MENTAL HEALTH RECOGNITION USING SPEECH PROCESSING

D.Jayakumar[1], N.Sathish Kumar[2], Devalla Charan Sri Sai[3], Galla Vishnu Sai Saketh[4], Dasararaju Gnanendra[5]

[1]Assistant Professor of CSE, R.M.D Engineering College, Kavaraipettai, Chennai, Tamil Nadu, India.

[2]Assistant Professor of AIML, R.M.D Engineering College, Kavaraipettai, Chennai, Tamil Nadu, India.

[3,4,5]Student (B.E), Department of CSE, R.M.D Engineering College, Kavaraipettai, Chennai, Tamil Nadu, India.

Abstract- From a long time ago, human emotion identification from the voice signal has been a research issue in applications involving human-machine interfaces. Emotions are crucial to the mental health of humans. It serves as a vehicle for communicating one's viewpoint or mental condition to others. The process of determining the speaker's emotional state from the speech signal is called speech emotion recognition (SER). Any artificial intelligence system with a small computational power can be taught to recognize few universal emotions such as Neutral, Anger, Happiness, Sadness, etc. The chromogram, the Mel scaled spectrogram, mel-frequency spectrum coefficients (MFCC) are the sources of characteristics that we are retrieving in this work. The emotion or mental health in this study is categorised using a deep neural network. Accuracy is boosted by incorporating the Resnet algorithm.

Keywords*: News, efficiently, performance, interest...*

Introduction:

There are many numbers of ways to communicate, however, speaking is one of the quickest and most natural forms of human communication. Speech can therefore be a quick and effective way for humans and machines to communicate. Humans naturally possess the capacity to completely understand the message they have just heard, by using all their senses. People can detect the emotional condition of their communication partners using all their accessible senses. Although emotional recognition comes naturally to humans, it is an extremely challenging assignment for machines. The goal of a system that recognizes emotions is to use emotional data to enhance human-machine interaction.

Overall, the quality of the feature extraction and efficiency of machine learning algorithms in classifying emotions were crucial to the success of voice emotion detection in this system. The goal of ongoing research in this area is to advance feature extraction methods, and developing more robust algorithms that can handle the variability in emotional expression across different speakers and cultures.

In the areas of image and speech recognition, Deep Neural Networks (DNN) have achieved amazing success. We discovered that the DNN has a significant advantage when processing speech emotion. Because of this, the librosa package in Python was suggested in this paper to implement the emotional elements that were taken automatically from the audio. To recover voice emotion features, a 5-layer deep network was trained using DNN. The softmax classifier layer is used to recognize emotional speech, and a high latitude characteristic is produced by combining speech emotion data from several consecutive frames. The accuracy of the speech emotion detection test was 73.38%, which is regarded as high when compared to other speech recognition models.

The system used conventional deep learning methods such as CNN and Resnet to classify emotions.

A speech emotion identification engine that contains a classifier to detect emotions from the speech signal and a signal processing unit that extracts pertinent information from the voice signal to make up the system's two primary components and a classifier that takes the voice signal and extracts the emotions. The average classification accuracy for most classifiers is lower for speaker-independent systems compared to that of speaker-dependent systems.

Automated emotion detection from the human speech is becoming more prevalent today because it improves interactions between humans and machines.

1. Related works:

Deep Neural Networks for Object Detection: Recent results on picture classification challenges have shown how well Deep Neural Networks perform (DNNs). This paper takes object detection one step further by utilising DNNs to precisely localise as well as categorizing them, as items of various kinds. We offer an easy-to-use and successful method for object recognition as a regression issue to object bounding box masks. We provide a multi-scale inference approach that may be used by a few network applications to swiftly and cheaply detect objects at high resolution. The method performs to the highest standard on Pascal VOC.

Summary: This journal discusses about the Deep Neural Networks theory and object detection using DNN.

A Study on Automatic Speech Recognition: Speech is a simple and practical method of human interaction, but in modern times, human interaction extends to our interactions with the various machines we use in our daily lives. The computer is the most crucial. So, both computers and people can communicate using this method. Interfaces are used for this interaction, which takes place relating to human-computer interaction (HCI). the key terms of Automatic Speech Recognition (ASR), a well-known area of AI that should be considered throughout any pertinent study, are summarized in this paper (Speech pattern, vocabulary quantity, etc.). Along with a brief outline of our suggestion, which may be viewed as a contribution to this area of study, we also present a synopsis of major recent research on speech processing. The conclusion refers to a few potential improvements for future work.

Summary: This article helps us in understanding and using the speech recognition by machines which improves Human Computer Interactions and is also useful in our project.

Speech Emotion Recognition: Voice emotion identification has long been a hot study subject due to its potential applications in human-machine interfaces. There are various techniques available for extracting emotions from speech signals, and this paper reviews some of the previous technologies that use classifiers to recognize emotions. These classifiers are intended to differentiate between various emotions, such as surprise, happiness, sadness, and anger.

The system samples. Variables such as energy, pitch, and word choice are used to identify the emotions in the speech signals. Several classifiers are used to analyse these variables and can recognize the speaker's emotional state., the linear prediction cepstrum coefficient (LPCC), Voice emotion identification has long been a hot study subject, as it has numerous applications in human-machine interfaces. Several techniques have been developed to extract emotions from speech signals, and this paper reviews some of the earlier technologies that employ classifiers for emotion recognition. To differentiate between various emotions including surprise, happiness, sadness, and anger, these classifiers are utilized.

A library of emotional speech samples is the foundation of the technology used to identify speech emotions. From these speech samples, characteristics including energy, pitch, and MFCC are retrieved. Several classifiers are employed to categorize the speaker's emotional state based on these features.

Summary: This paper highlights the significance and necessity of utilizing diverse features such as MFCC, Mel, and other relevant attributes in audio and speech analysis for predicting emotions accurately. These features are critical components of our application, which are designed to classify and identify our emotions based on audio signals.

Deep Belief Network and SVM: Deep Neural Networks (DNN) and Deep Belief Networks (DBNs) were used to construct this study. By combining several consecutive frames, a five-layer DBN was trained to extract speech emotion aspects, producing high-dimensional features in order to address the key issue of feature extraction for speech emotion identification. The gleaned characteristics were subsequently supplied into a multiple classifier system for voice emotion recognition using a original method. These findings emphasize the potency of utilizing DBNs in DNNs for increasing speech emotion detection and the efficiency of using DBNs in DNNs for speech emotion feature extraction.

Summary: This paper highlights the significance of. Through this study, we gained insight into the features and methods utilized in DNN models, which can be implemented in our own application to improve its accuracy and performance.

2. Methodology:

Proposed system:

Our proposed approach (DNN) model. To capture the relevant information from audio files, we utilize various features Database at Ryerson University is the dataset used here (RAVDESS). We have only selected the speech portion, which has 1440 audio files and 24 evenly distributed actors. The model divides the audio of the speech into 8 categories, including neutral, peaceful, joyful, sad, furious, afraid, disgusted, and astonished.

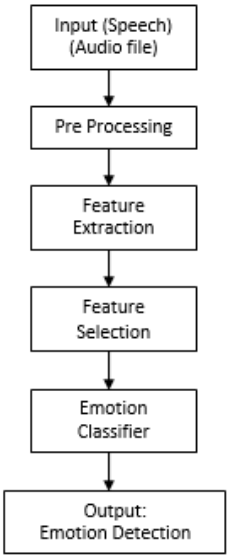


Figure 1: The block diagram of proposed method

3. Implementation:

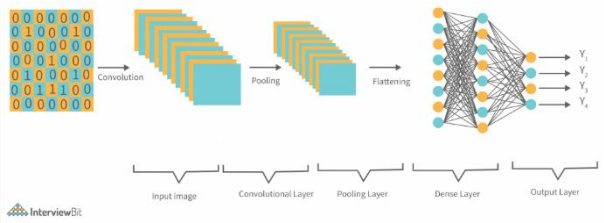
Using the algorithm mentioned below, the project was carried out.

CNN:

Convolutional neural networks (CNN, or ConvNet), to extract visual patterns from pixel pictures, multi-layer neural networks are used. CNN uses the term "convolution" to describe the mathematical operation. You can combine two functions in a specific A linear operation is utilized to exhibit how the structure of one function can be transformed by another function. In simpler terms, when two images are presented as matrices, they can be multiplied to produce a new output. the output in order to extract information from an image. While) cannot operate without convolutional layers. These layers, which form the basis of CNN architecture, enable it to effectively extract traits from incoming data.

The adoption of CNN artificial neural networks in a variety of industries has considerably improved computer vision applications because of its capacity to cater to the preferences of consumers. A backpropagation technique is used by data. It has several layers, including fully linked, pooling, and convolution layers. You can learn more about these words in the section that comes next.

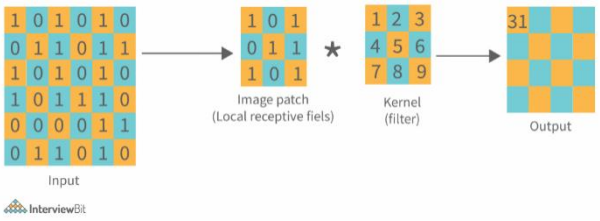
Typical CNN Architecture



The ConvNet's job is to keep elements necessary for obtaining an accurate forecast while condensing the images into a more digestible format. This is essential for developing an architecture that can scale to very large datasets and learn features.

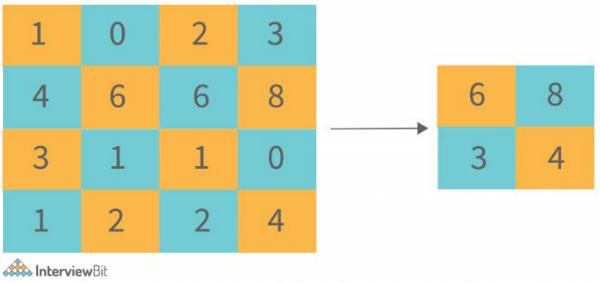
Convolutional neural networks, or ConvNets for short, are made up of three layers. Let's look more closely:

Convolutional operations are carried out by the convolutional layer (CONV), the basis of CNN. The part of this layer responsible for carrying out the convolution process (matrix) is the Kernel/Filter. The kernel adjusts the horizontal and vertical axes in accordance with the stride rate until the entire picture has been scanned. **Although it is smaller than the kernel, an image has more surface area. As a result, the kernel will span all three while having a short height and width. (RGB) channels in a picture.**



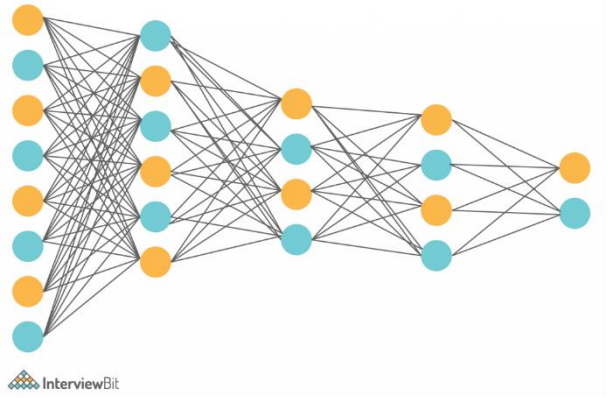
The operation of neural networks depends critically are used to process the output produced by linear processes like convolution. Historically, smooth non-linear the sigmoid and hyperbolic tangent (tanh) functions have been used, as they resemble activity of real neurons mathematically. However, currently, The (0, x). Deep neural networks have proven to perform better when using the ReLU function, which has made it a popular option for many computer vision applications.

Pooling Layer (POOL): This layer has responsibility for reducing dimensionality. As a result, less computer processing power is essential for processing the data. There are two categories of pooling: maximal and the average. The result of max pooling is the value that comes from the region of the picture that the kernel has covered the most. Average pooling results in the average of all the values in the area of the picture that kernel covers.



Fully Connected Layer (FC): In the fully connected layer (FC), which functions with a flattened input, every input is coupled to every neuron. Mathematical functional operations are then often carried out after the flattened vector has been routed through a few further FC levels. This is the moment where the classifying process begins. In CNN designs when FC layers are included, they are

often located at the end.



Along with the above layers, there are some additional terms that are part of a CNN architecture.

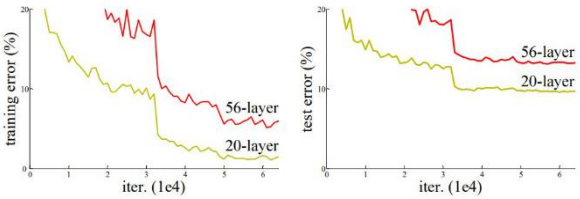
Activation Function: The activation function transforms the output real values of the last fully connected layer into a target class probability where each of the value ranges from 0 to 1 and the sum of all the values equals to 1. Using a softmax function, particularly useful for tasks that require identifying multiple objects or classes.

Dropout Layers: In order to prevent some neurons from contributing to the subsequent layer, the Dropout layer acts as a mask leaving all other neurons fully functional. A Dropout layer utilized to nullify some of the attributes of an input vector. As they stop the training data from being too well fitted, dropout layers are essential in CNN training. The initial set of training data has an excessively big influence on learning if they are absent. The following qualities would not be learned since they only appear in subsequent samples or batches:

Now that you have a firm grasp on CNN's fundamental components, let's look at some of the channel's well-known architectural designs.

Resnet:

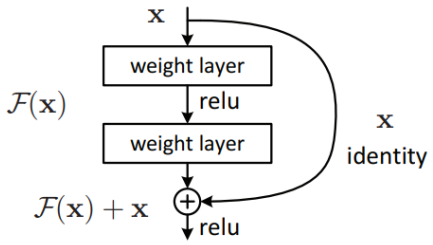
The early AlexNet, an architecture developed by CNN, won the ImageNet 2012 contest. In order deep learning difficulty. The result is the gradient is either zero or too large, leading to an increase in training and testing error.



Comparison of 20-layer vs 56-layer architecture

Residual Network: This design developed the Residual Blocks concept to deal with. The skip connection bypasses several intermediary levels in order to link layer activations to subsequent layers. As a result, a block is left over. to foster resentment.

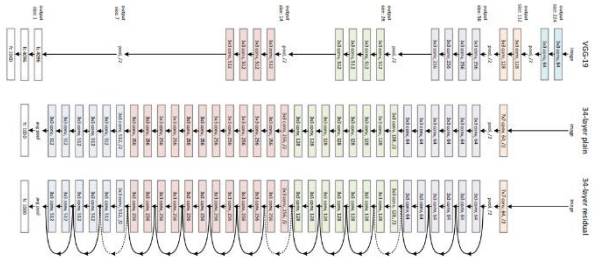
Rather than letting layers acquire the underlying mapping, this network lets the network suit the residual mapping. Hence, rather than applying, for example, the preliminary mapping of F(x), H(x): = H(x) - x, results in H(x):=F(x)+x, let network fit.



Skip (Shortcut) connection

This sort of skip link has the benefit of permitting regularisation to bypass any layer that degrades architectural performance. As an outcome, it is possible to train a very without having complications with growing or decreasing slopes. The 100-1000 layers of the CIFAR-10 dataset were the subject of the experiments conducted by the paper's authors.

Skip connections are widely employed in similar techniques referred as “highway networks” These are just better than ResNet architecture design that is inspired by the VGG-19. These shortcuts are then used to convert the design.



ResNet -34 architecture

Implementation:

Using the Tensorflow and Keras APIs, we were able to build the ResNet architecture (with Residual Blocks) from the ground up. Below are a few examples of real-world ResNet architecture. For this application, we make use of the CIFAR-10 dataset. There are 10 distinct categories in this collection of 60 000 3232 color photos, including vehicles, trucks, frogs, horses, deer, birds, cats, and cats. The keras.datasets API method can be used to assess this dataset.

1. Layer of Input: This is the layer where we feed the model data. Our input has the same number of characteristics as the number of neurons in the input layer.

2. **Layer of Hidden:** he characteristics of the input are sent to the underlying layer(s), where various processes/activities occur. There might be several hidden layers. The layers are subjected to mathematical procedures such as multiplication of the matrix, convolutions, pooling, and so on, as well as a function of activation.

3**. Layer of Output:** This level is used to create chance scores using sigmoid or softmax functions, that are subsequently transformed into our model outputs.

Padding:

- There are numerous ways to manage the edge pixels:
- Missing the edge pixels; padding with pixels with no value
- Padding for reflection

Reflection padding entails copying pixels taken from the picture's edge and pasting them outside the image, is the best technique. to make up of a 3x3 kernel, and three additional pixels must be reflected in a 7x7 kernel's outer shell.

The edge pixels are typically just ignored in research papers, It causes a minor loss of information and worsens when more deep layers of convolution are added. As an outcome of this, I've was unable to easily illustrate.

Strides:

When utilizing begin at (1, 1), travel to (1, 3), and finally to (1, 5), so forth. The kernel generates a single output that is the sum of each stride.

Image Input Layer:

This layer can be used, which also conducts data normalization. The input format option should be utilised to provide the image format. A picture's height, width, and number of colour channels all contribute to its overall size. For instance, a grayscale image has one channel, whereas a colour image has three.

Convolution Layer:

An activation is created by merely adding a filter to a source of data during a convolution. Applying Similar filter on a component repeatedly leads to an activation map known as a map of features are used to display the positions and strengths of a recognized feature within an input, like a photograph.

Pooling Layer:

It is common to introduce a Pooling layer intermittently between subsequent Convolution layers.
Its objective is to gradually reduce the spatial size of the representation, which is separately handled through the Pooling Layer, and then applies. In this case, maximum number of digits needed for each MAX operation is 4. (Little 2x2 an area in a depth slice). The value of the depth parameter remains unchanged.

Input Sequence Layer:

A sequence input layer sends data to a network. In most circumstances, the sequences of inputs and outputs are of various lengths (as in machine translation), hence the entire input sequence must be supplied before the intended result is anticipated. When someone says "sequence to sequence models" with no more clarification, it typically refers to a more complicated configuration. This is how it goes:

Because input and output sequences are typically of varying lengths (such as those used in machine translation), the entire input sequence must be supplied before the target can be anticipated. The term "sequence to sequence models",

- A RNN layer, or stack of them, functions as a encoder.
- Another RNN layer (or stack thereof) acts as "decoder": Given the target sequence's prior characters, it has been taught to anticipate the target sequence's subsequent characters.

LSTM Layer:

Recurrent neural networks struggle with short-term memory. They are going to struggle to transfer If a sequence is too lengthy, information from previous time steps is transferred to subsequent ones. RNN's may omit crucial information from the start if you're attempting to predict something from a piece of text.

Backpropagation vanishing gradient occurs in recurrent neural networks. Gradients are used to modify a neural network's weights. The gradient decreases as time passes, and this phenomenon is referred to as the "vanishing gradient problem".

Because of recent advances in data science, long short-term memory networks, or LSTMs, have been proven to be the best solution for virtually each of these prediction of sequences challenges.

LSTMs outperform RNNs and classic feed-forward neural networks in many areas. This is explained by their capacity to selectively remember patterns for lengthy periods of time.

Fully Connected Layer:

Each neuron on one layer connects with every neuron in every other layer via fully linked layers. The flattened matrix is sent through a fully linked layer to categorize the photographs.

Convolutional layers can only filter images, so in order to see the result, a dense layer is required. As the classification, the final set of convolutional neurons/filters output a number (or numbers if only one hot encoding is employed). Convolutional layers work to detect features that help classification by picking up edges and curves and then from there detecting shapes and from there picking out ears, for example. The last convolutional layer's neurons and filters outputs are merged, and the resultant 1D data is "flattened," or reduced to a single row from the original rows and columns. The 1D data is then sent into the fully connected layer's neuron(s), which execute a dot product of the input data and the neuron's weights to create a single number as output (one number per neuron).

Output Layers:

SoftMax and Classification Layers:

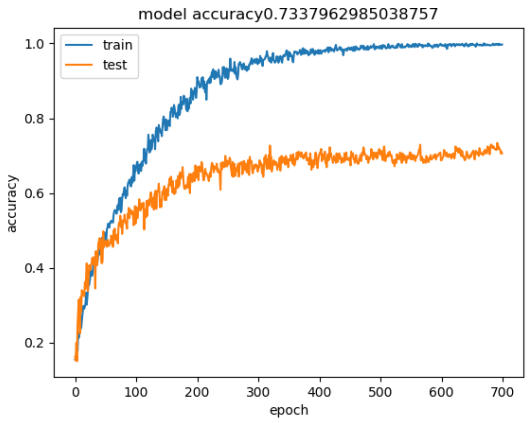
Use the classification layer to build a classification layer. For categorization issues, a layer of softmax must come after the last fully linked layer, then come a layer. of classification. The logistic sigmoid function, also known as the normalized exponential, is a multi-class generalization that is known as the softmax function. In traditional classification networks, the classification layer has to appear following the softmax layer. The categorization layer's railway network uses the Softmax function's values.

4. Results and Discussion:

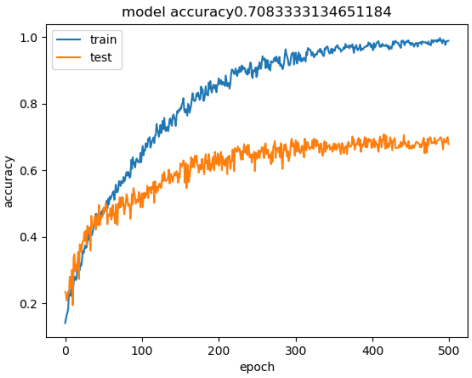
- The suggested technique offered a method for detecting emotion in human speech.
- The neural networks were used to achieve this strategy.
- We have successfully created a model for deep learning using the deep neural network architecture to forecast the emotions of the individual who speaks in audio.
- We have famed our project in a web-based application using the Flask architecture. The UI also includes user registration system.
- By applying the trained model, we managed to achieve a test accuracy of 73.4%.

Train vs test accuracies of our model over 700 epochs are shown below:

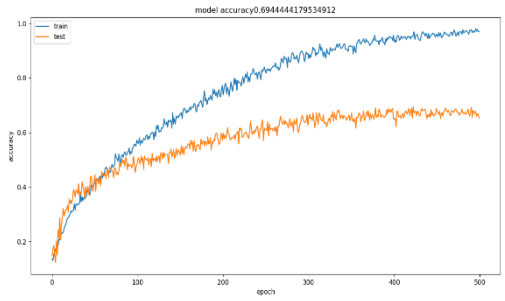
The model gets the best test accuracy of 73.4%



We previously tried other models too, with different accuracies.



This is another model which we developed using DNN.



5. Conclusion:

The suggested plan offered a method for identifying emotions in human speech. The neural networks have been used to implement this strategy. To accomplish this, accurately predict the addressee's emotions in an audio file, we have successfully created a model for deep learning with the deep neural network architecture. Our project is a well-known online application built with the Flask architecture. The user interface also has a user registration system. By applying the trained model, we managed to achieve a test accuracy of 73.4%.

Please be aware that emotion prediction is arbitrary, and that different people may grade the same music with different feelings. This is also the cause of the algorithm's occasionally inconsistent output when trained on human-rated emotions. Since the model was only trained using data from the RAVDESS dataset, the speaker's accent may also cause unexpected results.

6. References:

1. Szegedy, Christian & Toshev, Alexander & Erhan, Dumitru. (2013). Deep Neural Networks for Object Detection. 1-9.
2. Benk, Sal & Elmir, Youssef & Dennai, Abdeslem. (2019). A Study on Automatic Speech Recognition. 10. 77-85. 10.6025/jitr/2019/10/3/77-85.
3. Ashish B. Ingale & D. S. Chaudhari (2012). Speech Emotion Recognition. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2 Issue-1, March 2012
4. Chenchen Huang, Wei Gong, Wenlong Fu, Dongyu Feng, "A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM", *Mathematical Problems in Engineering*, vol. 2014, Article ID 749604, 7 pages, 2014. https://doi.org/10.1155/2014/749604
5. 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.
6. M. E. Ayadi, M. S. Kamel, F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", Pattern Recognition 44, PP.572-587, 2011.
7. I. Chiriacescu, "Automatic Emotion Analysis Based on Speech", M.Sc. THESIS Delft University of Technology, 2009.
8. P.Shobha Rani, Vamsidhar Enireddy, S.Finney Daniel shadrach, R.Anitha ,SugumariVallinayag ,T.Maridurai, T.Sathishf, .Balakrishnan "Prediction of human diseases using optimized clustering techniques", https://doi.org/10.1016/j.matpr.2021.03.068 https://www.sciencedirect.com/science/journal/22147853/46Materials Today: Proceedings
9. T. Vogt, E. Andre and J. Wagner, "Automatic Recognition of Emotions from Speech: A review of the literature and recommendations for practical realization", LNCS 4868, PP.75-91, 2008.
10. S. Emerich, E. Lupu, A. Apatean, "Emotions Recognitions by Speech and Facial Expressions Analysis", 17th European Signal Processing Conference, 2009.
11. P.Shen, Z. Changjun, X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine", International Conference On Electronic And Mechanical Engineering And Information Technology, 2011.
12. Pacha Shoba Rani ,A Vasantharaj, , Sirajul Huque, KS Raghuram, R Ganeshvkumar, Sebahadin Nasir Shafi "Automated Brain Imaging Diagnosis and Classification Model using Rat Swarm Optimization with Deep Learning based Capsule Network" Publication date2021/7/12 International Journal of Image and Graphics Pages 2240001 Publisher World Scientific Publishing Company.