SPEECH EMOTION RECOGNITION

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering/Information Technology

By

Shreyansh Puri (191363)

Under the supervision of

Dr. Shweta Pandit and Dr. Rajni Mohana

to

Department of Computer Science & Engineering and InformationTechnology

Jaypee University of Information Technology
Waknaghat, Solan-173234, Himachal Pradesh

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report entitled **" Speech Emotion Recognition"** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic recordof my own work carried out over a period from January 2023 to May 2023 under the supervision of **Dr. Shweta Pandit** (Assistant Professor (SG), Electronics and Communication Engineering) and **Dr. Rajni Mohana** (Associate Professor, Computer Science and Engineering/ Information Technology) .

I also authenticate that I have carried out the above mentioned project work under the proficiency stream **Machine Learning**.

The matter embodied in the report has not been submitted for the award of any other degree ordiploma.

Shreyansh Puri, 191363

This is to certify that the above statement made by the candidate is true to the best of myknowledge.

Dr. Shweta Pandit	Dr. Rajni Mohana
Assistant Professor (SG)	Associate Professor
ECE	CSE
Dated:	Dated:

PLAGIARISM CERTIFICATE

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: .....

Type of Document (Tick): PhD Thesis | M.Tech Dissertation/ Report | B.Tech Project Report | Paper

Name: _____ **Department:** _____ **Enrolment No** _____

Contact No. _____ **E-mail.** _____

Name of the Supervisor: _____

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): _____

_____

_____

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

Complete Thesis/Report Pages Detail:

- Total No. of Pages =
- Total No. of Preliminary pages =
- Total No. of pages accommodate bibliography/references =

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at .....(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

(Signature of Guide/Supervisor) **Signature of HOD**

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

<table>
<tr><th>Copy Received on</th><th>Excluded</th><th>Similarity Index (%)</th><th colspan="2">Generated Plagiarism Report Details (Title, Abstract & Chapters)</th></tr>
<tr><td></td><td rowspan="4">• All Preliminary Pages
• Bibliography/Images/Quotes
• 14 Words String</td><td rowspan="3"></td><td>Word Counts</td><td></td></tr>
<tr><td>Report Generated on</td><td>Character Counts</td><td></td></tr>
<tr><td rowspan="2"></td><td>Submission ID</td><td>Total Pages Scanned</td><td></td></tr>
<tr><td></td><td>File Size</td><td></td></tr>
</table>

Checked by
Name & Signature **Librarian**

.....

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com

ACKNOWLEDGEMENT

Firstly, I express my heartiest thanks and gratefulness to almighty God for His divine blessingmakes it possible to complete the project work successfully.

I am really grateful and wish my profound indebtedness to Supervisors **Dr. Shweta Pandit,** Assistant Professor (SG), Department of Electronics and Communication Engineering and **Dr. Rajni Mohana**, Associate Professor, Department of Computer Science and Engineering Jaypee University of Information Technology,Wakhnaghat. The supervisor's extensive knowledge and deep interest in the "Research Area" are essential to the success of this project. This endeavour was made possible by their never-ending patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at allstages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Shweta Pandit,** Department of ECE and **Dr. Rajni Mohana**, Department of CSE, for their kind help to finish my project.

I would also generously welcome each one of those individuals who have helped me straight forwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

Shreyansh Puri(191363)

CONTENTS

LIST OF ABBREVIATIONS

ML - Machine Learning

ANN - Artificial Neural Network

CNN - Convolutional Neural Network

MFCC - Mel Frequency Cepstrum Coefficient

ReLU - Rectified Linear Unit

MLP - Multilayer Perceptron Classifier

SAVEE- Surrey Audio-Visual Expressed Emotion

RAVEDESS- Ryerson Audio-Visual Database of Emotional Speechand Song

TESS- Toronto Emotional Speech Set

CREMA-D- Crowd-sourced Emotional Multimodal Actors Dataset

MFCC- Mel-Frequency Cepstral Coefficients

ZCR- Zero Crossing Rate

RMSE- Root Mean Square Energy

LIST OF FIGURES

LIST OF GRAPHS

LIST OF TABLES

ABSTRACT

Speech Emotion Recognition is a very interesting yet very challenging task of human computer interaction. Speech Emotion Recognition is the process of trying to identify affective and emotional states in speech. This makes use of the fact that tone and pitch in the voice frequently convey underlying emotion. In order to grasp human emotion, animals like dogs and horses also use this phenomenon. In this project, we tried to recognize emotion in short voice message. We had used four datasets (SAVEE, RAVDESS, TESS, and CREMA-D) in this project which contains ~7 types of main emotions: *Happy, Fear, Angry, Disgust, Surprised, Sad or Neutral.* In previous studies, we had seen that everyone has worked on separate datasets but in our project, we had combined the four datasets (SAVEE, RAVDESS, TESS, CREMA-D) into a single file and then we send input wav file as our input to the model. Then we had performed feature extraction techniques (MFCC, ZCR, RMSE) for reducing noise from the data and then organised the sequential data obtained in the 3D array form that the CNN model accepts. Using the Matplotlib library, we put the data into a graphical form, then after some repeated testing with various values reveals that the model's average accuracy is 71% at testing and 96% at the training phase.

CHAPTER-1 INTRODUCTION

1.1 Introduction

The most elementary way of communication in humans is Speech. To enrich interaction, one needs to know and understand the emotion of another person and how to react to it. Unlike machines, we humans can naturally recognize the nature and emotion of the speech. Can a machine also detect the emotion from a speech? Well this could be made possible using machine learning. Machines need a specific model for detecting the emotions of a speech and such a model can be implemented using machine learning.

Speech emotion recognition is a very useful and important topic in today's world. A machine detecting the emotion of a human speech can be proved useful in various industries. A verybasic usage of speech recognition is in the health sector where it can be used in detecting depression, anxiety, stress etc. in a patient. It can also be used in industries like the crime sector where emotions can be recognized from the speech to distinguish between victims and criminals.

Emotions can be of various types like happy, sad, angry, disguised etc. depending on the feeling and frame of mind of the person. In our study, we have used various datasets with different emotions. We have also combined four datasets to one dataset and then applied themodel so that the efficiency of the model can be improved and there can be a variety in the data points. This has also resulted in eliminating the overfitting condition in our model.

The Deep Neural Networks (DNN) classifiers provided a solution to circumvent the problem of feature selection. Using an end-to-end network, which accepts

raw data as input and outputs class labels, is the notion. There is no need to compute manually created features or determine the optimal parameters in terms of classification. The network itself does everything. To effectively partition the data into the desired categories, the network parameters are optimized. However, this very practical solution comes at the expense of a much higher requirement for labeled data samples than traditional classification methods.

Machine Learning

Machine Learning is a well known procedure of foreseeing or Classifying information to assist with people in pursuing important choices. In order to learn from previous experiences and analyze the verifiable data, ML computations are prepared over cases or models. Just structure models aren't sufficient. The model should be adequately advanced and tuned so thatit gives you precise results. In order to achieve the best results, streamlining strategies requiretweaking the hyper parameters.

As it repeatedly trains on the models, it gains the ability to detect designs, enabling more precise decision-making. When the ML model is familiar with any new data, it applies its learnt lessons to the new data and creates predictions for the future. Using various normalizedmethodologies, one can advance their models in light of the most recent exactness. In a similar vein, AI models learn how to adapt to novel models and deliver better outcomes.

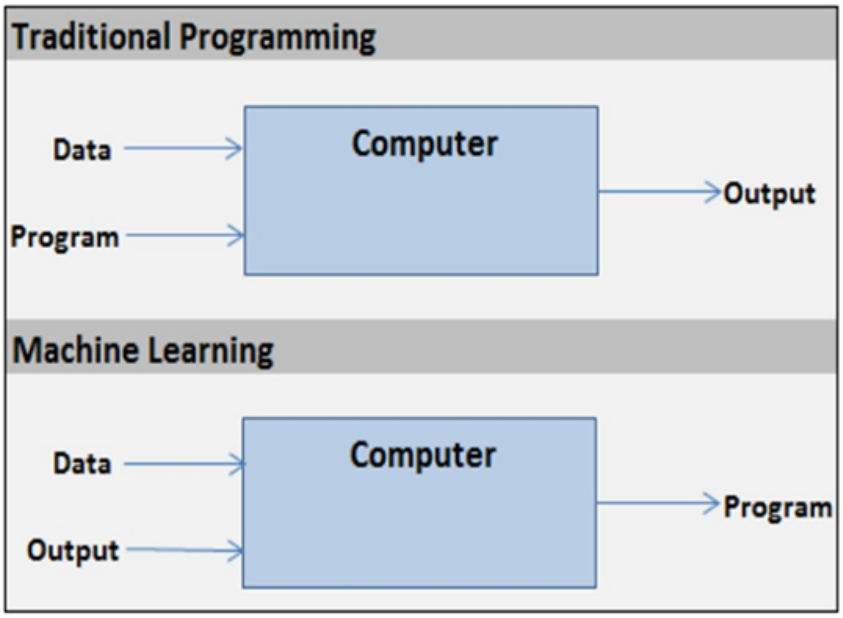


Fig 1.1: Traditional Programming vs Machine Learning [23]

Types of Learning:

Supervised Learning

Regulated learning is a type of AI in which machines are trained with carefully "marked" training data and on the basis of that data, the machines predict the outcome. The marked data suggests that some information is now labeled with the correct output. In regulated

learning, the preparation information given to the machines function as the boss that helps the machines to accurately foresee the result.

Managed learning is used most frequently in pragmatic machine learning. When you use a computation to get the planning capability from the input variable's contribution to the output, you are administering learning. The input variable is (x) and the result variable is (Y). Y = f(x) (x) The goal is to accurately prepare so that you can anticipate the outcome factors (Y) for the information when you receive fresh information (x).

Unsupervised Learning

Unsupervised Learning is an ML method, where you needn't bother with administering the model. All things considered, you really want to permit the model to chip away on its own to find data. It essentially manages the unlabelled information and searches for already undetected examples in an informational index with no prior marks and with at least humansupervision.In differentiation to administered discovering that generally utilizes human-named information, solo learning, otherwise called self-association, takes into account displaying of probability densities over inputs.

Calculations based on unsupervised learning generate designs from a dataset without makinguse of named or known results. It is the setup of a machine that uses data that is neither labelled nor organised and allows the calculation to proceed unguidedly on that input. In this case, the machine's task is to group unsorted data according to analogies, comparisons, and contrasts with essentially no prior knowledge preparation. In contrast to supervised learning, no educator is provided, implying that the machine won't receive any training. As a result, the machine is limited in its ability to find the hidden design in unlabeled data.

Reinforcement Learning

Reinforcement Learning is a machine learning technique that enables experts to learn in intelligent environment by experimenting with feedback from their own behaviors and experiences. Machine generally learns from prior experiences and makes an effort to provide the best solution for a given problem. It is the process of getting machine learning models ready to choose from a range of options. In contrast to regulated realization, where the input given to the expert is the proper arrangement of activities for carrying out an assignment, directed and support learning both use planning between information and results.

The most effective way at the moment for indicating a machine's creativity is reinforcementlearning. One use of AI is in support learning. It involves taking a sensible action to increasecompensation in a particular situation. It is used by various programming and computers todetermine how to behave or what action to take in a specific situation. Support gaining differsfrom directed learning in that there is no answer in support realizing, but the support expertdecides how to carry out the provided task. In managed learning, the preparation material includes the answer key, so the model is prepared with the correct response itself. It will surely benefit from a preparation dataset's absence.

1.2 Problem Statement

Feelings assume a fundamental part in correspondence, the location and examination of the equivalent is of imperative significance in the present computerized universe of distant correspondence. Feeling identification is a testing task, since feelings are emotional. We characterize a SER framework as an assortment of strategies that cycle and group discourse signs to identify feelings implanted in them. Such a system has a vast variety of application, such as intelligent voice-based assistants and expert guest conversation research. The goal of this work is to identify fundamental emotions in recorded conversation by breaking down the acoustic components of thesound data of reports. In this undertaking, we will foresee the feeling in the discourse of an individual's sound on the given dataset utilizing CNN and profound learning calculations. The dataset comprises 12,800 sound records of 12 male and 12 female voices with various feelings like blissful, outrage, miserable, shock, unbiased, dread, disdain. The significant objective of the proposed framework is figuring out Convolutional Brain Organization, and predicting Emotion in view of the model.

1.3 Objectives

Using voice data in practical applications like Automatic Speech Recognition (ASR)and Speech Emotion Recognition was the focus of this study (SER). In order to propose project ideas, we looked at open source Python packages and launched its ASR. The TESS dataset, SAVEE, RAVDESS, and CREMA-D were all utilised in the process of building a strong SER model. We will be able to start creating projects and learn SER fundamentals thanks to this practical experience.

Different audio processing methods are used by scientists to extract this buried information layer. It enables you to magnify and separate speech's tones and acoustic characteristics. It is not as simple to convert an audio signal to a numeric or vector representation as it is for a picture. How important information is preserved when the"audio" format is abandoned depends on the conversion mechanism. It will be challenging for the model to understand emotions and categorise samples if a certain data transformation misses smoothness and stillness.

The visualisation of audio signals based on their frequency content using Mel spectrograms is one way to translate audio data into numerical form. This may be modeled as an audio wave and sent into a CNN that is being trained to classify images. With the help of the Mel-Frequency Cepstral Coefficients, this may be recorded (MFCC). Based on the applications, each of these data formats offers benefits and drawbacks.

1.4 Methodology

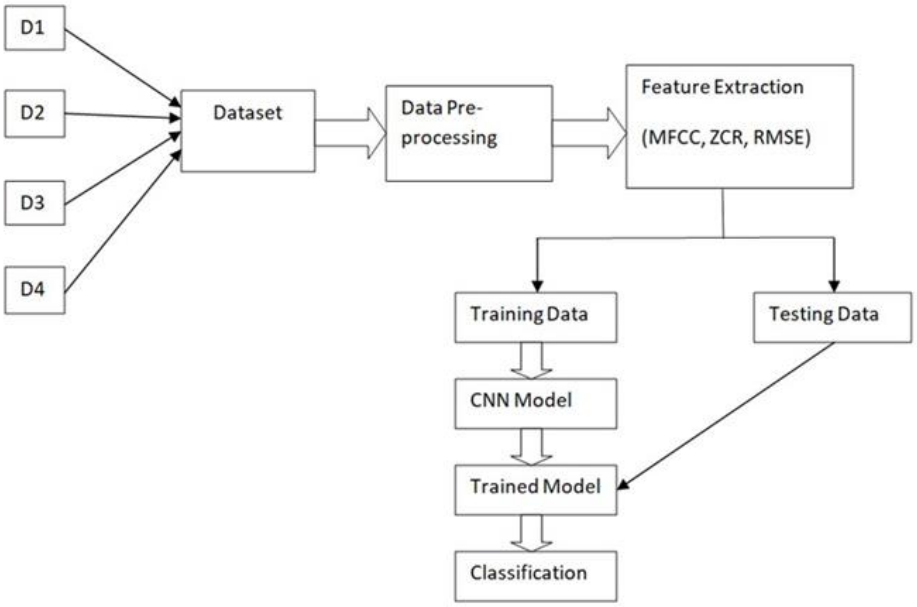


Fig 1.2: Flowchart of the model

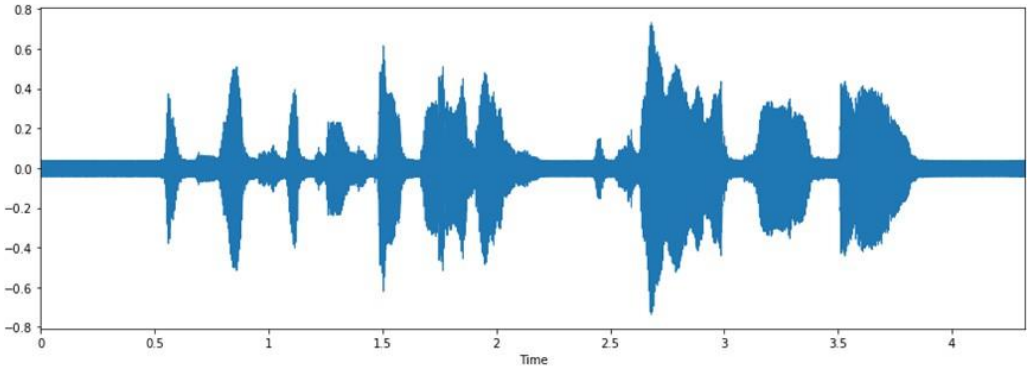
DATASETS

1] SAVEE

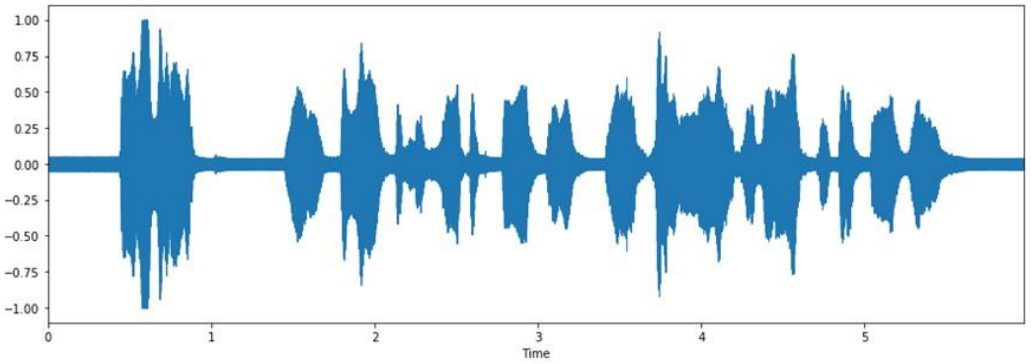
Surrey General media Communicated Feeling (SAVEE) information base has been recorded as a pre-imperative for the improvement of a programmed feeling acknowledgment framework. The data set comprises accounts from 4 male entertainers in 7 distinct feelings, 480 English expressions altogether. The sentences were looked over the standard TIMIT corpus and phonetically-adjusted for every inclination. The information was kept in a visual media lab with great general media gear, handled and marked. To check the nature of execution, the accounts were assessed by 10 subjects under sound, visual and general media conditions.

Characterization frameworks were assembled involving standard highlights andclassifiers for every one of the sound, visual and general media modalities, and speaker-free acknowledgment paces of 61%, 65% and 84% accomplished separately.

The first source includes four folders, each addressing a speaker. However, I combined all of them into a single organiser, so the first two letters of the filename are the initials of each speaker. For instance, "DC d03.wav" is the speaker DC's thirdexpression of nausea. The fact that they are male speakers is meaningless. This won'tbe a problem since the TESS dataset, which is almost all female, will balance thingsout.



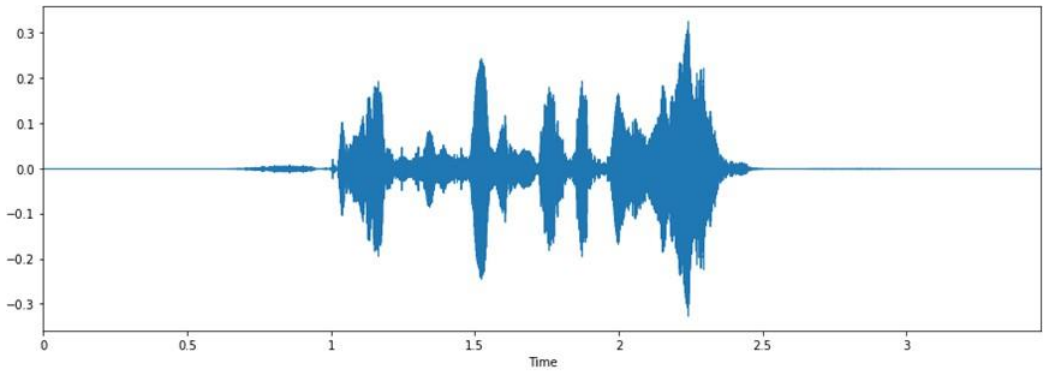
Graph 1: Wave plot of fearful audio of SAVEE dataset



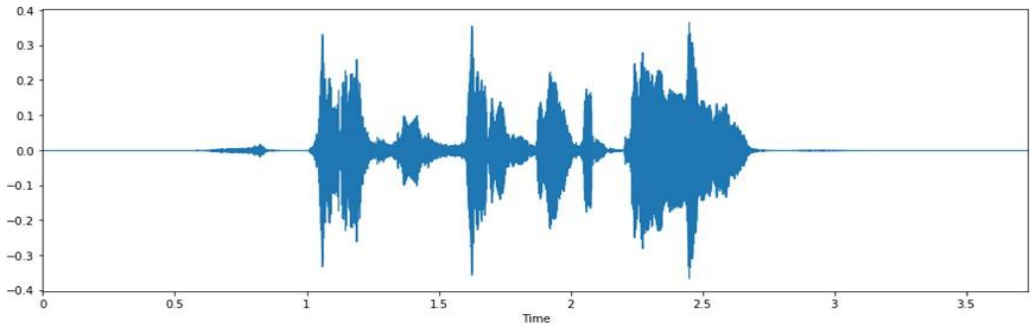
Graph 2: Wave plot of happy audio of SAVEE dataset

2] RAVDESS

The RAVDESS, Emotional Audio Database at Ryerson University (RAVDESS). There are 7356 files in the (RAVDESS) (total size: 24.8 GB). 24 professional actors:12 male and 12 female—perform 2 lexically similar phrases in the database with neutral American accents. Both in speech and in the lyrics, there are expressions of calmness, happiness, joy, sadness, anger, fear, surprise, and contempt. There are two emotional intensity levels (normal and strong) and one neutral expression created for each expression. Three modality forms are accessible for all conditions: Audio-only (16bit, 48kHz.wav), Audio-Video (720p H.264, AAC 48kHz,.mp4), and Video-only (480p H.264, AAC 48kHz,.mp4) (no sound). Note that Actor 18 doesn't have any song files.



Graph 3: Wave plot of fearful audio of RAVDESS dataset



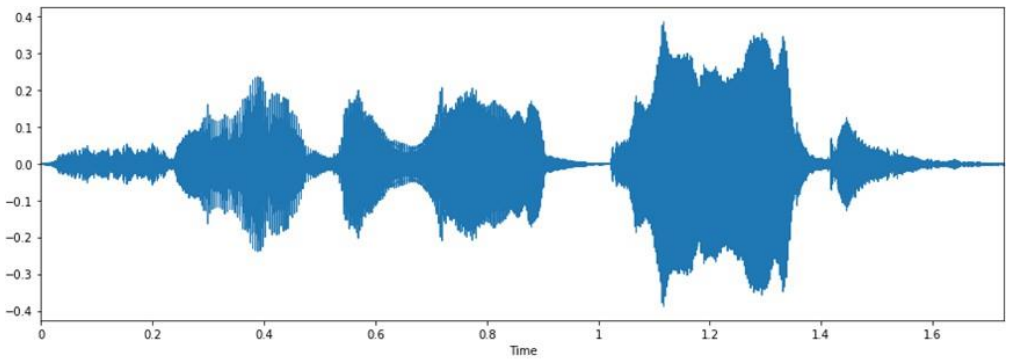
Graph 4: Wave plot of happy audio of RAVDESS dataset

3] TESS

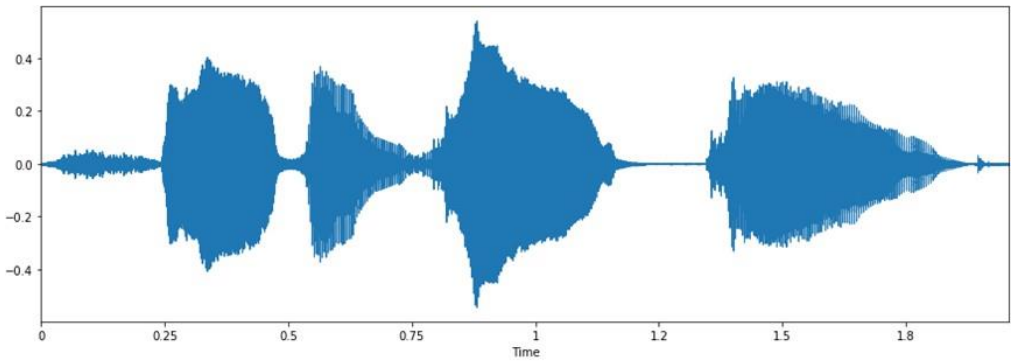
One of the four important datasets that I was fortunate to discover is the (TESS) dataset. It'sintriguing that this dataset solely includes females and yet the audio is of such good calibre.The other sample is primarily composed of male speakers, creating a slightly unbalanced representation. Consequently, in terms of generalisation, this dataset would serve as a very good training dataset for the emotion classifier (not overfitting).

Two actresses (26 and 64 years old) recited a set of 200 target words in the carrier phrase "Say the word _," and recordings of the set evoking each of the seven emotions were created.There are a total of 2800 audio files.

Each of 2 female performers and their emotions are included within an own folder in the dataset, which is organised thus way. Additionally, all 200 target word audio files are contained within that.



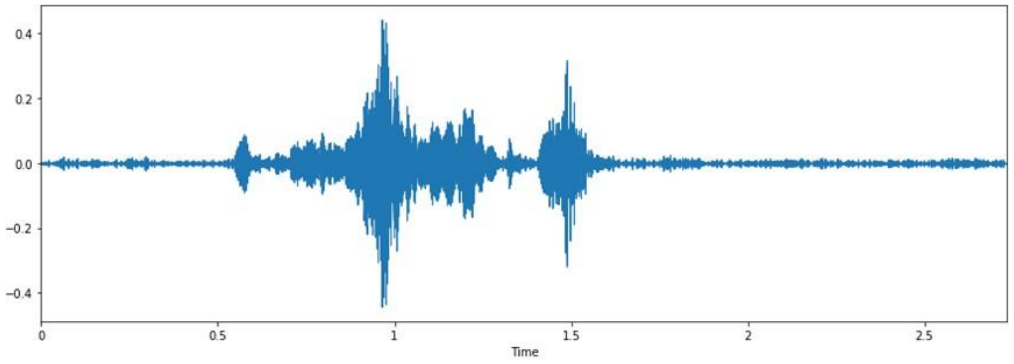
Graph 5: Wave plot of fearful audio of TESS dataset



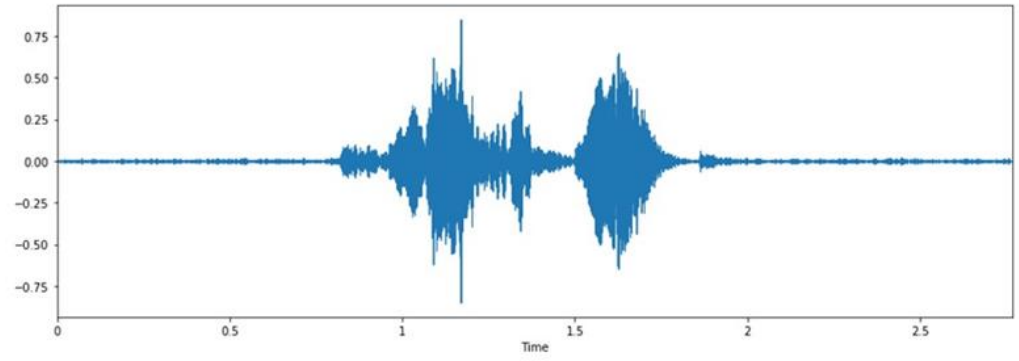
Graph 6: Wave plot of happy audio of TESS dataset

4] CREMA-D

One of the four important datasets that I was fortunate to discover is the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) dataset. The intriguing thing about this dataset is how diverse it is, which aids in the training of a model that can be applied to other datasets. There is a lot of information loss as a result of the restricted number of speakers used in many audio datasets. There are several speakers on CREMA-D. Because of this, using the CREMA-D dataset will help guarantee that the model does not overfit. 91 performers contributed 7,442 original footage to the data collection known as CREMA-D. The performers included in these video, who ranged in age from 20 to 74 and represented a diversity of races and ethnicities, included 48 men and 43 women. A selection of 12 phrases were read by the actors. Six distinct emotions (angry, disgust, fear, happy, neutral, and sad) and four different emotion degrees were used to deliver the phrases (Low, Medium, High, and Unspecified).



Graph 7: Wave plot of fearful audio of CREMA-D dataset



Graph 8: Wave plot of happy audio of CREMA-D dataset

Feature Extraction

Feature extraction assists with lessening how much excess information from the informational index. Extraction of highlights is a vital part in dissecting and tracking down relations between various things. The information given of sound can't be perceived by the models straightforwardly to change over them into a reasonable organisation highlight extraction is utilised.

The State of the Discourse signal figures out what sound emerges. If the shape is resolved precisely, then the right portrayal of the sound being created is acquired. The occupation of Mel Recurrence Cepstral Coefficients' (MFCC's), ZCR(), RMSE() is to address it accurately. MFCCs are utilised as information highlights. Stackingalso, changing over sound information into MFCCs design is done by python bundle librosa.

MFCC

MFCC is typically used as components in discourse recognition frameworks, such as those that can recognise numbers spoken into a phone. MFCCs are also increasingly keeping track of participation in applications for music data recovery, such as kind categorization, sound closeness measurements, and so forth.

RMSE

The Root-Mean-Square (RMS) Energy is quite similar to the AE-Amplitude Envelope. Rather than beginning identification, be that as it may, it endeavours to see tumult, which can be utilised for occasion recognition. Besides, it is substantially more powerful against exceptions, meaning on the off chance that we fragment sound, we can distinguish new occasions, (for example, another instrument, somebody talking, and so on) considerably more dependably.

As we window across our wave structure, we square the amplitudes inside the window and summarise them. When that is finished, we will separate by the edge length, take the square root, and that will be the RMS energy of that window.To extricate the RMS, we canjust utilise librosa.feature.rms.

Zero Crossing Rate

Zero Crossing Rate is essentially the time a waveform crosses the flat time pivot. This component has been essentially utilised in acknowledgment of percussive versus pitched sounds, monophonic pitch assessment, voice/unvoiced choice for discourse signals, and so on. The zero-crossing rate can be used as a fundamental pitch location calculation for monophonic apparent signs. Voice movement identification (VAD), which decides if human discourse is available in a sound portion, likewise utilises zero-crossing rates.

CNN Implementation

One of the most convincing developments in computer vision has been the use of convolutional neural networks. They have performed significantly better than typical PC vision and have produced cutting-edge outcomes. These neural networks have demonstrated real success across a wide range of actual contextual studies and applications.

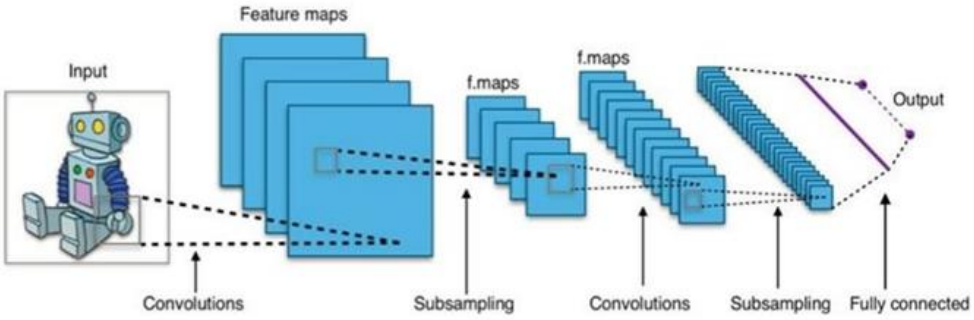


Fig 1.3: General Architecture of Convolutional Neural Network [5]

This figure shows that the picture as a contribution to the organisation, which goes throughvarious convolutions, subsampling a completely associated layer, yields something.

1) The convolution layer analyses the output of neurons which are connected to close locations or open fields in the information, processing each neuron's output of a single itemin between their loads and the small responsive field they are connected to in the information volume. Every computation triggers the information picture's element map to be extracted. As a result, picture yourself taking a 3x3 network and sliding it around an image that is addressed as a 5x5 network of values. You replicate the benefits of your 3x3 window at eachplace of that network by the aspects of the image that are now hidden by the window. You will then be given a single number that corresponds to each of the attributes in that windowof the images. This layer is used for sifting; when the window goes over the image, you lookfor patterns there. Channels, which are replicated by the values the convolution produces, are the reason this works.

2) Subsampling's goal is to obtain an information depiction by reducing its components, which aids in reducing overfitting. Max pooling is one of the subsampling techniques. Withthis method, you choose the highest pixel value from a location based on its size. Overall, maximal pooling detracts most from the portion of the image that is already covered by the bit. For instance, a

maximum pooling layer with a 2 x 2 size will select the 2 x 2 district withthe highest pixel power value. The pooling layer therefore behaves quite similarly to the convolution layer at that time, which is more accurate than incorrect. You can also move a window or a piece over a picture; the main difference is that the picture window or piece's capability isn't straight.

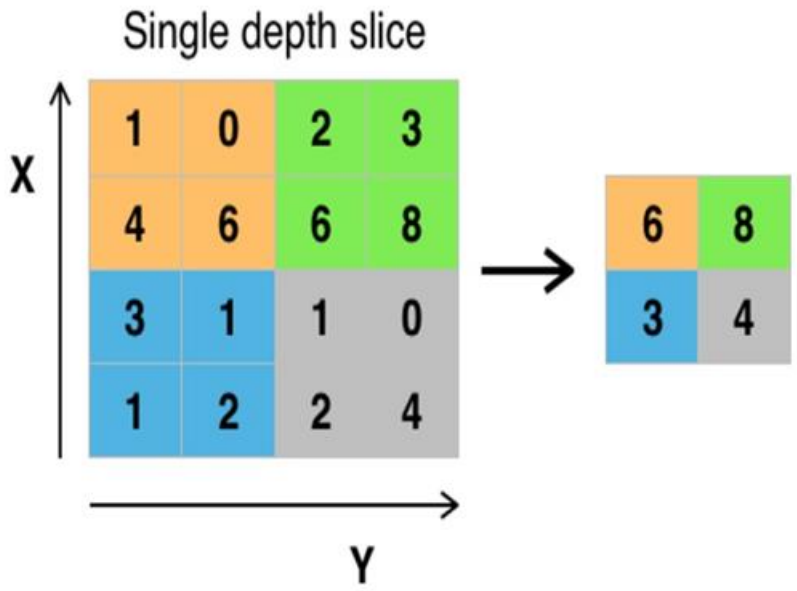


Fig 1.4: Max-Pooling [12]

3) Convolutional layers advance the undisputed level components, and the fully associated layer's job is to align them and combine all of the highlights. It sends the leveled output to the result layer, where you use a sigmoid or a softmax classifier to predict the name of the information class.

Activation Function

An enactment capability is a vital element of a counterfeit brain network, they essentially conclude regardless of whether the neuron ought to be actuated. They acquire non-straight properties with our organisation whose primary design is to change over an input sign of a hub in ANN to a result signal. That result signal presently can be utilised as a contribution to the next layer in the stack.

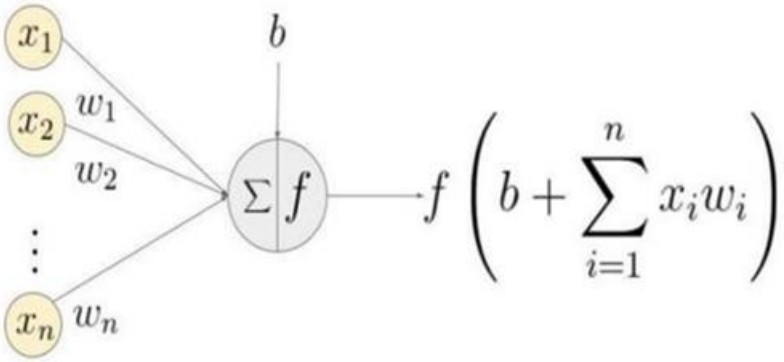
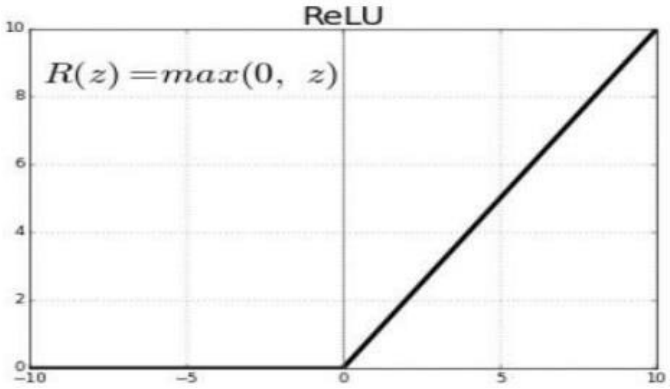


Fig 1.5: Working of an Activation Function [19]

Working

In the illustration, the input signal vector (x1, x2,..., xn) is multiplied by the weights (w1, w2,..., wn). Accumulation (i.e. summation with addition of bias comes next. Finally, thissum is subjected to an activation function, f.

ReLU (Rectified Linear Unit)



Graph 9: ReLU Activation Function

Equation:

f(x) = max (0, x)

Range: (0 to infinity)

Softmax Activation Function

A softmax capability is likewise a kind of sigmoid capability yet it is exceptionally helpful to deal with characterization issues having various classes.

$$\sigma(\overrightarrow{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

The softmax capability is displayed above, z is vector of contributions to the

yield layer (on the off chance that you have 10 result units, there are 10 components in z).

Softmax capability is undeniably utilised in the result layer of classifier where weare attempting to achieve the probabilities to characterise the class.

ReduceLROnPlateau

When a statistic stops improving, lower the learning rate.

Models typically profit from reducing the learning rate by a factor of 2–10 to improvelearning. This callback screens a certain amount, and if no progress is observed after a certain length of time (referred to as "persistence"), the learning rate is lowered.

Sequential model

A Sequential model is suitable for a plain pile of layers where each layer has precisely one info tensor and one result tensor.

Creating a Sequential model:-

```
model = keras.Sequential(
  [
    layers.Dense(2, activation="relu"),
    layers.Dense(3, activation="relu"),
    layers.Dense(4),
  ]
)
```

Python Libraries

Librosa- A Python library for analysing music and audio is called librosa.

NumPy- A Python package used for cluster operations is called numPy. It can also operate in the areas of straight polynomial maths, fourier transform, and grids.

Matplotlib- A cross-platform library for data visualisation and graphical charting is calledmatplotlib.

TensorFlow- TensorFlow is a Python-friendly open-source toolkit for mathematicalcomputation that accelerates and simplifies AI.

Keras- An open-source programming framework called keras provides fake brain organisations with a Python point of contact. Keras functions as the TensorFlow libraryconnecting point intended to provide fast experimentation .

1.5 Organisation

Chapter 1 summarises the basic introduction of our project topic, what is the problem statement, objectives behind the project and what are the different methodologies we had used in our project.

Chapter 2 describes the literature survey, where different research papers related to this project work have been included.
At most sixteen research papers are added in this report clearly based on speech emotion recognition.

Chapter 3 summarises the system development where various analytical and developmentalanalysis is explained along with the design and algorithms of the model development. Model development technique is explained further by adding various feature extractiontechniques. Their mathematical explanation is also shown using the formulas used.

Chapter 4 provides an account of the performance analysis where we have mentioned the most suitable model for this project after calculating the accuracy of the model. It also provides the outputs and step by step results at various stages.

Chapter 5 presents a brief summary of conclusions, and future work based on the research done during the implementation of the project.

In the end, References is added where all the research papers are mentioned that were neededfor the better implementation of the model.

CHAPTER-2 LITERATURE SURVEY

Md. Rayhan Ahmed et al.[1],used four deep neural network-based models built using LFABS. Model-A uses seven LFABs followed by FCN layers and a softmax layer for classification. Model-B uses LSTM and FCNs , Model-C uses GRU and FCNs and Model-D combines the three individual models by adjusting their weights. From each of these audio files, they hand-craft five categories of features- MFCC, LMS, ZCR, RMSE. did. data set. These features are used as inputs to a one-dimensional (1D) CNN architecture to further extract hidden local features in these speech files. To obtain additional contextual long-term representations of these learned local features via the 1D CNN block, we extended our experiment by incorporating LSTM and GRU after the CNN block , giving us more improved accuracy. After running DA, we observe that all four models perform very well on the SER task of detecting emotions from raw speech audio.Amongst all four models, the ensemble Model-D achieves the state-of-the-art weighted average accuracy of 99.46% for TESS dataset.

Dr. Nilesh Shelke et al.[2] used RAVDESS, TESS and SAVEE datasets for classification. Their purpose is to mandate the modernization of current plans and technology enabling EDS and to implement assistance in all areas of computers and technology. Analytics complement emotions extracted from databases, layers, and model libraries created for emotion recognition from speech. It mainly focuses on data collection, feature extraction, and automatic emotion detection results. The intermodal recognition computer system is considered a unimodal solution because it offers higher sorting accuracy. Accuracy depends on the number of emotions detected, the features extracted, the classification method, and the stability of the database.

A novel paradigm for emotion identification in the presence of noise and interference was put out by Shibani Hamsa et al. [3]. In order to examine the

speaker's emotions, our methodtakes into account the speaker's energy, time, and spectral factors. However, we suggest adopting the novel wavelet packet transform (WPT)-based cochlear filter bank rather of the gammatone filter bank and short-time Fourier transform (STFT) frequently employed in the literature. To do. When tested on three different speech corpora in two different languages,our system—which combines this representation with a random forest classifier—performsbetter than other existing algorithms and is less prone to stressful noise. All metrics(Accuracy, Precision, Recall, and F1 scores) in the RAVDESS and SUSAS datasets score above 80%.

A data imbalance processing approach based on the selective interpolation synthetic minority oversampling (SISMOTE) methodology is suggested by Zhen-Tao Liu et al. [4] toreduce the influence of sample imbalance on emotion identification outcomes. In order to minimise duplicate characteristics with inadequate emotional representation, a feature selection approach based on analysis of variance and gradient-enhanced decision trees (GBDT) is also provided. The results of speech emotion detection tests on the CASIA, Emo-DB, and SAVEE databases demonstrate that our technique produces an average of 90.28% (CASIA), 75.00% (SAVEE), and 85.82% (based on the findings) (Emo-DB). It demonstrates its precision in recognition. Utilizing voice emotion recognition is superior to some cutting-edge technologies.

To accomplish efficient speech emotion identification, Apeksha Aggarwal et al. [5] have presented two alternative feature extraction strategies. First, utilising superconvergence to extract two sets of latent features from voice data, bidirectional feature extraction is presented. Principal Component Analysis (PCA) is used to produce the first set of features for the first set of features. A second method involves extracting the Mel spectrogram picturefrom the audio file and feeding the 2D image into his pre-trained VGG-16 model. In this study, several algorithms are used in comprehensive experimentation and rigorous comparative analysis of feature extraction approaches across two

datasets (RAVDESS ANDTESS). The RAVDESS dataset offered significantly more accuracy.

Aseef Iqbal [6] used a real-time emotion detection system to analyse the tonal features of live-recorded speech and identify emotions. 34 audio characteristics, including MFCC, energy, spectral entropy, and others, are retrieved. To categorise emotions, this system essentially employs a model trained via gradient boosting. SVM and KNN, two of his other classifiers, are also used to assess their efficacy on test audio samples. The system is trained using two of his datasets, namely the RAVDESS and SAVEE databases. The method considers four emotions: neutrality, neutrality, pleasure, and neutrality.SVM and ANN exhibit 100% accuracy for both Anger and Neutral for RAVDESS (male). However, gradient boosting surpasses his SVM and ANN in both happiness and melancholy. SVM obtains her 100% accuracy in rage in RAVDESS (female), exactly like in the male version. Except for the sadness, SVM functions nicely overall. At 87% and 100%, respectively, KNN also performs well on the furious and neutral dimensions. Anger and neutral don't mix well with gradient enhancing. When compared to other classifiers, ANN scores horribly on the happiness and sadness metric. SVM and ANN do extremely well on neutral and rage on the combined male and female data sets, but not on gradient boosting. KNN's performance with happiness and melancholy is quite lacking.

A voice analysis-based emotion recognition system was proposed by Noushin Hajarolasvadi and Hasan Demirel [7]. In order to extract an 88-dimensional vector of audio characteristics, including Mel-frequency cepstrum coefficients (MFCC), pitch, and intensity for each frame, they first partition each audio signal into overlapping frames of identical duration. For every frame, a spectrogram is created concurrently. The speech signal is retrieved from each audio signal by applying k-means clustering to the extracted characteristics of all the frames. This is the last preprocessing step. Then, 3D tensor keyframes

are used to represent the relevant series of spectrograms. Instead of using the entire set of spectrograms corresponding to speech frames, they selected the k best frames to represent the entire speech signal. They then compared the proposed 3D-CNN results s.with the 2D-CNN results and demonstrated that the proposed method outperforms pre-trained 2D meshes.

Using cluster-based genetic algorithms, Sofia Kanwal and Sohail Asghar [8] devised a feature optimization strategy. This method is based on three databases: the Ryerson Audio- Visual Database of Speech and Song (RAVDESS), the Berlin Emotional Speech Database (EMO-DB), and the Surrey Audio-Visual Comparing Datasets of Expressed Emotions (SAVE). The findings demonstrate that the suggested strategy significantly enhanced speech sentiment categorization. The recognition rates of EMO-DB are as follows: 89% for normal speakers, 86% for men, 88% for women, 82% for general speakers, 4% for men.

Key sequence segment selection based on repeated dial-based functional network (RBFN) similarity measurements inside a cluster is a new approach to SER that Muhammad Sajjad et al. [9] presented. The STFT technique is used to turn the chosen sequence into a spectrogram, which is then input to a CNN model to extract distinctive and significant characteristics from the spoken spectrogram. In the suggested approach, we process important segments rather than entire utterances to lower the complexity of model and normalise CNN features prior to actual processing to facilitate perception of spatiotemporal information. To increase detection accuracy and shorten model processing time, the proposed approach is tested on several standard datasets like IEMOCAP, EMO-DB, and RAVDESS. The suggested SER model's efficacy and robustness were tested against existing SER techniques with accuracy values of 72.25%, 85.57%, and 77.02% for up to using IEMOCAP, EMO-DB, or RAVDESS data sets.

The preparation of incoming audio samples using filters to denoise speech samples was the main topic of Anusha Koduru et alstudy .'s [10]. Discrete Wavelet Transform, Energy etc. methods are used to extract features in the next stage. In the feature selection stage, redundant data is culled from the features and sentiment is derived from the features using a global feature algorithm. It makes use of a machine learning classification method. These feature extraction algorithms have been tested for emotions that are common to all people, such as anger, happiness, sadness, and neutrality. The findings show that the decision tree has an accuracy of 85%, the SVM has an accuracy of 70%, and the LDA has an accuracy of 65%. The decision tree distinguishes the sentiment from the audio samples more precisely than SVM and LDA, according to the findings of the analysis.

K.A.Darshan, DR.B.N.Veerappa [11] the purpose of this paper is to document the development of speech recognition systems using CNNs. Design a model that can recognize the emotion of an audio sample. Various parameters are changed to improve the accuracy ofthe model. This paper also aims to find the factors that affect model accuracy and the key factors needed to improve model efficiency. The whitepaper concludes with a discussion ofvarious CNN architectures and parameter accuracies needed to improve accuracy, as well as potential areas for improvement.

S.R. Harini Murugan [12] there is a disconnect between acoustic characteristics and human emotions, and because it mainly relies on discriminating acoustic features extracted for the detection job presented, automated speech emotion recognition is a difficult procedure. People express their emotions in a variety of ways that are significantly distinct from one another. Looking at various things emphasises pitch shifts, which are caused by the varying intensities of spoken emotions. Therefore, recognising speech emotions is a difficult problem for visual computation. As a result, vocal emotion detection is based on (CNN) algorithm, which employs several

modules and uses classifiers to discriminate between emotions including happy, neutrality, and sorrow. The LIBROSA software is used to extract features from the voice samples that make up the data set for the Speech Emotion Recognition System. Performance in classification is based on retrieved characteristics. The audio signal's emotional content can then be identified.

K Ashok Kumar et al. [13] proposed a CAS (Computer Aided Services) 's challenging module that recognized emotions from audio signals. SER (Speech Emotion Recognition) uses several schemes, including different classification and speech analysis methods, to extract emotion from signals. This manuscript presents an overview of the method and examines the contemporary literature using existing models of language-based emotion recognition. This literature review presents the contribution of speech to emotion recognitionand extracts features used to determine emotion.

According to Linqin Cai et al. [14], it is challenging for the existing emotion detection system to satisfy the need of emotion detection in one mode given the rapid expansion of social media. We suggested a multimodal emotion recognition model for voice and text in this research. We merged CNN and LSTM in the form of binary channels to learn acoustic-emotional properties by taking into account complementarity between various modes. In themeanwhile, we effectively captured text characteristics using a Bi-LSTM network (bidirectional long-short-term memory). In order to learn and identify the fusion characteristics of, we also used a deep neural network. The results of both voice sentiment analysis and textual sentiment analysis were used to establish the final emotional state. The concept put out in the IEMOCAP database was then validated using a multimodal fusion experiment. The overall recognition accuracy for text increased by 6.70% when compared to a single modal, and the speech emotion recognition accuracy increased by

13.85%. The experimental findings demonstrate that multimodal detection accuracy in the test dataset is superior to single-modal detection accuracy and superior to other published multimodal models.

Thaweesak Yingthawornsuk [15] proposed MFCC, which are taken from spoken speech signals, are a method for speech recognition. Prior to training and testing speech samples using maximum likelihood classifiers (ML) and support vector machines (SVM), principal component analysis is used as a supplement to the feature dimensionality reduction state. based on a database of 40 spoken utterances from an experimental study that was recorded in acoustically controlled spaces, SVMs containing more randomised MFCC samples were selected from the database, 16 ordered MFCC extract clearly showed an improvement in recognition rate. were compared with ML.

A SER system based on several classifiers and feature extraction techniques has been created, according to Leila Kerkeni et al. [16]. Speech signals are processed to obtain modulation spectrum (MS) characteristics and mel-frequency cepstral coefficients (MFCC),which are then used to train different classifiers. The most pertinent feature subsets were found using feature selection (FS). For the sentiment categorization challenge, a number of machine learning paradigms were employed. Their effectiveness is then contrasted with thatof multivariate linear regression (MLR) and support vector machine (SVM) methods, which are often employed in the field of speech signal emotion identification. As test data sets, the Berlin and Spanish databases are employed. This study demonstrates that when speaker normalisation (SN) and feature selection are used, all classifiers in the Berlin database achieve 83% accuracy.

Ruhul Amin Khalil et al. [20], highlights some current literature using these

techniques for speech-based emotion recognition and gives an outline of deep learning techniques. It has offered a thorough analysis of the deep learning methods for SER. Recent years have seen a lot of study on deep learning methods like DBM, RNN, DBN, CNN, and AE. Based on the classification of several natural emotions including happy, joy, sadness, neutral, surprise, boredom, disgust, fear, and rage, these deep learning algorithms and their layer-wise designs are briefly explained. These approaches provide both simple model training and the effectiveness of shared weights.

Raghav Mittal et al. [21], analyses speech in order to draw out its emotions. The feature vector is made up of audio signal elements that represent unique characteristics of the speaker, such as tone, pitch, and energy. This information is crucial for setting up the classifier model to accurately identify a given emotion. To build the dataset for the research effort, MFCC were extracted from the speech signals. The classifier model is then given the feature vector that was extracted from the training dataset. The extraction method will be applied to the test dataset before the classifier makes a decision regarding the hidden emotion in the test audio. Four separate datasets—namely, SAVEE, RAVDESS, TESS, and CREMA-D—have been used in the dataset preparation process. In this work, the CNN model, LSTM, random forests, and support vector machines have all been employed to categorise emotions. The 2D CNN model fared better than all the other models, with accuracy close to 70% on the test dataset.

In order to compare the results, Harshit Dolka et al. [22] concentrate on training an ANN Model for SER utilising Mel Frequency Cepstral Coefficients (MFCCs) feature extraction. Although the number of emotions varies in different datasets, the model can categorise audio files based on a total of eight emotional states: happy, sad, angry, surprised, disgust, calm, and neutral. On the TESS data set, the proposed model provides an average accuracy of 99.52 percent; on the RAVDESS data set, 88.72 percent; on the CREMA data set, 71.69%; and on the SAVEE data set, 86.80 percent.

Minji Seo et al. [23] in their study have found SER approaches have demonstrated performance reduction when trained and tested using various datasets. Using many corpora and languages, cross-corpus SER research pinpoints speech emotion. To enhance generalisation, recent cross-corpus SER research has been carried out. We pretrained the log-mel spectrograms of the source dataset using our created visual attention convolutional neural network (VACNN), which has a 2D CNN base model with channel- and spatial-wise visual attention modules, in order to enhance the cross-corpus SER performance. In order to help the refined model during training, we extracted the feature vector from the target dataset using a bag of visual words (BOVW). Visual words represent local information in the image, so by creating a frequency histogram of visual words, the BOVW aids VACNN in learning both global and local features in the log-mel spectrogram. RAVDESS, EmoDB, and the SAVEE, the suggested method demonstrates an overall accuracy of 83.33%, 86.92%, and 75.00%, respectively.

Taiba Majid Wani et al. [25] have altered the Deep Stride Convolutional Neural Networks (DSCNN), a recently discovered alternate network design of convolutional neural networks, by using less convolutional layers to speed up computation while preserving accuracy. Additionally, we used the spectrograms produced by the SAVEE speech emotion dataset to train the cutting-edge CNN model and the proposed DSCNN. Four emotions—angry, joyful, neutral, and sad—were taken into consideration during the evaluation procedure. The evaluation findings demonstrate that the proposed architecture, DSCNN, outperforms CNN with a prediction accuracy of 87.8% (compared to 79.4% for CNN).

S.Ramesh et al. [26] to establish a new database and identify their emotions, the SAVEE and TESS datasets were combined. Our key goal is to characterise

their emotions using this solid dataset. We have suggested a fresh machine learning algorithm for this use. The features from the voice signal datasets are first extracted using Mel-frequency cepstral coefficients. Finally, a mix of the naive Bayes machine learning method and the grey wolf optimizer was suggested for classification. According to the findings, our suggested categorization system performs better than current machine learning.

J.Ancilin et al. [27] have performed the extraction of the Mel frequency cepstral coefficient in this study is modified in two ways: the magnitude spectrum is used in place of the energy spectrum, and the discrete cosine transform is excluded. The logarithmic magnitude spectrum on a non-linear Mel scale frequency is known as the Mel frequency magnitude coefficient. With a multiclass support vector machine as the classifier, the Mel frequency magnitude coefficient and three conventional spectral features—the Mel frequency cepstral coefficient, log frequency power coefficient, and linear prediction cepstral coefficient—are tested on the Berlin, Ravdess, Savee, EMOVO, eNTERFACE, and Urdu databases.

Yeşím Ülgen Sönmez et al. [28] In their study, a brand-new, less computationally complex SER approach has been created. On the databases RAVDESS, EMO-DB, SAVEE, and EMOVO, this technique known as 1BTPDN is used. The raw audio data is first transformed using a one-dimensional discrete wavelet transform to produce the low-pass filter coefficients. Textural analysis techniques, a one-dimensional local binary pattern, and a one-dimensional local ternary pattern are used to extract the features from each filter. The most prominent 1024 features out of 7680 features are chosen using neighbourhood component analysis, and the other features are ignored. These 1024 features are chosen as the input for the classifier, a support vector machine based on a third-degree polynomial kernel. In the databases RAVDESS, EMO-DB, SAVEE, and EMOVO, the success rates of the 1BTPDN were 95.16%, 89.16%, 76.67%, and 74.31%, respectively. In comparison to numerous textural, acoustic, and deep learning

state-of-the-art SER approaches, the recognition rates are higher.

Misbah Farooq et al. [29] found the advantages of a deep convolutional neural network (DCNN) for SER. Modern speech emotive datasets are mined for features for this purpose using a pretrained network. The best and most discriminative features for SER are then chosen by using a correlation-based feature selection technique on the retrieved features. We employ support vector machines, random forests, the k-nearest neighbours approach, and neural network classifiers for the classification of emotions. The Berlin Dataset of Emotional Speech (Emo-DB), Surrey Audio Visual Expressed Emotion (SAVEE), Interactive Emotional Dyadic Motion Capture (IEMOCAP), and Ryerson Audio Visual Dataset of Emotional Speech and Song (RAVDESS) are four publicly accessible datasets that are used in the experiments for speaker-dependent and speaker-independent SER. For speaker-dependent SER trials, our suggested technique obtains an accuracy of 95.10% for Emo-DB, 82.10% for SAVEE, 83.80% for IEMOCAP, and 81.30% for RAVDESS. Moreover, compared to other handcrafted features-based SER approaches, our method produces the best results for speaker-independent SER.

Title	Year	Dataset	Models	Performance Metrics
Ensemble 1D	2021	TESS EMO-DB RAVDESS SAVEE CREMA-D	DNN CNN Ensemble	Precision Recall Accuracy
Novel approach	2022	TESS SAVEE RAVEDESS		(Survey)
Wavelet Packet	2020	RAVEDESS SUSAS ESD	Gradient Boosting Random Forest	Accuracy
Selective Interpolation	2020	SAVEE EMO-DB	SVM Decision Tree KNN LR Naive Bayes	Precision Recall Accuracy
Two way feature	2022	RAVEDESS TESS	Decision Tree Random forest MLP Proposed	Accuracy
A real time emotion	2019	RAVEDESS SAVEE	SVM KNN Gradient Boosting	Accuracy
3D CNN	2019	SAVEE RML	SVM Random Forest 2D CNN Proposed	Accuracy
SER using clustering	2021	RAVEDESS SAVEE	SVM	Accuracy
Clustering based Deep BiLSTM	2020	RAVEDESS EMO-DB	CNN RCNN proposed	Precision Recall F! Score
feature extraction algorithm	2020	RAVEDESS EMO-DB	SVM MLP DT KNN	Accuracy
SER using ML	2020	SAVEE	CNN	Accuracy

		EMO-DB		
SER using CNN	2020	RAVEDESS TESS	CNN	Accuracy
Using speech signal	2019	RML SUSAS	Random forest SVM	Precision Accuracy
Audio-textual emotion recognition	2019	TESS	CNN LSTM	Accuracy Recall
Speech Recognition using MFCC	2020	SAVEE EMO-DB	SVM	Accuracy
Automatic SER	2019	CREMA-D TESS	MLR SVM	Accuracy

Table 1: Literature survey

CHAPTER-3 SYSTEM DEVELOPMENT

System Design

First, we had combined the four datasets (SAVEE, RAVDESS, TESS, CREMA-D) into a single file and then we send input wav file as our input and then we had performed feature extraction techniques (MFCC, ZCR, RMSE) for reducing noise from the data and to increase the accuracy of the learned models by extracting data from the input file.

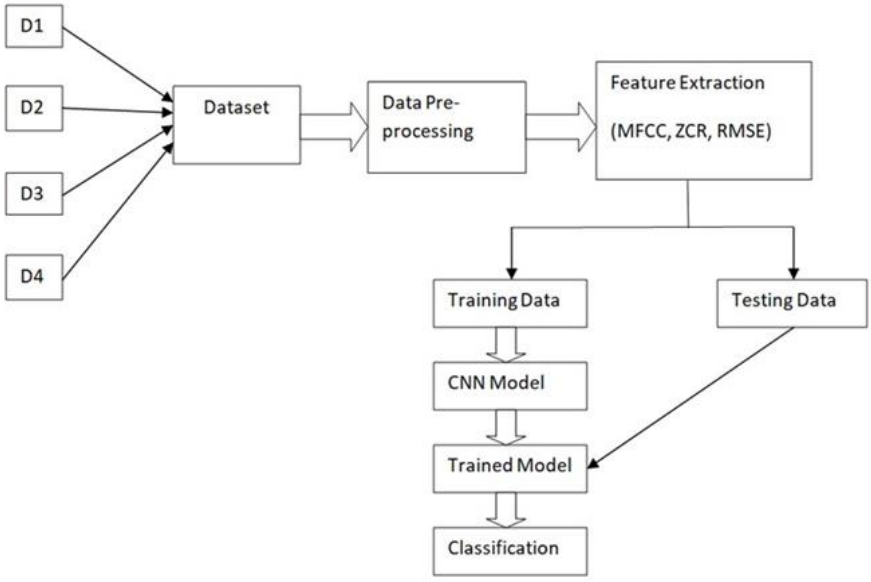


Fig 3.1: Flowchart of the model

This period of the general edge work diminishes the dimensionality of the information by eliminating the repetitive information. Obviously, build preparation and deduction speed. It additionally serves to decrease the quantity of highlights in a dataset by making new elements from existing ones. Next we are preparing the model with a Convolution Neural Network Algorithm with various feelings.

Feature Extraction Techniques:-

Working with datasets containing hundreds (or even thousands) of characteristics is getting more and more typical these days. A machine learning model may get overfitted if the number of features matches or exceeds (or even exceeds!) the number of observations contained in a dataset. Either regularisation or dimensionality reduction techniques (Feature Extraction) must be used to prevent this kind of issue. A dataset's dimensionality in machine learning is determined by the number of variables that were utilised to represent it.

Feature extraction of the audio dataset can be done on the basis of two domains. Thetwo domains for the feature extraction are time domain and frequency domain.

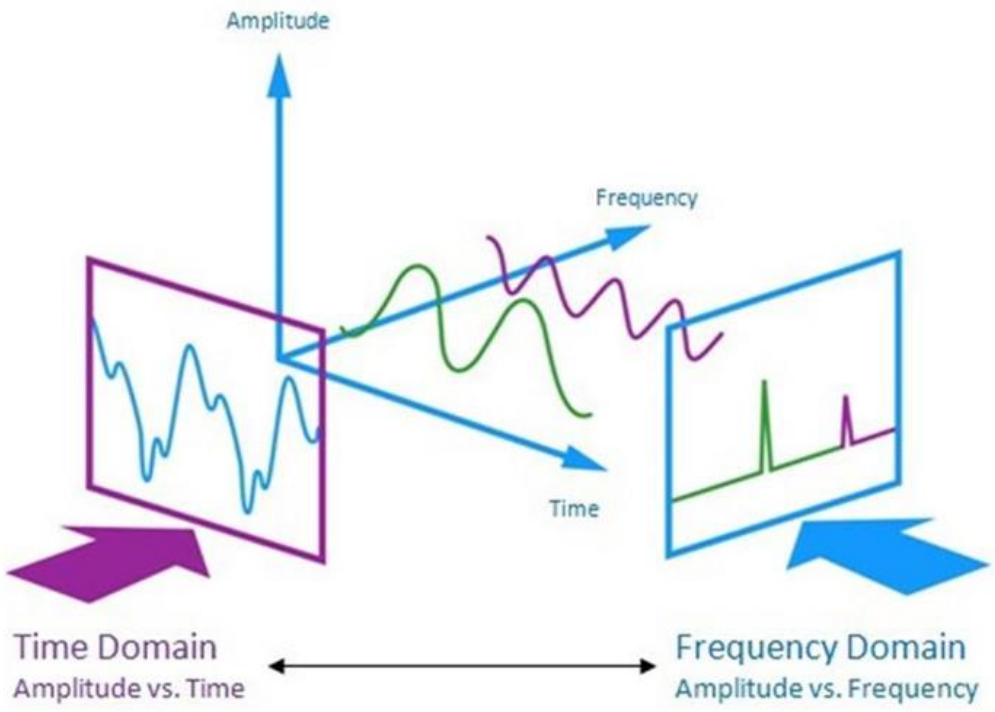


Fig 3.2: Feature Extraction Technique [18]

Mel-frequency cepstral coefficients (MFCC)

MFCC technique is given below:

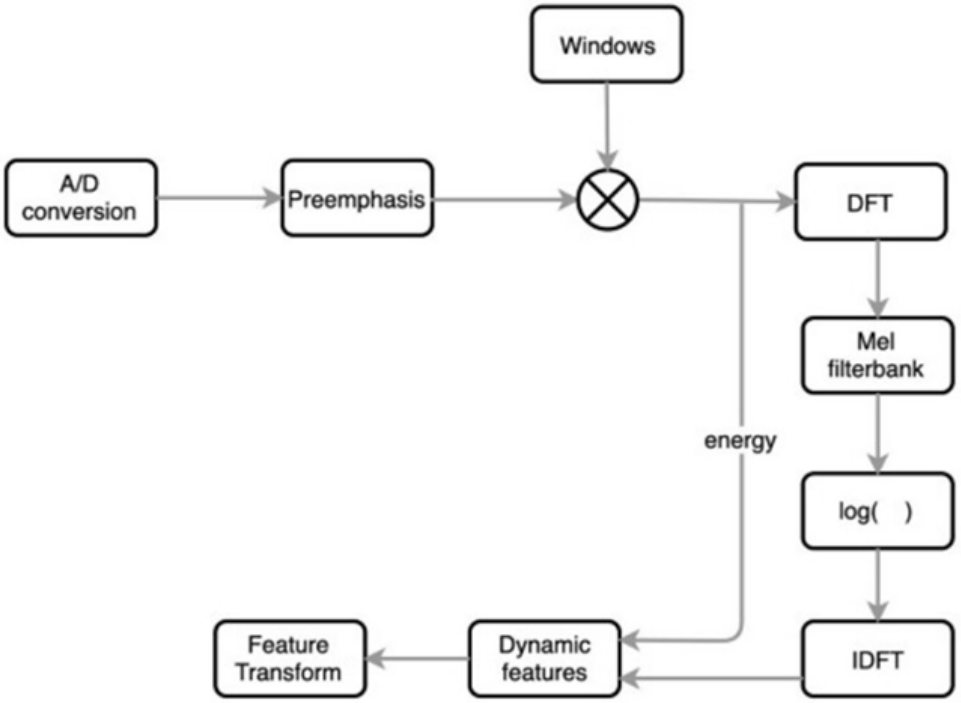
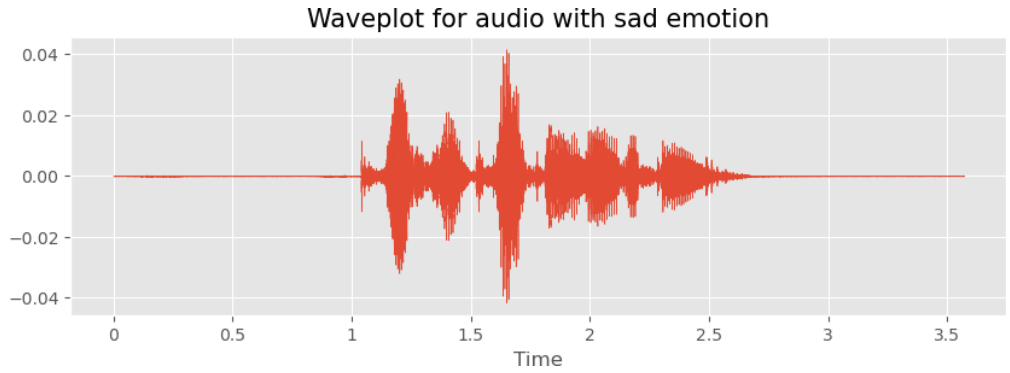


Fig 3.3: MFCC Process [17]

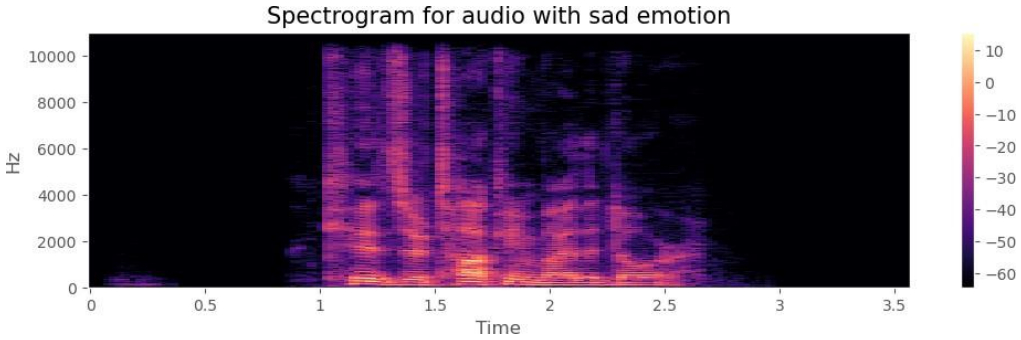
Windowing

The MFCC technique aims to develop the acoustic sign components that may be used to detect telephones in conversation. Anyhow, the supplied sound sign will include a huge number, so we will divide it into many fragments, each of which will be 25 ms wide and separated by 10 ms, as shown in the image below. An individual typically uses four phones to communicate, each of which has three states, for a total of around 36 states every second, or 28 milliseconds per state, which is close to our 25 millisecond boundary.

We'll take 39 features out of each chunk. Additionally, if we chop the signal off at its edges directly during signal splitting. We will therefore use Hamming/Hanning windows to chop the signal instead of a rectangular window, which won't cause noise in the high-frequency range.



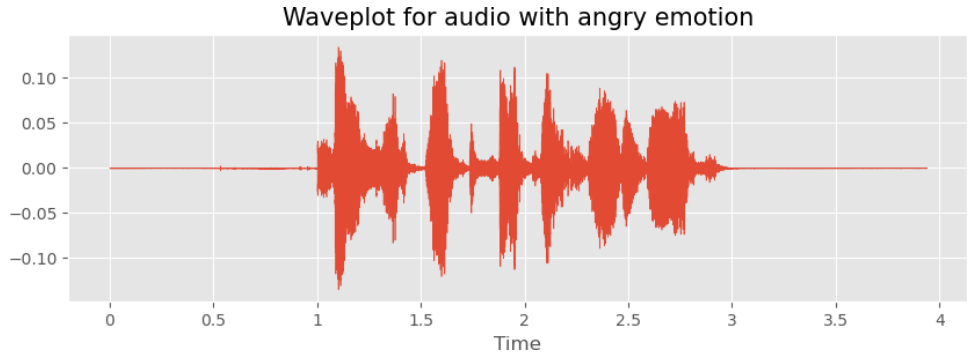
Graph 10



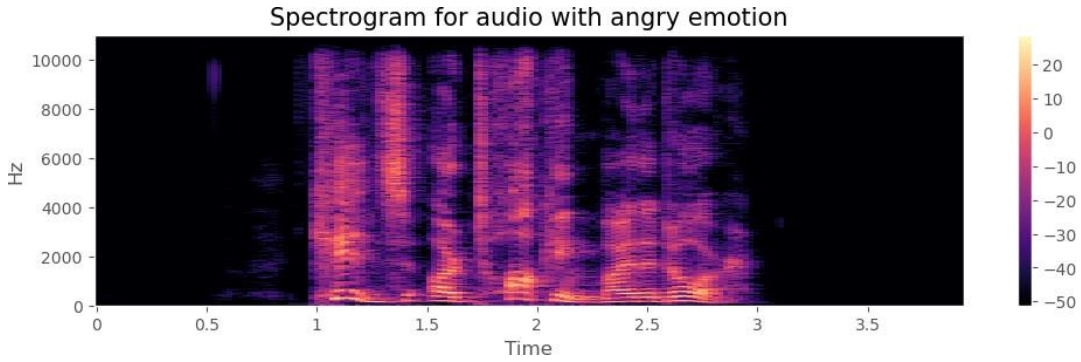
Graph 11

Mel Filter Bank:

$$mel(f) = 1127ln(1 + \frac{f}{700})$$



Graph 12



Graph 13

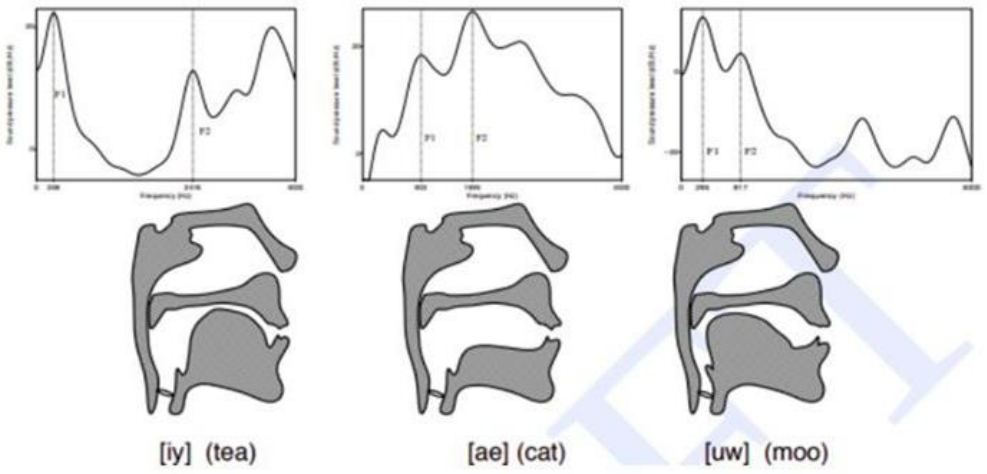
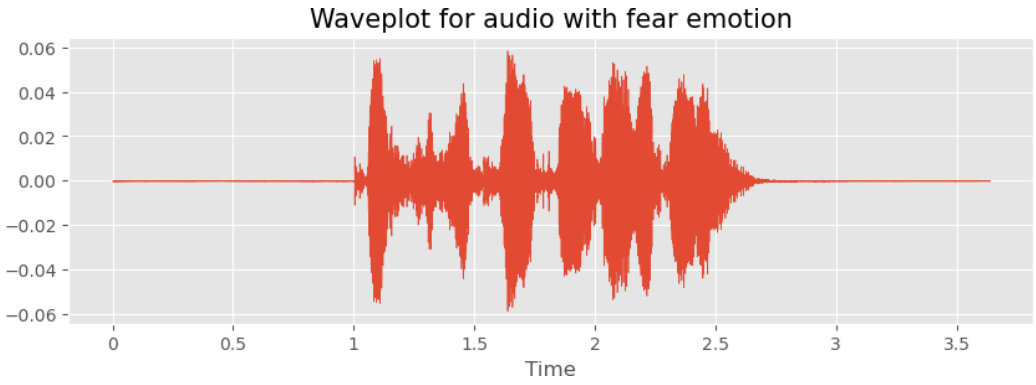


Fig 3.4: MFCC Technique [17]

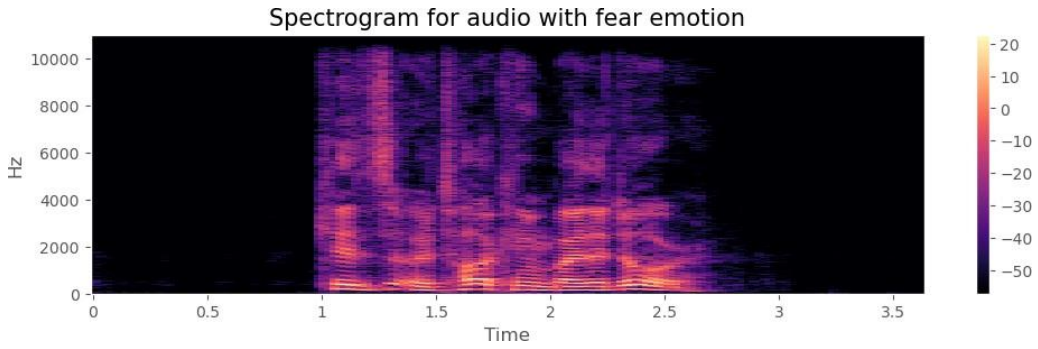
IDFT:

Dynamic Features:

Alongside these 13 elements, the MFCC strategy will consider the main request subordinate and second request subsidiaries of the highlights which comprise another 26 elements. So generally MFCC strategy will produce 39 highlights from every sound sign example whichare utilised as contribution for the discourse acknowledgment model.



Graph 14



Graph 15

Zero Crossing Rate (ZCR)

The Zero-Crossing Rate (ZCR) of a sound edge is the pace of sign-changes of the sign duringthe casing. At the end of the day, it is the time the sign changes esteem, from positive to negative as well as the other way around, separated by the length of the casing. The ZCR is characterised by the accompanying condition:

$$Z(i) = \frac{1}{2W_L}\sum_{n=1}^{W_L}|sgn[x_i(n)] - sgn[x_i(n-1)]|,$$

where sgn (.) is the sign function, i.e.

$$sgn[x_i(n)] = \begin{cases} 1, & x_i(n) \geq 0, \\ -1, & x_i(n) < 0. \end{cases}$$

The zero intersection rate (ZCR) measures how frequently the waveform crosses the zero pivot. It tends to be gotten counting how often both following circumstances are satisfied for a sign :

$$|X(t) - X(t+1)| \geq \in,$$

where $\in$ is a threshold to avoid miscounting zero crossing due to noise.

Root-Mean-Square Energy (RMSE)

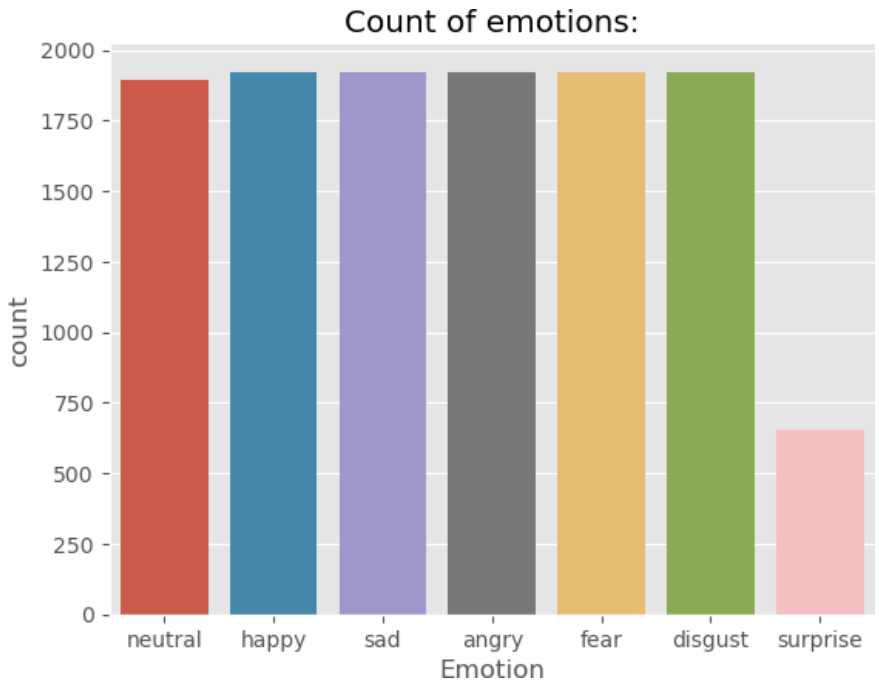
The formal definition of RMS Energy:

$$RMS_t = \sqrt{\frac{1}{K}. \sum_{k=t.K}^{(t+1)(K-1)} s(k)^2}$$

The Root-Mean-Square (RMS) Energy is very like the AE. Rather than beginning identification, in any case, it endeavours to see tumult, which can be utilised for occasion location. Besides, it is substantially more powerful against exceptions, meaning on the off chance that we section sound, we can distinguish new occasions (such an another instrument, somebody talking, and so on) significantly more dependably.

CHAPTER-4 EXPERIMENTS & RESULT ANALYSIS

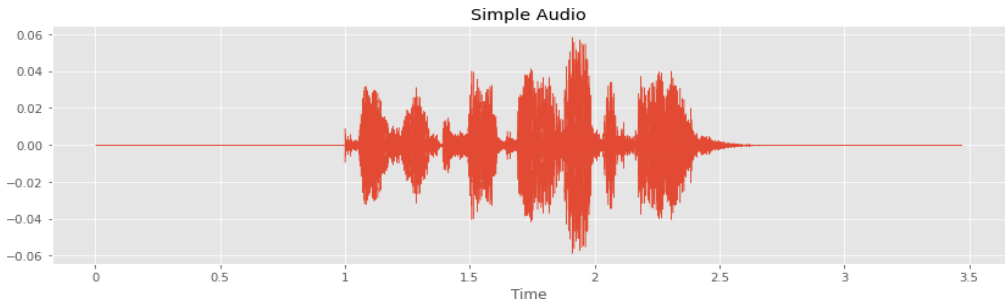
Accordingly, in this model, we obtained model accuracy of around 96% at training data and 71% on the testing data and value loss of about 15% after analysing 48000 samples of data using the MFCC method of feature extraction and the CNN model for training and testing purposes. By using additional feature extraction methods, such as ZCR and RMSE, as well as by providing the model with more data—in our case, we used 48000 samples of data, so the accuracy of the model would increase if the number of samples used were increased. We can further increase the model's accuracy by increasing the number of epochs.



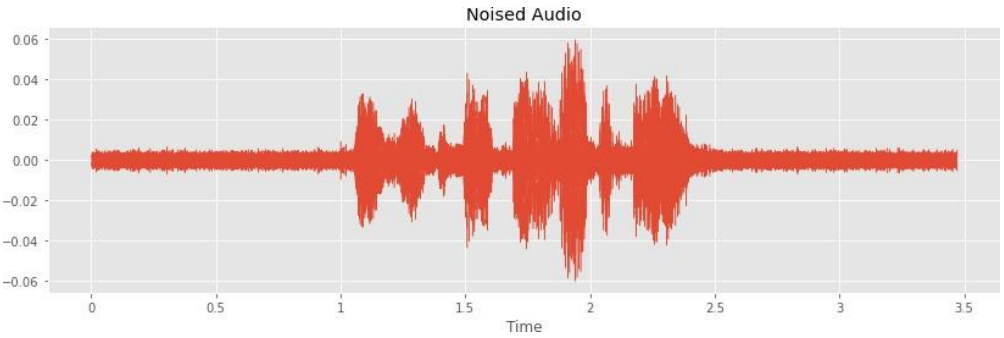
Graph 16

The above figure depicts the count values of different emotions in the whole dataset after combining the all four datasets. It is clearly shown that the surprise emotion data points are the least in numbers compared to the other emotion data points.

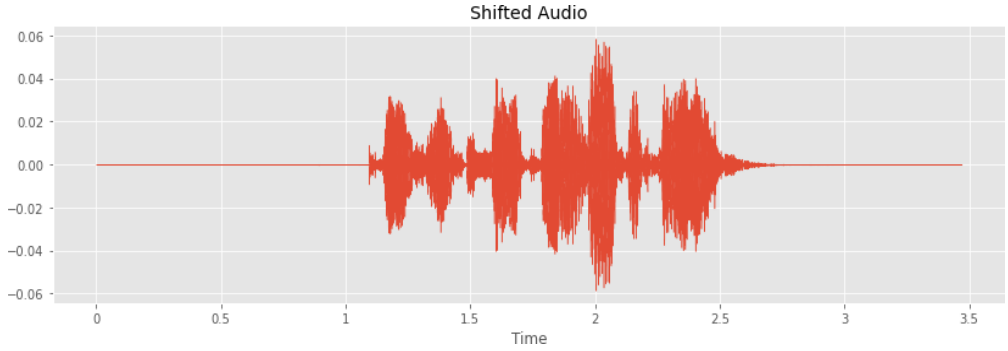
We have also done changes with the audio files by adding noise, elevating the pitch, stretching the audio etc. The following graph shows the different wave plots of the changedaudio samples.



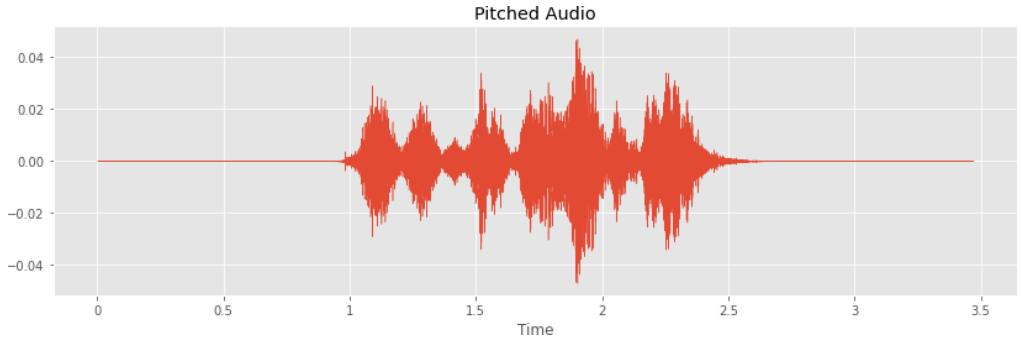
Graph 17



Graph 18



Graph 19



Graph 20

In [49]:

```
model = models.Sequential()
model.add(layers.Conv1D(512, kernel_size=5, strides=1,
                        padding="same", activation="relu",
                        input_shape=(X_train.shape[1], 1)))
model.add(layers.BatchNormalization())
model.add(layers.MaxPool1D(pool_size=5, strides=2, padding="same"))

model.add(layers.Conv1D(512, kernel_size=5, strides=1,
                        padding="same", activation="relu"))
model.add(layers.BatchNormalization())
model.add(layers.MaxPool1D(pool_size=5, strides=2, padding="same"))

model.add(layers.Conv1D(256, kernel_size=5, strides=1,
                        padding="same", activation="relu"))
model.add(layers.BatchNormalization())
model.add(layers.MaxPool1D(pool_size=5, strides=2, padding="same"))

model.add(layers.Conv1D(256, kernel_size=3, strides=1, padding='same', activation='relu'))
model.add(layers.BatchNormalization())
model.add(layers.MaxPooling1D(pool_size=5, strides = 2, padding = 'same'))

model.add(layers.Conv1D(128, kernel_size=3, strides=1, padding='same', activation='relu'))
model.add(layers.BatchNormalization())
model.add(layers.MaxPooling1D(pool_size=3, strides = 2, padding = 'same'))

model.add(layers.Flatten())
model.add(layers.Dense(512, activation='relu'))
model.add(layers.BatchNormalization())
model.add(layers.Dense(7, activation="softmax"))

model.compile(optimizer="rmsprop", loss="categorical_crossentropy", metrics=["acc", f1_m])
```

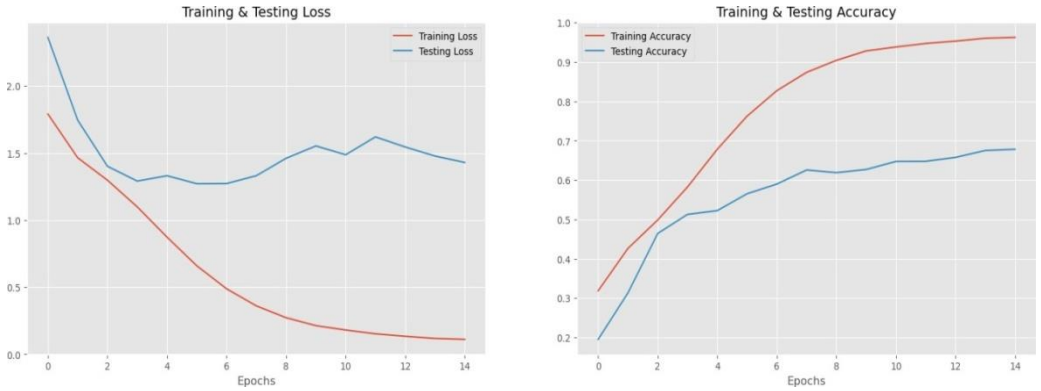
This is basically the CNN function we used for predicting the emotions. There are around 6levels in the model consisting of 3 layers. The layers include the relu function, batch normalisation and max pooling.

```
In [59]:  EPOCHS = 15
          batch_size = 32

In [60]:  history = model.fit(X_train, y_train, validation_data=(X_val, y_val),
                              epochs=EPOCHS, batch_size=batch_size,
                              callbacks=[earlystopping, learning_rate_reduction])
```

```
Epoch 12/15
1095/1095 [==============================] - 122s 111ms/step - loss: 0.1655 - acc: 0.9438 - f1_m: 0.9436 - val_loss: 1.6174
- val_acc: 0.6948 - val_f1_m: 0.6957 - lr: 0.0010
Epoch 13/15
1095/1095 [==============================] - 123s 112ms/step - loss: 0.1521 - acc: 0.9481 - f1_m: 0.9481 - val_loss: 1.5497
- val_acc: 0.7022 - val_f1_m: 0.7025 - lr: 0.0010
Epoch 14/15
1095/1095 [==============================] - 121s 111ms/step - loss: 0.1434 - acc: 0.9519 - f1_m: 0.9521 - val_loss: 1.6115
- val_acc: 0.6984 - val_f1_m: 0.6999 - lr: 0.0010
Epoch 15/15
1095/1095 [==============================] - 120s 109ms/step - loss: 0.1389 - acc: 0.9528 - f1_m: 0.9529 - val_loss: 1.5004
- val_acc: 0.7122 - val_f1_m: 0.7143 - lr: 0.0010
```

We also run around 15 epochs on our model which also helped in increasing the accuracy ofthe model.



Graph 21: Training testing loss and accuracy.

The accuracy of our model is around 96% on the training dataset and 71% on the testing dataset.

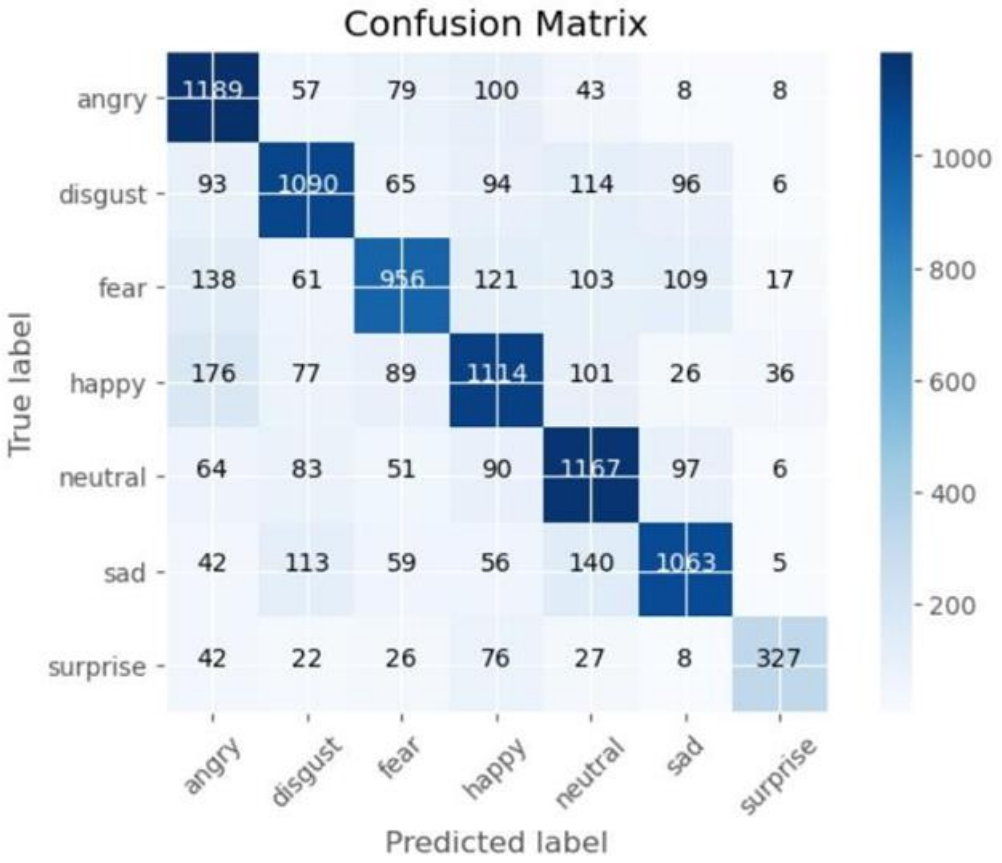


Fig 4.1: Confusion Matrix of the Mode

CHAPTER-5 CONCLUSIONS

5.1 Conclusions

In the project, we tried to use deep learning to analyse certain speech samples. In order to illustrate the various human emotions, we first loaded the datasets using the Librosa libraryand depicted them in the form of various wave plots and spectrograms. Then, we used the MFCC feature extraction method to analyse the acoustic characteristics of all of our samples and organised the sequential data obtained in the 3D array form that the CNN model accepts. Using the Matplotlib library, we put the data into a graphical form, then after some repeated Testing with various values reveals that the model's average accuracy is 71% at testing and 96% at the training phase.

5.2 Future Scope

So the discourse feeling acknowledgment is an extremely fascinating subject and there is something else to find in the field, in our model the future work will incorporate the improvement of exactness of the model to come by improved results, we can likewise prepare the model to give aftereffects of the discourse that is longer in term, like in this model we can perceive the feeling just for brief length of time. In future we will ready to stack the more drawn out example dataset and the model will arrange various feelings in various timeframe. Its future work can likewise incorporate the recording of on time information through a receiver with the goal that there is no need of stacking the dataset; we will just train the model and afterward information can be recorded to give the feelings of that individual's voice.

References

[1] Ahmed, M. R., Islam, S., Islam, A. K. M. M., & Shatabda, S. (n.d.). *An Ensemble 1D- CNN-LSTM-GRU Model with Data Augmentation for Speech Emotion Recognition*.

[2] Shelke, N., Wadyalkar, V., Kotangale, D., Kuyate, N., Nerkar, A., & Gour, N. (n.d.). A NOVEL APPROACH TO EMOTION DETECTION FROM SPEECH.

[3] Hamsa, S., Shahin, I., Iraqi, Y., & Werghi, N. (2020). Emotion Recognition from Speech Using Wavelet Packet Transform Cochlear Filter Bank and Random Forest Classifier. *IEEEAccess*, *8*, 96994–97006.

[4] Liu, Z. T., Wu, B. H., Li, D. Y., Xiao, P., & Mao, J. W. (2020). Speech emotion recognition based on selective interpolation synthetic minority over-sampling technique in a small sample environment. *Sensors (Switzerland)*, *20*(8).

[5] Aggarwal, A., Srivastava, A., Agarwal, A., Chahal, N., Singh, D., Alnuaim, A. A., Alhadlaq, A., & Lee, H. N. (2022). Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning. *Sensors*, *22*(6).

[6] Hajarolasvadi, N., & Demirel, H. (2019). 3D CNN-based speech emotion recognition using k-means clustering and spectrograms. *Entropy*, *21*(5).

[7] Kanwal, S., & Asghar, S. (2021). Speech Emotion Recognition Using Clustering BasedGA-Optimised Feature Set. *IEEE Access*, *9*, 125830–125842.

[8] Mustaqeem, Sajjad, M., & Kwon, S. (2020). Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. *IEEE Access*, *8*, 79861–79875.

[9] Iqbal, A., & Barua, K. (2019). A Real-time Emotion Recognition from Speech using Gradient Boosting. *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1–5.

[10] Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology*, *23*(1), 45–55.

[11] Veerappa, B. (2020). SPEECH EMOTION RECOGNITION. *International Research Journal of Engineering and Technology*.

[12] *Speech Emotion Recognition Using CNN Speech Emotion Recognition Using Convolutional Neural Network (CNN) View project Fire Safety in Indian Coal Mines using Machine Learning Techniques View project Harini Murugan SRMIST*. (n.d.).

[13] Kumar, A., & Iqbal, J. L. M. (2019). Machine Learning Based Emotion Recognition using Speech Signal. *International Journal of Engineering and Advanced Technology (IJEAT)*, *9*, 2249–8958.

[14] Cai, L., Hu, Y., Dong, J., & Zhou, S. (2019). Audio-Textual Emotion Recognition Based on Improved Neural Networks. *Mathematical Problems in Engineering*, *2019*.

[15] Suksri, S. (n.d.). *Speech Recognition using MFCC Arm Support for Rehabilitation Viewproject*.

[16] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Ali Mahjoub, M., & Cleder, C. (2020). Automatic Speech Emotion Recognition Using Machine Learning. In *Social Mediaand Machine Learning*. IntechOpen.

[17] Uday Kiran. (2021) *MFCC Technique for Speech Recognition,* https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-

recognition/

[18] Elaine Rodrigues Ribeiro, André Luiz Cunha. (2020), *Historical traffic flow data reconstrucion applying Wavelet Transform*

[19] Venkat Markapuri, George LaVessi, Robert Stewart, Dan Wagner. (2020), *https://www.researchgate.net/figure/Single-Neuron-Activation-1_fig1_342378314*

[20] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H.,; Alhussain, T. (2019). *Speech Emotion Recognition Using Deep Learning Techniques: A Review.*

[21] Mittal, R., Vart, S., Shokeen, P; Kumar, M. (2022). *Speech Emotion Recognition.*

[22] Dolka, H., M, A. X. v, Juliet, S. (2021). *Speech Emotion Recognition Using ANN on MFCC Features.*

[23] John J. Lee., *https://nulib-oer.github.io/empirical-methods-polisci/machine-learning.html*

[24] Seo, M., Kim, M. (2020). *Fusing Visual Attention CNN and Bag of Visual Words for Cross-Corpus Speech Emotion Recognition*.

[25] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Mansor, H., Kartiwi, M., Ismail, N. (2020). *Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks*.

[26] Ramesh, S., Gomathi, S., Sasikala, S., Saravanan, T. R. (2021). *Automatic speech emotion detection using hybrid of gray wolf optimizer and naïve Bayes*.

[27] Ancilin, J., Milton, A. (2021). *Improved speech emotion recognition with Mel frequency magnitude coefficient*.

[28] Sonmez, Y. U., Varol, A. (2020). *A Speech Emotion Recognition Model Based on Multi-Level Local Binary and Local Ternary Patterns*.

[29] Farooq, M., Hussain, F., Baloch, N. K., Raja, F. R., Yu, H., Zikria, Y. bin. (2020). *Impact of Feature Selection Algorithm on Speech Emotion Recognition Using Deep Convolutional Neural Network*.