITERATURE REVIEW

S. No	Author	Title	Published Year	Methodology/ Dataset	Limitations
1	Sonam Gupta	Psychological Analysis for Depression From Social Networking Sites	2022	Twitter dataset ML and Lexicon approach.	It can help depression detection only through social Media. A large number of world do not use it. So there may be chance of wrong results drawn.
2	Monika Gahalawat	Explainable depression Detection via head motion patterns	2023	Black Dog dataset, Kineme Discovery	Experiments are restricted to classification tasks involving a discretisation of depression scores.
3	Adam Valen Levison	Using audio data to facilitate depression Risk Assessment in PHC	2023	Facebook data TPOJ	Data is only from 24 individuals which is small sample size.
4	Fuxiang Tao	Improving Speed and Performance of Depression Detection	2023	Androids Corpus, LSTM+SVM	Using only linear layers in approach.
5	Nawshed	Deep temporal modelling of clinical depression through Social Media Text	2023	Mixed datsets, DESVA algorithm and Depression Score algorithm.	Depression is calculated on a highly imbalanced dataset

In the paper [1] , authors have chosen and trained the twitter data for the study. On datasets, they have carried out data preparation and extracted the raw data. Following the extraction of values from the datasets, cross-validation and training of the twitter data have been completed. To sample the data in accordance with psychology, the sampling procedure is applied to the dataset. Once the classifier is put into practice, the real data will be categorized and subjected to a variety of learning techniques. The data are divided into two categories in the classification technique: balanced data and imbalanced data. The data that are not missing any values might be oversampled in unbalanced data. The authors have used LSTM for their project to classify the tweets into positive or negative so finally classifying them into depressed or not.

In the paper [2], authors used head motions in order to detect the depression. Datasets they examined are two datasets : clinically validated data collected at the Black Dog Institute – a clinical research facility focusing on the diagnosis and treatment of mood disorders such as anxiety and depression (referred to as BlackDog dataset) – and the Audio/Visual Emotion Challenge (AVEC2013) depression dataset. They didn't rely on single feature to detect the depression, the employed different ML methods which include Logistic Regression, Random forest, Support Vector Machine, Extreme Gradient Boosting and Multi Layer Perceptron.

In the paper [3], authors followed unique way of collecting data from patients. They designed special PHQ-9 questions and asked the patients for their answers. All the answers are carefully recorded and used as dataset for their research. Later they have extracted the features and used TPTO to train the model. But their limitation is that they considered minimum number of patients for their dataset which does not generalize the outputs

The expert [4] discusses a study focusing on how depression impacts the relationship between features extracted from speech. It emphasizes that leveraging this understanding can improve the training speed and effectiveness of depression detection models using Support Vector Machines (SVMs) and Long Short-Term Memory networks (LSTMs). The research

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 3, March 2024 |

utilized the Androids Corpus, which includes 112 speakers, including 58 diagnosed with depression. Findings revealed that models trained with feature correlation matrices, as opposed to feature vectors, experienced enhanced training speed and performance, resulting in an error rate reduction of 23.1% to 26.6% across various models. This improvement is attributed to the increased variability of feature correlation matrices among individuals with depression, suggesting their potential use as markers for depression.

In paper [5] discusses the creation of a model aimed at detecting clinical depression using a user's social media posts. This model utilizes a specialized Depression Symptoms Detection (DSD) classifier, trained on clinician-annotated tweets focused on depression symptoms. It leverages features such as depression scores, temporal patterns, and user activity levels, including periods of "no activity" or "silence" on social media platforms. To assess the efficacy of these features, the study constructs three distinct datasets derived from established benchmark datasets for user-level depression detection. Accuracy evaluations are conducted, considering various features and their performance across different levels of temporal granularity. The findings underscore the effectiveness of semantic-oriented representation models while also highlighting the potential enhancement in overall detection performance through the incorporation of clinical features. This improvement is particularly notable when there is a congruence between the training and testing data distributions and an abundance of data available in a user's social media timeline.

III.METHODOLOGY OF PROPOSED SURVEY

The CNN method, also known as Convolution Neural Network or Conv Net, is the foundation of the suggested model. It is a deep learning algorithm that accepts an image as input, weights the input values, and then assists in classifying the output image. CNN is utilized for data analysis, pattern identification, computer vision, and NLP task solving in addition to picture classification.
CNN, also called Multilayer Perceptron, is a kind of Artificial Neural Network (ANN). The design of the neurons in the human brain served as the model for the CNN algorithm. CNN operates on the convolution operation principle.

1.Dataset:
A verified multimodal database of emotive speech and music is called RAVDESS. 24 professional actors with neutral North American accents who vocalize lexically matched statements make up the gender-balanced database. Expressions of calmness, happiness, sadness, anger, fear, surprise, and disgust are all present in speech, and similar emotions are present in songs. Every expression has two emotional intensity levels in addition to a neutral version. There are formats for face-and-voice, face-only, and voice-only circumstances. Each of the 7356 recordings in the set was scored ten times for emotional validity, genuineness, and intensity. 247 people who matched the description of untrained study volunteers from North America supplied ratings. A second group of seventy-two participants contributed test-retest data.

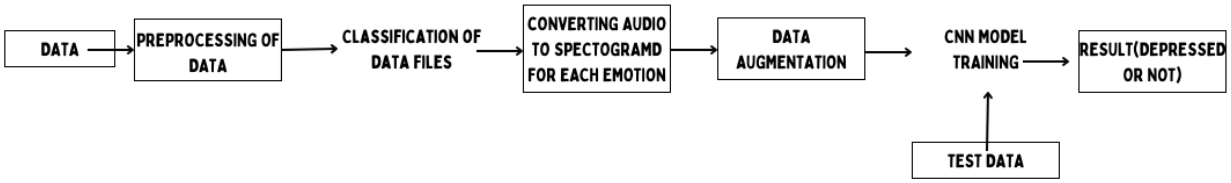
voice-only formats which are .wav files are used for detecting depression. Here is the filename identifiers as per the official RAVDESS website:
- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion
01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised.
- Emotional intensity (01 = normal, 02 = strong).
NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).
So, here's an example of an audio filename. 02-01-06-01-02-01-12.mp4
This means the meta data for the audio file is:
- Video-only (02)
- Speech (01)
- Fearful (06)
- Normal intensity (01)
- Statement "dogs" (02)
- 1st Repetition (01)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 3, March 2024 |

- 12th Actor (12) - Female (as the actor ID number is even)

2.Methodology:



Using criteria like depression and non-depression categories, first group files from the RAVDESS dataset into distinct folders.

Depression = {4:'Sad', 5: 'Angry', 6: 'Fearful', 7: 'Disgust'} Non-depression = {1: 'Neutral', 2: 'Calm', 3: 'Happy' , and 8: 'surprised'}.To the categorized data, we applied the following data augmentation process:
-The technique of introducing tiny perturbations to our initial training set to generate fresh synthetic data samples is known as data augmentation. We can use noise injection, time shifting, pitch and speed changes, and other techniques to generate syntactic data for audio. Our goal is to improve our model's generalization capacity and make it invariant to those perturbations. The original training sample's label must be preserved while adding fluctuations for this to function. In picture data the image can be augmented by rotating, zooming in, and other techniques.

Variability in audio data can arise from a variety of sources, including background noise, speaker changes, and environmental factors, as well as recording settings. We replicate these real-world fluctuations in the training data by using data augmentation techniques such pitch variation, temporal stretching, shifting, and noise addition. Audio signals may have distortions, background noise, or other differences in real-world situations. We train the model to become more robust and resilient to noise and other environmental influences by adding such fluctuations to the training data. As a result, the model is better able to generalize to new data since it can identify underlying patterns and traits in the audio signals.

It is performed in 4 stages:

1.Adding random noise to the audio signal
2.Stretching the audio signal in time
3.Shifting the audio signal in time
4.Changing the pitch of the audio signal

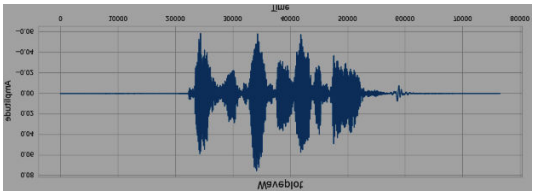


Fig 1. Original audio

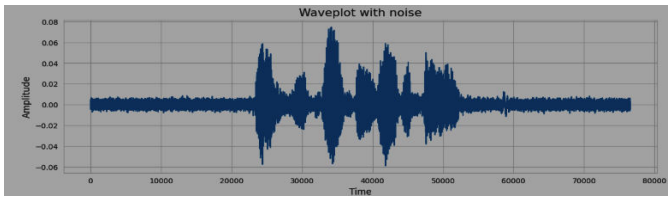


Fig 2. After adding noise

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 3, March 2024 |

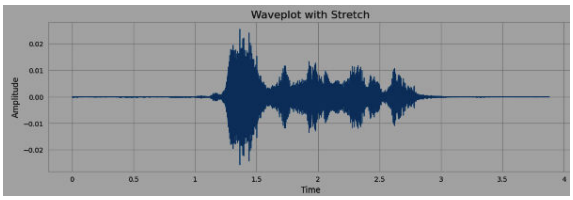


Fig 3. After Stretching

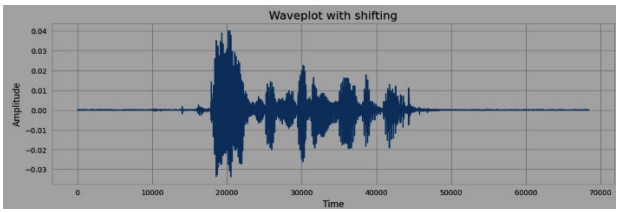


Fig 4. After Shifting

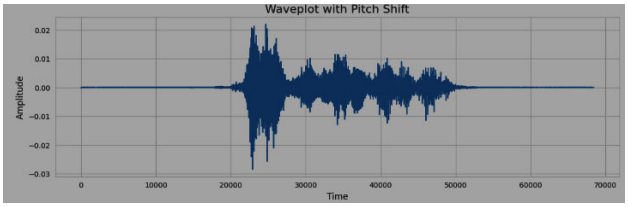
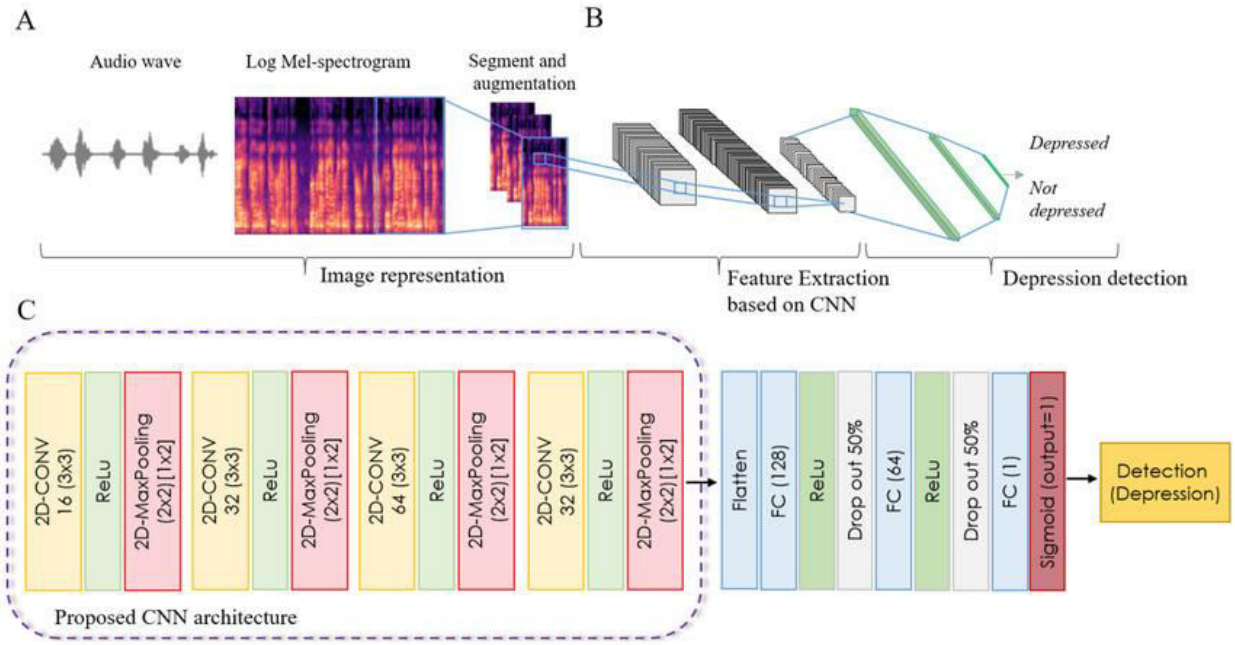


Fig 5. After path Shift

2a. Feature Extraction:



A crucial step in studying and determining relationships between various entities is feature extraction. Since the models cannot directly interpret the audio data that is presented, we must transform it into a format that the models can understand. To do this, we used feature extraction. Three axes-time, amplitude, and frequency—represent the three dimensions of the audio stream . The following five features are being extracted:

-Zero Crossing Rate

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 3, March 2024 |

-Chroma_stft
-MFCC
-RMS(root mean square) value
-Mel Spectrogram to train our model.

i. Zero Crossing Rate (ZCR):
The zero crossing rate is the rate at which the signal changes its sign. It measures the number of times the signal crosses the horizontal axis (zero amplitude) within a given time frame. ZCR indicates the frequency of the signal and can be useful for differentiating between voiced and unvoiced segments of speech or identifying percussive sounds.

ii. Chroma STFT:
Chroma features represent the energy distribution of pitch classes (e.g., musical notes) within an audio signal. Chroma features are useful for capturing the harmonic content and tonal characteristics of music and speech signals. They are often used in tasks such as music genre classification and chord recognition.

iii. Mel-Frequency Cepstral Coefficients (MFCC):
MFCCs are a widely used feature representation for audio signals, particularly in speech and audio processing tasks. They capture the spectral characteristics of the signal by representing the short-term power spectrum of the audio signal. MFCCs are computed by taking the Discrete Cosine Transform (DCT) of the log Mel-spectrogram of the signal. They are known to be effective for speech recognition, speaker identification, and emotion recognition tasks.

iv. Root Mean Square (RMS) value:
RMS represents the root mean square amplitude of the signal, which provides a measure of the signal's energy. It is calculated by taking the square root of the mean of the squared values of the signal samples. RMS is useful for measuring the overall loudness or intensity of the signal.

v. Mel Spectrogram:
The Mel spectrogram is a representation of the short-term power spectrum of the audio signal in the Mel-frequency domain. It is computed by applying a Mel filter bank to the STFT of the signal and then taking the logarithm of the resulting energies.
Mel spectrograms capture the spectral characteristics of the signal in a perceptually relevant frequency scale and are widely used in speech and audio processing tasks. Later we split the data into train and test data. The data is given to the CNN model.

Architecture of CNN model:

1.**Input Layer**: This layer receives the input data, which consists of audio features extracted from the audio samples. In the context of the code, it's not explicitly defined as a layer but rather implied by the input data fed into the network.

2.**Convolutional Layers** (`Conv1D`): These layers apply a set of filters to the input data, capturing local patterns and features. Each filter is convolved with the input data, producing a feature map. The ReLU (Rectified Linear Unit) activation function is commonly used after each convolutional operation to introduce non-linearity into the network, allowing it to learn complex patterns and representations effectively.

3.**Max Pooling Layers** (`MaxPooling1D`): These layers down sample the feature maps produced by the convolutional layers, reducing the spatial dimensions while retaining the most important information. Max pooling is achieved by selecting the maximum value within each pool size, providing translation invariance and reducing computational complexity.

4.**Flatten Layer** (`Flatten`): This layer converts the multi-dimensional feature maps into a one-dimensional vector, suitable for input to the fully connected layers. It essentially "flattens" the output of the preceding convolutional and pooling layers into a single vector.

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 3, March 2024 |

5.**Fully Connected Layers** (`Dense`): These layers perform classification based on the learned features from the convolutional layers. Each neuron in a fully connected layer is connected to every neuron in the previous layer, allowing the network to learn complex relationships between features. The ReLU activation function is commonly used in the hidden layers to introduce non-linearity, while the softmax activation function is typically used in the output layer for multi-class classification tasks. Soft max ensures that the output probabilities sum up to 1, making it suitable for probability estimation across multiple classes.

6.**Dropout Layer** (`Dropout`): Dropout is a regularization technique used to prevent overfitting in neural networks. It randomly drops a fraction of the units (neurons) from the network during training, effectively forcing the network to learn more robust features and reducing reliance on any specific set of neurons. In the provided code, dropout layers are applied after some of the fully connected layers to prevent overfitting. These layers and activation functions collectively form the CNN architecture, enabling the network to learn hierarchical representations of the input audio data and perform classification tasks effectively.

IV.RESULTS AND DISCUSSIONS

Audio files are not utilized directly. Therefore, the audio files must be transformed into spectrograms in order to extract features. These are the results obtained:

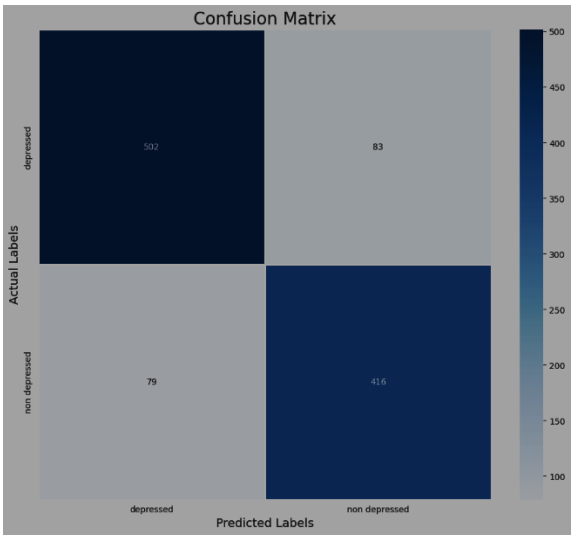


Fig. Confusion Matrix

	precision	recall	f1-score	support
depressed	0.86	0.86	0.86	585
non depressed	0.83	0.84	0.84	495
accuracy			0.85	1080
macro avg	0.85	0.85	0.85	1080
weighted avg	0.85	0.85	0.85	1080

Fig. Training Testing Accuracy

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 7.580| A Monthly Double-Blind Peer Reviewed Journal |

| Volume 11, Issue 3, March 2024 |

V CONCLUSION AND FUTURE WORK

The project resulted an accuracy of 84.6%. In conclusion, this research demonstrates the successful application of Convolutional Neural Networks (CNNs) for audio classification tasks, particularly in distinguishing between depression and non-depression states. By leveraging data augmentation techniques to increase the variability and size of the training dataset, and extracting a comprehensive set of features from the audio samples, the CNN model achieved commendable performance in accurately classifying audio samples. The results highlight the efficacy of deep learning techniques in audio classification, showcasing the importance of meticulous data preparation, feature extraction, and model architecture design. Moving forward, this research paves the way for further exploration and application of CNNs in diverse domains such as emotion recognition, speech analysis, and healthcare diagnostics, offering promising avenues for future research and practical applications. The project's drawback is that this paper only able to categorize eight different emotions; however, the project can expand on these to obtain results that are more accurate.

REFERENCES

[1] Sonam Gupta, ”Psychological analysis for Depression from Social Networking sites”, Hindawi, Volume 2022,Article ID: 439538 ,2022.

[2] Juan Aguilera, Delia Irazú Hernández Farías, Rosa María Ortega-Mendoza, and Manuel Montes-y Gómez. 2021” Depression and anorexia detection in social media as a one-class classification problem” Applied Intelligence 51 (2021), 6088–6103,2021.

[3] Sarmad Al-gawwam and Mohammed Benaissa, 2018, “Depression detection from eye blink features”, In 2018 IEEE international symposium on signal processing and information technology (ISSPIT). IEEE, 388–392.

[4] Sharifa Mohammed Alghowinem, Tom Gedeon, Roland Goecke, Jeffrey Cohn, and Gordon Parker, 2020, “Interpretation of depression detection models via feature selection methods”, IEEE transactions on affective computing (2020).

[5] Yashna Nainani, Kaveh Khoshnood, Ashley Feng, Muhammad Siddique, Clara Broekaert, Allie Wong, Koustuv Saha, Roy Ka-Wei Lee, Zachary M Schwitzky, Lam Yin Cheung, et al,2022,”Categorizing Memes about Abortion”, (2022).

[6] Ermal Toto, ML Tlachac, Francis Lee Stevens, and Elke A Rundensteiner,2020, ”Audio-based depression screening using sliding window sub-clip pooling”, In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 791–796.

[7] Tran Hien Van, Abhay Goyal, Muhammad Siddique, Lam Yin Cheung, Nimay Parekh, Jonathan Y Huang, Keri McCrickerd, Edson C Tandoc Jr, Gerard Chung, and Navin Kumar, 2023,”How is Fatherhood Framed Online in Singapore?’, arXiv preprint arXiv:2307.04053 (2023).

[8] Raymond chiong, Gregorius Satia Budhi 2021, “A textual-based featuring approach for depression detection using machine learning Classifiers and Social media texts”, 2023.

[9] Falzana Arefin Nazira, Sharna Rani Das, Sadah Anjum Shanto, M.F.Mridha 2021, “Depression detection using Convolutional Neural Networks”,2022 , Volume:2022.

[10] Yung Teck Kiong 2022, “An Initial Study of Depression Detection on Mandarin Textual through BERT Model”,2020.

[11] Wheidima Carneiro de Melo, Eric Granger, Abdenour Hadid 2019, DEPRESSION DETECTION BASED ON DEEP DISTRIBUTION LEARNING.

[12] Aloshban, N., Esposito, A., Vinciarelli,” A.: What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech. Cognitive Computation pp.”, 1–14 (2021).

[13] Alsarrani, R., Esposito, A., Vinciarelli, A.:” Thin slices of depression: Improving depression detection performance through data segmentation”, In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6257–6261 (2022).

[14] Wu, P., Wang, R., Lin, H., Zhang, F., Tu, J., Sun, M.: “Automatic depression recognition by intelligent speech signal processing: A systematic survey. CAAI Transactions on Intelligence Technology (2022)”.

[15] Lim, B., Arık, S.O., Loeff, N., Pfister, T.: Temporal fusion transformers for ¨ interpretable multi-horizon time series forecasting. International Journal of Forecasting 37(4), 1748–1764 (2021).

[16] Farruque, N., Goebel, R., Sivapalan, S., Zaiane, O.: “Depression symptoms modelling from social media text: A semi-supervised learning approach”. arXiv preprint arXiv:2209.02765 (2022).

Impact Factor: 7.580

INTERNATIONAL JOURNAL
OF MULTIDISCIPLINARY RESEARCH
IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

+91 99405 72462 +91 63819 07438 ijmrsetm@gmail.com