Implementation of Speech Emotion Recognition Using RAVDESS dataset

Harshit Jain
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
harshitjaintata@gmail.com

Dhrubo Bhattacharjee
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
dhrubob026@gmail.com

Devansh Tiwari
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
devanshdt2500@gmail.com

Dr.Preksha Pareek
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
preksha.pareek@sitpune.edu.in

Dr.Pooja Kamat
Artificial Intelligence and Machine Learning
Symbiosis Institute of Technology
Symbiosis International (Deemed university), Pune, India
pooja.kamat@sitpune.edu.in

***Abstract*—Speech Emotion Recognition (SER) is a hot topic in academia and industry. Although researchers have done a tremendous amount of work in this field, there are still the issues of speech feature choice and the correct application of feature engineering that remains to be solved in the domain of SER. In this research, we have used different classification techniques like KNN classifier , Decision Tree , Support Vector Machine , Naive Bayes , Multi-Layer Perceptron , CNN , LSTM and compare the analysis among them.The results are compared using these classification techniques using Ryerson Audio-Visual Database of Speech and Song (RAVDESS).**

***Index Terms*—Audio, Emotion, RAVDESS, Feature Extraction, Feature Scaling , KNN classifier, Decision Tree, Support Vector Machine,Naive Bayes, Multi-Layer Perceptron, CNN and LSTM.**

I. Introduction :

Speech is one of the essential means by which humans communicate. Because people express themselves through their emotions, it is only natural for them to use emotion in speech to express their feelings clearly.Speech contains both linguistic and nonlinguistic information. A speech signal carries information such as the intended message, the speaker's identity, and the speaker's emotional condition. Effective Communication via language and voice has facilitated the exchange of ideas, messages, and intellect. Acoustic fluctuation and the words themselves are the two most important components in voice-based signalling. Acoustic properties such as tone, timing, voice quality, and expression of the speech signal strongly match with the elementary emotion due to the effects of nervous system excitation, rise in heart rate, and so on. The fluctuation of these factors serves as the foundation for recognising speech emotions. The procedure for determining a Speech emotion recognition is the detection of a speaker's sentiments from their oration output.Finding these sentiments shows deeper complexities and aids in issue solving in the present. Emotion identification from speech is the single most difficult threat in the realm of human computer relation. The development of effective sentiment identification systems is thus advantageous, and the goal of a good sentiment identification system is to be able to replicate human intellect in the same way that people can recognise sentiments such as anger, sorrow, and joy while conversing. Despite extensive research, there are still a number of challenges to overcome in the field of sentiment identification from speech, such as incomplete databases, poor recording quality, cross-database performance, and difficulties with speaker independent recognition due to each individual's unique speaking style. Speech Emotion Recognition (SER) systems are pattern identification systems which comprises three parts: (1) speech signal acquisition, (2) feature extraction, and (3) emotion recognition using classifiers(fig.1). Speech features, qualitative features, spectral

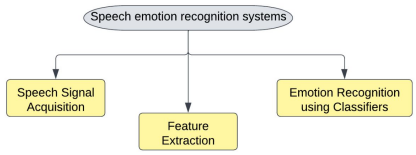

Fig. 1.

features, and Teager energy operator (TEO)-based features are the four categories of features. Audio- based emotion identification systems can help in intellectual health evaluation. Speech-processing mechanics can be used to diagnose and assess the intensity of illnesses. It can also help with speech therapy and its goal is to help people with speech issues. Furthermore, applications such as health care and counselling can help from such self-operating solutions. Speech recognition systems are especially convenient where man and machine interaction (MMI) is needed such as, in web movies, computer tutorial applications, online learning, call centre communications and so on, because the reaction in such systems is based on the user's emotions. The purpose of such technologies in contact centre talks would be to detect the caller's urgency and emotional state. This would help to upgrade the performance of call centres, particularly those providing health care support to the elderly and emergency call centres. It is also used in vehicle systems to determine the driver's mental state, which is linked to the possibility of reckless driving and subsequent collisions. The drivers and other road users' safety can be ensured by the help of these devices. It can also be used in forensic and criminal investigations to detect lies. Furthermore, the study demonstrates its use in identifying school violence based on children's speech. Finally, it can improve other AI applications such as marketing, intelligent toys, and music playing that is tailored to the caller's mood.

II. Contribution of this research :

Speech emotion recognition has been the subject of a significant amount of research (SER). We provide a quick overview of the research on emotion recognition from audio in this section. The foundation of many current research methodologies is two distinct classification strategies. The first is the use of traditional classifiers like SVM and artificial neural networks (ANN), and the second is the use of deep learning- based classifiers like convolutional neural networks (CNN) and deep neural networks (DNN) (Akc ay and Ouz). [2020] USML Research paper 5 SVM was utilised as a classifier for anger recognition using both linguistic (probabilistic and entropy-based models of words and phrases) and acoustic (pitch, loudness, and spectral characteristic) feature modelling (Polzehl et al. [2011]). According to this study, audio modelling works better than linguistic modelling. With the WoZ database and roughly 79 percent of the IVR datasets, accuracy was attained. The single task (ST), multi-task feature selection/learning (MTFS/MTFL), and group multi-task feature selection/learning (GMTFS/GMTFL) models were all used in Zhang et al. (2016) to solve the binary classification issue. For the purpose of classifying the different emotions, four models were utilised after feature extraction of the acoustic low level descriptors (LLDs). The RAVDESS dataset was used for testing, and the highest accuracy possible was 64.29 percent.Several scholars have employed support vector machines as a classification method. Using three separate datasets, namely IITKGP-SEHSC, RAVDESS, and Berlin EMODB, feature extraction utilising MFCCs, Spectral Centroids, Delta and Delta-Delta MFCCs, together with a bagging ensemble with SVM as a classifier, was employed for speech detection. Using the suggested methods, 75.69 percent accuracy was attained on the RAVDESS dataset (Bhavan et al. 2019). Another work by Tomba et al. (2018) sought to identify stress using MFCC characteristics, mean energy, and mean intensity in speech analysis. Accuracy values of 78.75 percent and 89.16 percent were obtained on the RAVDESS dataset using SVM and neural networks. In Deb Dandapat (2016), residual sinusoidal peak amplitude (RSPA), a relatively new feature, was chosen as a feature for emotion classification. Using a sinusoidal model, the RSPA feature is assessed using the speech signal's LP residual. On the EMO-DB dataset, the SVM classifier was once more used and tested, yielding a maximum accuracy of 74.4 percent. Also, the capacity of conventional speech representations like the mel spectrogram, magnitude spectrogram, and Mel-Frequency Cepstral Coefficients (MFCC's) to capture emotions has been tested using architectures like the convolutional neural network (CNN) and long short-term memory(LSTM). Convolutional neural networks and bidirectional long short term memory networks were employed, and Pandey etal CNN.'s + BLSTM architecture with MFCC as input for EMO-DB attained the best accuracy of 82.35 percent. (2019). In Jannat et al(2018) .'s evaluation of a convolutional neural network model using RAVDESS, the accuracy of the only audio tests is fairly poor at 66.41 percent. One 1D CNN LSTM network and one 2D CNN LSTM network were built in Zhao et al(2019) .'s research in order to learn regional and global emotion-related characteristics from speech and the log-Mel spectrogram, respectively. In the Berlin EmoDB of USML Research paper 6 speaker-dependent and speaker-independent experiments, accuracy was 95.33 percent and 95.89percent, while on the IEMOCAP database of speaker-dependent and speakerindependent experiments, accuracy was 89.16percent and 52.14percent, respectively. 63.5percent accuracy with the IIEMOCAP corpus was achieved using recurrent neural network (RNN) architectures in Mirsamadi et al (2017). In order to categorise the mel spectrograms derived from the voice samples of the RAVDESS dataset, Popova et al. (2018) utilised a fine-tuned DNN. The accuracy of the classifier used by the authors, the VGG-16 network, was 71 percent. In Denget alproposal, .'s a sparse autoencoder technique for feature transfer learning for speech emotion recognition was put out (2013). The datasets were accurate on average by 51.6 percent (original) and 59.9 percent (reconstructed). Deng et al. created the semisupervised autoencoder (SS-AE) to learn from both labelled and unlabeled data (2018). A well-liked unsupervised deep denoising autoencoder is extended. SS-AE-Skip, a variation of SS-AE that adds skip connections from the lower layer to the upper one, was also put into practise. The average UAR for SS-AE and SS-AE-Skip is 42.7 percent and 42.8 percent, respectively. By simultaneously learning common knowledge from labelled and unlabelled data, Deng et al. (2017) also integrate Universum learning to a deep autoencoder, which reduces the intrinsic mismatch between the training and test data.The accuracy

Section-3	Proposed Work- 3.1- Dataset Description 3.2- Pre-processing 3.3- Feature Extraction 3.4- Standardization 3.5- Classifiers
Section-4	4.1- Result Analysis 4.2-Conclusion 4.3- Future Scope
Section-5	5.1- Reference

TABLE I

Gender	Count	Trials per Actors	Total Count
Male	12	60	720
Female	12	60	720
Total			1440

Fig. 2. Total Count Of Files

of the Universum Autoencoder, which is 59.3 percent, is comparable to the SVM UAR's accuracy of 54.1 percent. After MFCC feature extraction, the model in Aouani and Ben Ayed (2018) constructs a stack and a simple auto encoder. With a classification rate of 69.84 percent and 68.25 percent utilising 39 MFCC, respectively, the experimental findings demonstrate that the DSVM technique beats the conventional SVM. Also, with a classification rate of 73.01 percent, the auto-encoder approach exceeds the conventional. SVM. The rest of the papers is organized as follows:(TABLE 1:)

III. Proposed Work:

A. 3.1 Dataset description -

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) is a database of emotional speech and song. The RAVDESS dataset contains a total of 7356 files (24.8 GB) of audio and video, spoken and music. It is a multimodal and dynamic set of North American English facial and vocal expressions. The collection features 24 performers who deliver two lexically matched lines in a neutral North American accent. Neutral, calm, happy, sad, angry, terrified, disgusted, and surprised are the eight speech emotions. Each and every expression has two strength levels: strong and normal, with a neutral expression as an extra. All data is provided in three modes: audio-video, audio solo, and video only. In this investigation, only the audio files, comprising 1440 files (24 actors * 60 trials per actor), are used. Figures 2 and 3 depict the distribution of RAVDESS dataset wave files and filename identifiers, respectively.The audio file naming strategy used in this dataset is depicted in Figure 4.

B. 3.2 Speech Processing -

The speech of the target speaker as well as background noise, the voices of non-target speakers, and reverberation are all present in the recoded audio signals. Interference sounds are automatically decreased by using various speech improvement technologies such as speech processing . Speech

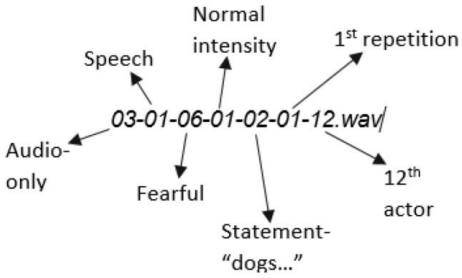

Fig. 3. filename identifier

Modality	01 = full-AV, 02 = video-only, 03 = audio-only
Vocal Channel	01 = speech, 02 = song
Emotion	01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry,06 = fearful, 07 = disgust, 08 =
Intensity	01 = normal, 02 = strong (Note: Strong intensity for neutral emotion is not there)
Statement	01 = "Kids are talking by the door" 02 = "Dogs are sitting by the door"
Repetition	01 = 1st repetition, 02 = 2nd repetition
Actor	01 to 24 Male: Odd numbered actors Female: Even numbered actors

Fig. 4. Naming convention of Audio File

processing involves changing or extracting principal information from signals by determining their basic properties. The following processes comprise speech processing.

1) 3.2.1 Pre Processing -: The step for pre-processing follows data collection. The gathered data would be utilized to train the classifier in an SER system. While only a few of these pre-processing approaches are employed for feature extraction, others take care of feature normalization so that fluctuations in the speech recordings do not affect the feature extraction process.

2) 3.2.2 Data augmentation -: We can generate new synthetic data samples by making tiny changes to our existing training set, a process known as data augmentation. To generate semantic data for audio-noise injection, time shifting, tone and speed variations, and other techniques can be used. We aim to increase generalization and make our model insensitive to these perturbations. While librosa (Library for Recognition and Organisation of Speech and Audio) helps with pitch and speed modulation, numpy provides a straight- forward method for dealing with noise injection and modifying time. Different augmentation steps that we used were: Different augmentation steps that we used were:

1) Noise injection(Fig.6)
2) Trimming the audio(Fig.7)
3) Shifting of audio(Fig.8)

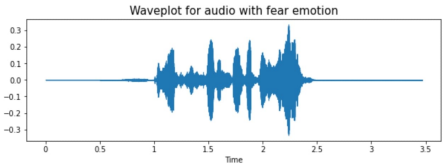

Fig. 5. Raw Audio waveplot

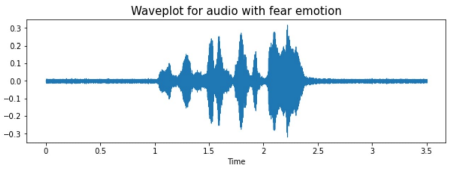

Fig. 6. Noise injected audio waveplot

4) Changing the pitch of the audio(Fig.9)

C. 3.3 Feature Extraction -

The majority of SER's attributes are speech-related. The rate of recognition increases with each well-planned combination of traits that accurately represents every emotion. A variety of attributes have been employed in SER frameworks. There is no commonly agreed feature arrangement for exact and specific classification. The studies that are currently available are all experimental. As we all know, speech conveys information and emotion in varying periods. As a result, depending on our requirements, we can extract either global or local features. There are other extraction methods like:

1) Zero Crossing Rate:: It is the rate at which the signal crosses the horizontal axis during the duration of a particular frame.
2) Energy: The sum of squares of the signal values, normalized by the specific frame length.
3) Entropy of Energy:: It measures disorder or impurity in the system.

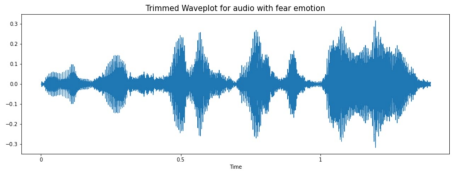

Fig. 7. Trimmed audio waveplot

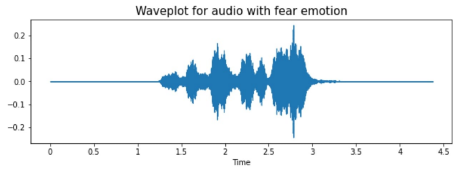

Fig. 8. shifted audio waveplot

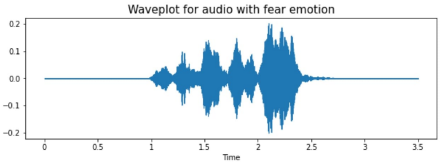

Fig. 9. Changed pitch audio waveplot

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

where $\operatorname{sgn}(\cdot)$ is the sign function, i.e.

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

Fig. 10. Zero-Crossing Rate Equation

4) Spectral Centroid:: It indicates the center of mass of the spectrum.
5) Spectral Spread: It uses broader bandwidth than the primary message while keeping the similar signal power.
6) Spectral Entropy: : It measures spectral energy distribution of a signal.
7) Spectral Flux:: It measures the rate at which the power spectrum of a signal is changing.
8) Spectral Rolloff:The frequency below which 85 percent of the magnitude scattering of the spectrum is intense.
9) MFCCs: Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are non linear but distributed according to the mel-scale.

However, we concentrated on:

1) 3.3.1 ZCR (Zero-Crossing Rate) - : The Zero-Crossing Rate (ZCR) of an audio frame quantifies how frequently the sign of the signal shifts throughout the frame. In other words, it is calculated by dividing the number of times the signal's value shifts from positive to negative by the duration of the frame.(Fig. 10)

2) 3.3.2 MFCC (Mel-Frequency Cepstral Coefficients) - : The Mel Frequency Cepstral Coefficient is the most often used spectral characteristic in automatic speech recognition(MFCC). MFCC seeks to recreate how our ears work by processing speech waves linearly at low frequencies and logarithmically at high frequencies. The envelope of the short-time power spectrum, which characterizes the geometry of the vocal tract, is represented by MFCCs. During framing, the characteristics of the voice signal are divided into frames for proper analysis. Because speech signals are not stationary, their properties vary every 15 milliseconds. To generate MFCC, the short-time discrete Fourier transform is employed to separate the utterances into distinct segments before translating them into the frequency domain. The Mel filter bank is used to determine many sub-band energies. The logarithm of the relevant subbands is then computed. Finally,

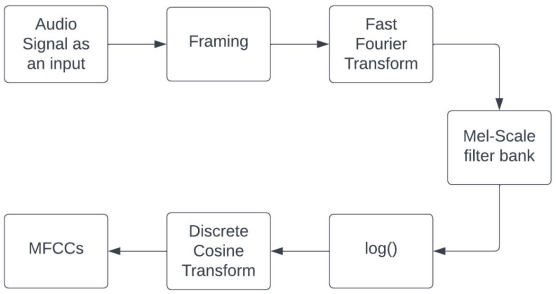

Fig. 11. MFCC Flow chart

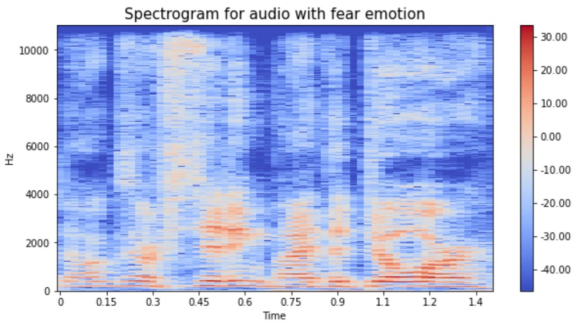

Fig. 12. Mel Spectogram

the inverse Fourier transform is employed to compute MFCC. Mel(f) = 2595 log10 1+f/1000, where f represents the actual frequency of speech. The link between speech frequency and Mel scale can be calculated as follows: Frequency (Mel) = [2695 log (1 + f(Hz))/700]. (Fig. 11)

3) 3.3.3 Mel Spectrogram - : : We use the Mel spectrogram to provide our models with sound data that is close to what a human would hear. Filter banks are applied to raw audio waveforms to generate the Mel spectrogram. Following this technique, each sample has a 128 x 128 shape, indicating the use of 128 bits filter banks and 128-time steps per clip.An Example of Mel Spectrogram is shown in (Fig. 12): From all these, MFCC fits best for feature extraction.

D. 3.4 Standardization -

It's conceivable that you will run over datasets with bunches of mathematical commotion worked in, like change or in an unexpected way scaled information, so a decent preprocessing is an unquestionable requirement before contemplating AI. A decent preprocessing answer for this sort of issue is frequently alluded to as standardization. Standardization is a preprocessing strategy used to change constant information to make it look typically dispersed. In scikit-learn this is much of the time an essential step in light of the fact that many models expect that the information you are preparing on is regularly disseminated, and in the event that it isn't, your gamble biasing your model.You can standardize your information in various ways, scaling is one of them.

1) 3.4.1 Standard Scaler - : Standard scaling is an standardization process commonly used in machine learning for data prepocessing and statistical analysis. It's purpose is to transform numerical data so it has mean of zero and a standard deviation of one. This process involves subtracting the mean of the data from each data point and then divide it by standard deviation. This transformation centers the data at zero and scales it to have unit variance.After this scaling process, the data would have mean of 0 and SD of 1.

E. 3.5 Classifiers -

Speech emotion recognition classifies the underlying emotions for any utterance. There are two methods for classifying SER: (a) conventional classifiers, and (b) deep learning classifiers. With the SER system, a variety of classifiers have been used, however it is challenging to determine which performs best. As a result, current research is largely practical. SER systems often make use of a number of conventional classification algorithms. A new class input was foreseen by the learning process, which calls for labelled data that identifies the relevant classes and samples by roughly approximating the mapping function. The remaining data is used to test the classifier's performance after training. There are several classifiers that we have use that is:

1) 3.5.1 KNN Classifier (K-Nearest Neighbour)- : : A new instance is classified using KNN, a supervised learning technique, based on the nearby training examples that are present in the feature space. It just relies on memory and does not fit using any models. The class that is most prevalent among a test data's k closest neighbours is chosen when it is entered. KNN classifier is non-parametric approach used for classification. It is not necessary to have any prior knowledge of the structure of the training set's data. If the current training set is expanded to include the new training pattern. Any ties may be severed arbitrarily. The neighbourhood categorization serves as the prediction value for the new query instance in the KNN method. The local structure of the data has an impact on the KNN algorithm The results showed that KNN algorithm achieved 77 percent accuracy in classifying emotional states in the RAVDESS.

2) 3.5.2 Decision Tree)- : Decision trees are a popular technique for categorization tasks. They represent decisions and their outcomes using a tree-like paradigm, with each node in the tree representing a decision based on a feature of the dataset and each branch reflecting the outcome of that decision. Decision trees operate by recursively dividing the dataset into smaller subsets depending on the attributes that best distinguish the classes. Based on voice recordings, decision trees can be used to classify the emotional states of speakers in the RAVDESS dataset. The programme can predict the emotional states of fresh recordings by training on a portion of the dataset with labelled emotional categories. Based on the impurity of the classes in the subsets, the decision tree method selects the appropriate feature to split the dataset

at each node. By applying Decision Tree algorithm it gives 49 percent accuracy on RAVDESS dataset.

3) 3.5.3 Support Vector Machine(SVM)- : The technique works by determining the best hyperplane for separating the various classes in the dataset. The decision boundary that maximises the margin is determined as the separation between the hyperplane and the nearest data points from each class. SVMs can be used to classify the emotional states of speakers based on their voice recordings in the context of the RAVDESS dataset. The algorithm can be trained on a subset of the dataset that has been labelled with emotional categories, and then the trained model can be used to predict the emotional states of new recordings. The results showed that KNN algorithm achieved 89 percent accuracy in classifying emotional states in the RAVDESS.

4) 3.5.4 Naïve Bayes- : A popular probabilistic algorithm in machine learning for classification tasks is called Naive Bayes. It is based on the Bayes theorem, which says that you may determine the probability of a hypothesis or class label given the data by multiplying the likelihood of the data given the hypothesis by the prior probability of the hypothesis, then dividing that result by the probability of the data. The algorithm makes the basic assumption which frequently proves to be correct in practice that the features in the dataset are independent of one another. Naive Bayes can be used to classify the emotional states of speakers based on their voice recordings in the context of the RAVDESS dataset. The system can be trained on a subset of the dataset with labelled emotional categories and then used to predict the emotional states of additional recordings. The result showed that by applying Naïve Bayes algorithm it gives 46 percent accuracy.

5) 3.5.5 Multi-Layer Perceptron (MLP)- : The Multi-Layer Perceptron (MLP) is a neural network method that is extensively used in machine learning for classification applications. It is made up of several layers of artificial neurons, with each neuron in one layer linked to all neurons in the preceding layer. Each layer's neurons apply a non-linear activation function to the weighted sum of the previous layer's inputs. The projected class label is represented by the output of the final layer. MLP can be used to categorise speakers' emotional states based on their voice recordings in the context of the RAVDESS dataset. A portion of the dataset with labelled emotional categories can be used to train the algorithm, and the learned model can then be applied to forecast the emotional states of new recordings. By applying Multi-Layer Perceptron (MLP) algorithm it gives 84 percent of accuracy on RAVDESS dataset.

6) 3.5.6 Convolutional Neural Networks(CNN)- : Convolutional Neural Networks (CNN) are a type of neural network that is extensively employed in machine learning for image categorization problems. The algorithm is made up of layers such as convolutional layers, pooling layers, and fully connected layers. The convolutional layers process the input image by applying filters that extract features such as edges and textures. The pooling layers reduce the size of the feature maps by down sampling the output of the convolutional layers.

Classifiers	Accuracy
KNN Classifier	77%
Decision Tree	49%
Support Vector Machine	89%
Naive Bayes	46%
MLP	84%
CNN	93%
CNN and LSTM	82%

TABLE II

The retrieved features are then used to classify the image by the fully linked layers. CNN can be used to classify speakers' emotional states based on audio recordings. The method can transform voice recordings into spectrograms, which are two-dimensional representations of sound waves. The CNN can then learn to classify the spectrograms based on the dataset's emotional categories. In the RAVDESS, the CNN algorithm obtained among the best accuracy of 93 percent among all the classifiers which are applied on the dataset.

7) 3.5.6 Long Short-Term Memory (LSTM)- : Machine learning problems involving sequence categorization frequently use recurrent neural networks of the Long Short-Term Memory (LSTM) variety. The network can read, write, and delete specific pieces of information from the memory thanks to the algorithm, which is made up of memory cells and gates. The gates, which control the flow of information into and out of the memory cells, include input gates, output gates, and forget gates. LSTM can be used to categorise the emotional states of speakers using their voice recordings in the context of the RAVDESS dataset. The system can extract a series of acoustic features from the voice recordings, including pitch, duration, and spectral properties.

IV. Analysis :

A. 4.1 Result Analysis -

We used KNN classifiers which gave us the accuracy of 77% , Decision Tree did not perform well on the dataset and gave us the accuracy of 49%, Naive Bayes also didn't perform well and gave the accuracy of 46% ,the best among all the Machine Learning algorithms was Support Vector Machine which performed comparatively better on the dataset giving us the accuracy of 89%. Then we applied Multi-Layer Perceptron which gave us the accuracy of 84%. The best model amongst all the others was CNN which performed exceptionally well and gave the accuracy of 93%. Finally we applied CNN and LSTM together which gave us the accuracy of 82%.(Table 2.)

B. 4.2 Conclusion -

This is the first time, to the best of our knowledge, that we have used the RAVDESS dataset to apply Machine Learning approaches to our model. Machine Learning approaches cant provide better accuracy so we had to use Neural Networks in which CNN provided the best accuracy. We can use other Deep Learning models for our dataset.

C. 4.3 Future Scope -

In the future, we plan to look more into deep learning algorithm, in such a way that our model will give us the highest possible accuracy. The next challenge that we need to solve is Language Independency.Right now,we are expecting that our model perform good atlest on all forms of English accent.

V.

[2] [4] [3] [1]

REFERENCES

[1] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4), 2021.

[2] Apeksha Aggarwal, Akshat Srivastava, Ajay Agarwal, Nidhi Chahal, Dilbag Singh, Abeer Ali Alnuaim, Aseel Alhadlaq, and Heung-No Lee. Two-way feature extraction for speech emotion recognition using deep learning. *Sensors*, 22(6), 2022.

[3] Mustaqeem and Soonil Kwon. A cnn-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, 20(1), 2020.

[4] Tanvi Puri, Mukesh Soni, Gaurav Dhiman, Osamah Khalaf, Malik Alazzam, and Ehtiram Khan. Detection of emotion of speech for ravdess audio using hybrid convolution neural network. *Journal of Healthcare Engineering*, 2022:1–9, 02 2022.