Extraction of signal features in Voice Signals to train Machine Learning-based Classifier algorithms for Emotion Detection

Simran Somani[1], Bhagyashree Shah[1], Bhisaji C. Surve[2]

[1]Student of MBA TECH (IT), MPSTME, NMIMS University, Mumbai, India
[2]Asst. Professor, Dept. of IT, MPSTME, NMIMS University, Mumbai, India
Emails: simran.somani@nmims.in; bhagyashree.shah@nmims.in;bhisaji.surve@nmims.edu

Abstract

This research aims to detect human emotions using speech signals through the development and implementation of methodologies, namely the frequency domain synthesis. To achieve improved results, various machine learning and deep learning models were applied for implementation and their resulting model performance was analyzed. The research findings revealed that each model exhibited different accuracy rates for different emotions but weighted accuracy is best for deep learning based model. This study provides valuable insights into the feasibility and effectiveness of utilizing different methodologies and models for emotion detection through voice signals synthesis. The audio signals are synthesized for Mel-Frequency Cestrum Coefficients (MFCC), Chroma, and MEL characteristics, which are then used as features to train the various machine learning-based classifiers. Python libraries like Librosa, Sklearn, Pyaudio, Numpy, and sound files are used to analyze voice modulations and identify emotions.

Keywords: MFCC; Emotion Detection; Machine Learning; Neural network

1. Introduction

Emotion is a natural element of every living being including plants as some scientific experiments demonstrate it. In the case of Human beings, it is vital as a human is a social animal and human has various mean to express their emotions. The various means by which human can express emotions are by voice in their speech, facial expression, or text used in written communication. Emotion detection can be a very useful use case for technology-based tool development. In Verbal talks audio signal can be processed to detect emotions in the speech, for facial detection video signal can be processed, and in textual matter it is natural language processing can be employed.

It is a very challenging task to detect exact emotional states from audio signals due to various reasons. Hence identification of the best representative of the respective emotions as a feature to be used for machine learning is the vital task. There are various applications of speech emotion detection in human interactions as customer service, teaching-learning, medical analysis, voice synthesis, personal interviews, forensics etc.

Another challenge is the presence of different languages, accents, phrases, speaking styles, and speakers because these factors affect most of the extracted attributes, such as pitch and energy, in a direct manner. Additionally, because each emotion correlates with a different component of speech signals, it is conceivable to have many instances of a given emotion in a single speech signal. As a result, establishing the limits between different emotional components is a very difficult process. Studies have considered the process of classifying emotions across multiple languages but audio-based speech synthesis will be language-independent.

2. Literature survey

A method for calculating the MFCC coefficients , STE (Short-Term Energy) , and Audio Pitch to identification of irritation, pleasure, and depressed using audio samples is performed by Girija Deshmukh et al.[1] using the entire

Doi : https://doi.org/10.54216/JCHCI.070202
Received: October 28, 2023 Revised: January 24, 2024 Accepted: March 21, 2024

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset and employing a multi-class SVM classifier.

In [2], Peng Shi developed and demonstrated that as compare to SVM and ANN algorithms DBN (Deep Belief Networks) performed well and there is about 5% accuracy improvement.

The J. Uma Maheswari et al [3] uses GLCM (Gray Level Co-occurrence Matrix) and MFCC (Mel Frequency Cepstral Coefficient) as features of audio. Reserachers employed KNN (K-nearest Neighbour) and PRNN (Pattern Recognition Neural Network) on the audio features.

The audio properties such as sound, format, and phoneme, which are detected through MFCC, can be use as signal features for training and predictions of emotions is expressed in According to M.S.

Asaf Varol et al. in [6], The reserachers suggest that various machine-learning procedures can be employed on numerous types of datasets with varied varieties of tests in order to yield better success rates.

3. Research Methodology

The research journey is executed through the following steps:

Research Framework:

Step-by-step process of machine learning based emotion detection is proposed as followed in figure 1.

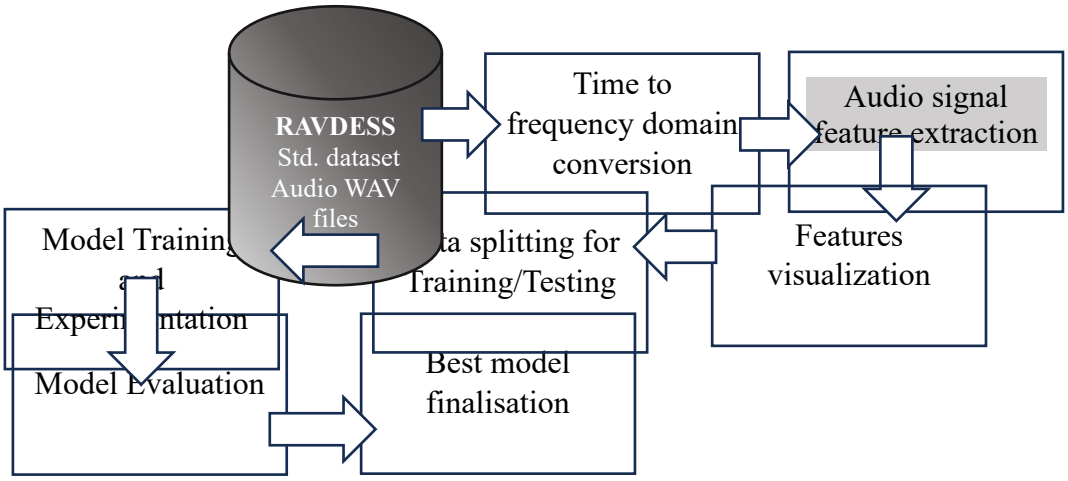


Figure 1: Framework

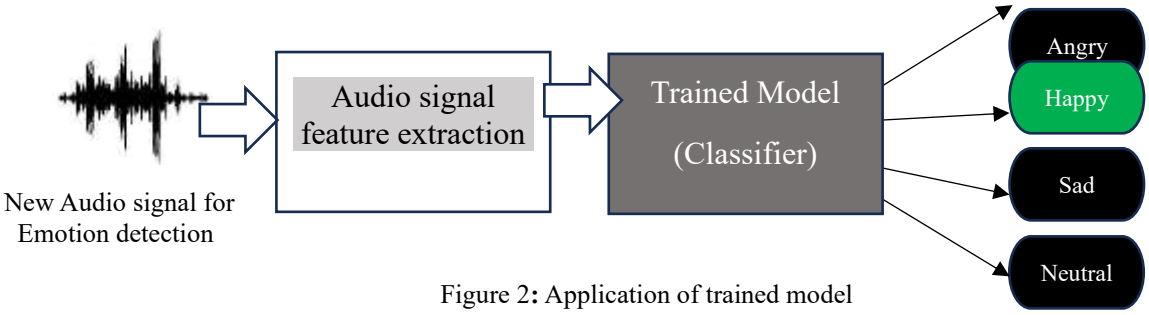


Figure 2: Application of trained model

After going through respective steps and once the trained model of classifier with good overall performance is developed. It can be deployed to online App as Emotion Detection of speech signals in various contextual applications.

3.1 Study of Dataset

The RAVDESS dataset was selected as it contains 247 untrained Americans song files as well as voice which are classified to 8 diverse emotions at 2 intensity levels as Sad, Angry, Calm, Happy, Horrible, Hatred, or Astonishment beside with a standard normal voice for each actor.

Doi : https://doi.org/10.54216/JCHCI.070202
Received: October 28, 2023 Revised: January 24, 2024 Accepted: March 21, 2024

Given that there are 12 men and 12 women among the 24 professional actors in the dataset, it is gender-balanced. Each audio clip was produced in a controlled setting, and it contains the same sentences delivered with an American accent. There are also two other categories of files:

- 1440 files are comprised in the speech file
- 1012 files are comprised in the song file

The files all have a 48 KHz sample rate with 16-bit bitrate and, and they are all in the WAV raw audio file format. Since the audio files in the dataset are all uncompressed, lossless audio files, they have not been altered from the original recording or lost any information or data.

File names in this data set are encoded in numbers as shown figure:

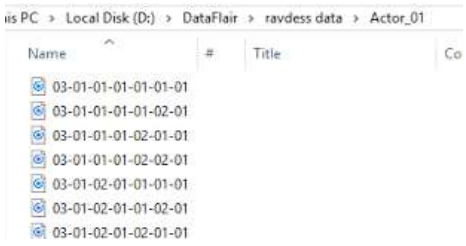


Figure 3**:** Audio files with encoded names (Data set)

DATA File names are encoded as: 03(audio)-01(speech)-06(fearful)-01(Normal intensity)-02(Statement "dogs")-01(1st Repetition)-12(12th Actor).

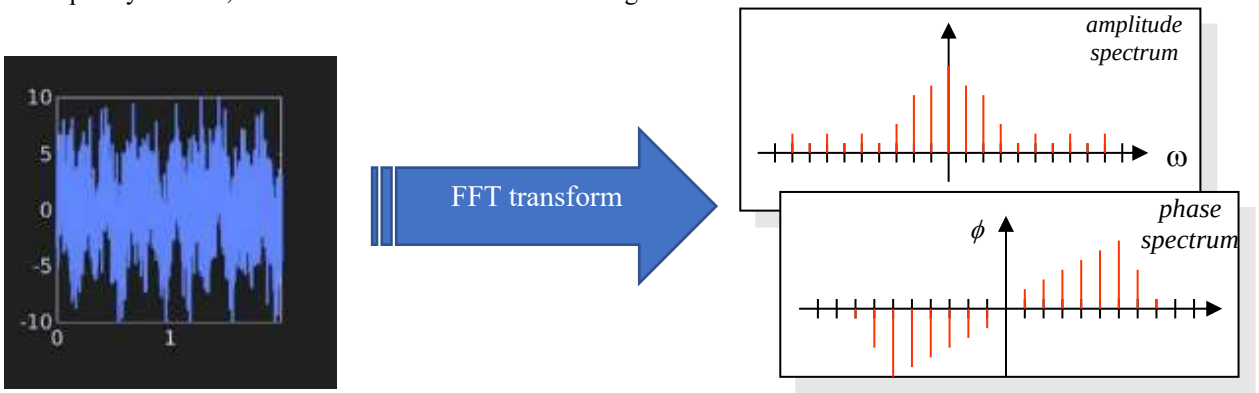
In this research work, only audio speech files are use from data set and that only four major emotions as “HAPPY”, “SAD”, “ANGRY” and “NEUTRAL” are considered for training.

3.2 Time to frequency domain Transformation of Audio signal

The Time audio signal displays the sound wave's strength i.e. loudness expressed as amplitude with respect to the Time axis. Amplitude = 0 corresponds to silence or pause.

The same audio signal in the frequency domain captures various frequency components in the audio spectrum and give bar chart of each component in terms of amplitude which gives better insight in terms of audio characterisation. The following figure give three different audio tunes in terms of time domain vs frequency domain representation.

Fourier Transform is a mathematical means to convert the time domain to the frequency domain.
In frequency domain, it is to extract features from audio signal.



Time domain Audio signal Frequency domain of Audio signal

Figure 4**:** Time vs frequency domain signal transforms

Fourier Transform is a mathematical means to convert the time domain to the frequency domain.
Python implementation of Fourier Transforms: Fourier Transforms are a mathematical idea that can separate this signal into its frequency spectrum. The NumPy function “np.fft.fft” is useful to compute a 1-D discrete Fourier Transform. The time domain continuous audio signal as training data is first converted into a discrete Fourier format using the FFT (Fast Fourier Transform) algorithm.
Implementing Fourier Transform to Audio signals with the aim of representing sound to its intensity (decibels (dB)):

Doi : https://doi.org/10.54216/JCHCI.070202
Received: October 28, 2023 Revised: January 24, 2024 Accepted: March 21, 2024

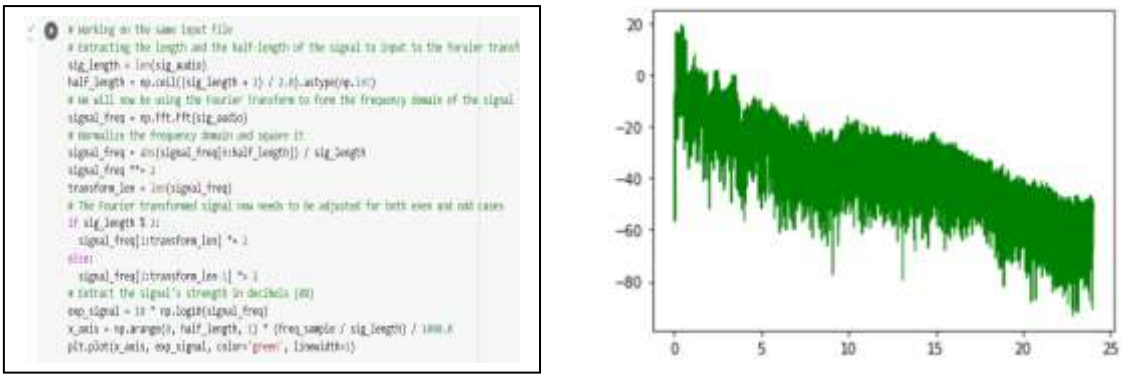


Figure 5: Code snippet and Audio Frequency domain representation

3.3 Audio Signal Features Extraction

The audio files from the film are extracted using the Python module Librosa, which is designed for analyzing audio and music. The load function accepts the visual route as an input and outputs the sample rate and audio stream. Both datasets share the same sampling rate, but the speaker in the RAVDESS dataset waits one second before speaking in that particular dataset. To homogenize the datasets, the silent portions of the signal at the beginning and finish are deleted. Note that when the peak amplitude is 30 dB lower than the maximum peak amplitude, silence is considered.

Data preparation is the process of putting raw data into a format that is logical. Prior to employ machine learning or data mining methods, the data's quality must be ensure.

Some of the features are as follows:

- **Mel scale** - examines how people perceive frequency, is a scale of pitches that listeners perceive to be equally spaced apart from one another.
- **Pitch** - It is the sound's high or low pitch. A higher pitch indicates a higher frequency, depending on the situation.
- **Frequency** - measures wave cycles per second, is the rate at which sound vibrates.
- **Chroma** - an audio representation in which the spectrum is projected onto 12 bins that each stand for one of the twelve unique semitones (or chroma). the result of adding the log frequency magnitude spectrum for all octaves.
- **Fourier transforms** - employed to move between the time and frequency domains. Signal variations throughout time can be seen in the temporal domain. In the frequency domain, the frequency displays the percentage of the signal that is included inside each specific frequency band.

3.4 Features Extraction

Physical features and perceptual features are the two categories of audio features. Physical properties, such as the energy function, spectrum, cepstral coefficients, fundamental frequency, etc., are mathematical measurements that are calculated directly from the sound wave. In terms of how humans perceive sounds, perceptual qualities include things like loudness, brightness, pitch, timbre, rhythm, etc. To categorize the emotions in the audio files, this work employs the Zero-Crossing rate, the Mel spectrogram, the chroma, the delta and delta-delta MFCCs, and the Mel Frequency Cepstral Coefficients. Using Librosa, a 525-length feature vector can be generated for every single speech piece.

3.4.1 Spectrograms

Spectrograms show the signal strength at various frequencies with time. The Mel-spectrogram is extracted from the power spectrum using Mel-Spaced filter banks. By being more discriminative at lower frequencies and less discriminative at higher frequencies, the Mel-scale seeks to emulate the non-linear human ear perception of sound.

The complete spectrum is projected onto 12 bins, which stand in for the 12 distinct semitones of the musical octave, in the Chroma representation of music audio. Analysing the dispersal of chroma, even without the entire frequency (the original octave), can provide valuable musical statistics about the audio and may even capture perceived musical similarity that is not visible in the original spectra because in music, notes exactly one octave apart are perceived as being particularly similar.

Doi : https://doi.org/10.54216/JCHCI.070202
Received: October 28, 2023 Revised: January 24, 2024 Accepted: March 21, 2024

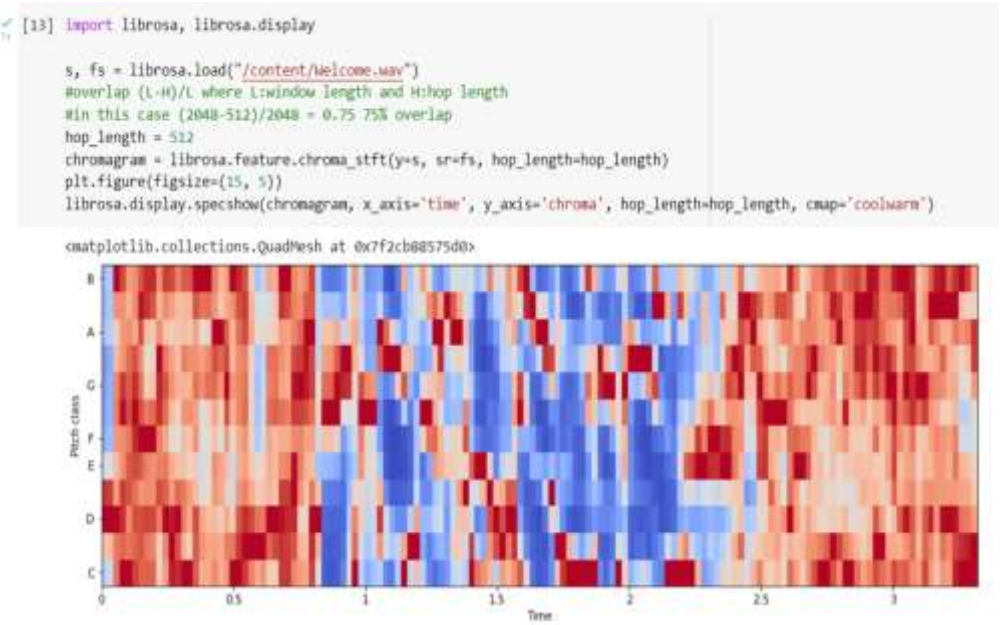


Figure 6: Chromagram

The Mel-frequency cepstrum is very good at recognising audio and modelling the subjective pitch and frequency content of the audio stream. The triangular band pass filter bank filters the FFT power coefficients before computing the Mel-frequency Cepstral Coefficients (MFCCs) (Mel-Spaced filter bank). Differential and acceleration coefficients are other names for the delta and delta-delta MFCCs. Understanding the dynamics of the power spectrum, or the trajectories of MFCCs through time, is the goal of employing differential and acceleration coefficients.

However, the MEL scale was created by the following ways:

- **Voice sound perception in humans**: An adult human's fundamental hearing ranges from 85 Hz to 255 Hz, and this is further gender-specific (85Hz to 180 Hz for Male and 165 Hz to 255 Hz for females). The human ear also processes harmonics that are above these fundamental frequencies.
- **MEL Scale:** Humans evaluate a pitch scale (a scale of audio signals with different pitch levels) based on the equality of their distances. In essence, it is a scale based on how people see things.
- **MEL-spaced Filter bank:** The first step in calculating the power (strength) of each frequency band is to identify the various feature bands that are accessible (done by MFCC). Once these segregations have been formed, we divide the frequencies and separate them using filter banks. The model should be trained with audio signal features that match human hearing properties which will enhance model performance.

The formula for the mapping MEL scale to human actual frequency is as below:

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

where **ln** is log of a base to 'e'

MFCC (Mel-Frequency cepstral coefficients) is a major technique employed to extract the features of the audio signal which is mapped in respective emotions. In this technique, raw audio signals are converted into a compressed illustration that figures out temporal and frequency domain information. By applying inverse discrete Fourier transforms; MFCC model extracts out first 12 coefficients of the signal and apart from these 12 it also picks up the energy of the signal sample as the feature so that there is total 13 features of audio signal by MFCC function which characterise the audio signal. This process helps in to capture the phones.

In addition to these 13 features, the MFCC technique will have the first-order derivative and second-order derivatives of the features which constitute another 26 features which will help in understanding how the transition is occurring Python library functions provide us with straightforward methods for creating filters and applying MFCC to sound.

3.5 Data Preparation

The researcher first formulates data for training. Data must be statistical, the most common example being real values. Categorical data, such as a sex trait with the values "male" and "female", and emotion elements such as "HAPPY",

Doi : https://doi.org/10.54216/JCHCI.070202
Received: October 28, 2023 Revised: January 24, 2024 Accepted: March 21, 2024

"SAD", "ANGRY" and "NEUTRAL" are converted to a real-valued representation by means of "One Hot Encoding". All other features of the audio signal are extracted as numbers through respective transfer functions. The total data set is split into 80:20 ratio where 80% is used for training and 20% for testing purposes.

4. Machine Learning Models

Researchers used a machine learning-based classifier which takes various numerical data as features for training; extracted from the RAVDESS dataset. Broadly, there are two categories of machine learning algorithms 1. Non neural network based 2. Neural network-based (Deep Learning).

In this research, the researchers first experimented with three conventional classifiers SVM, Decision Tree, and Random Forest as type 1 category and then researchers worked with neural network-based Deep learning algorithms as MLP (Multilayer Perceptron) and CNN (Convolution Neural Network).

Training and Testing are two vital processes involved in Machine learning. For training purposes, researchers use 80% of the overall data set while 20% is assigned for testing to evaluate models.

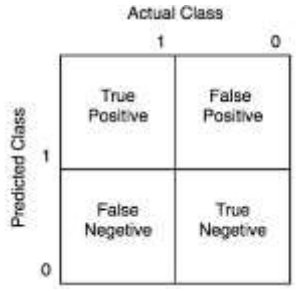
Prediction

Once the model has been trained it can be used to make various predictions. We then make predictions on test data in order to estimate model performance on unseen data, accuracy, and other merit ratios for the model are checked using the testing dataset.

Model Evaluation

In order to detect four emotions based on training, model performance is evaluated using a confusion matrix and relevant various ratios to ensure the usefulness of model.

Confusion Matrix: This four-quadrant matrix expression relationship between the Actual vs Predicted class.



Testing Accuracy
The base metric used for model evaluation is often *Accuracy*, describing the number of correct predictions overall predictions:

$$= \frac{\text{True Positive + True Negative}}{\text{True Positive + True Negative + False Positive + False Negative}}$$

Figure 7: Confusion Matrix

Precision: It is a measure of how many of the positive predictions made are correct (true positives).

Formula: **True Positive / (True positive + False Positive)**

Recall / Sensitivity: It is a measure of how many of the positive cases the classifier correctly predicted, over all the positive cases in the data. Formula: **True Positive / (True positive + False Negative)**

F1-Score: F1-Score is a measure combining both precision and recall, which is basically Harmonic means of two.

$$2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}}$$

All the above merit ratios for model evaluation should be higher and near to 1 is better.

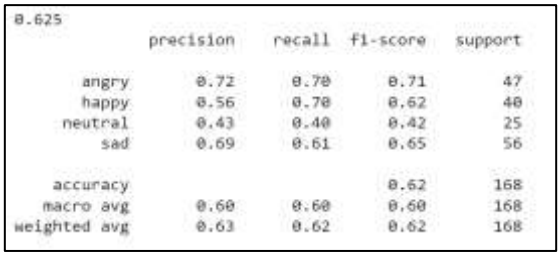
4.1 Non-Deep Learning based Models

4.1.1 Support Vector Algorithm (SVM)

Doi : https://doi.org/10.54216/JCHCI.070202
Received: October 28, 2023 Revised: January 24, 2024 Accepted: March 21, 2024

SVM (Support Vector Machine): The algorithm learns to create the best line or decision boundary that can segregate n-dimensional space into different classes; this boundary is called as "Hyper plane". Then the extreme points/vectors that help in creating the hyperplane, which are called as, support vectors, and hence algorithm is known as the Support Vector Machine.

Results: Overall model accuracy: **62.5%.** Exhibits different merit ratios for different emotions as shown below:



0.625

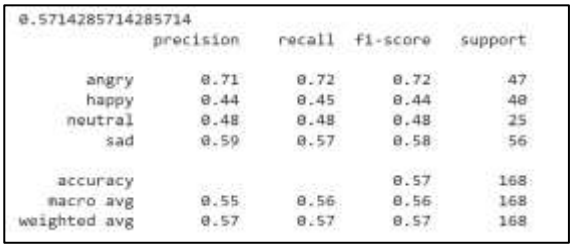
	precision	recall	f1-score	support
angry	0.72	0.70	0.71	47
happy	0.56	0.70	0.62	40
neutral	0.43	0.40	0.42	25
sad	0.69	0.61	0.65	56
accuracy			0.62	168
macro avg	0.60	0.60	0.60	168
weighted avg	0.63	0.62	0.62	168

Figure 8*:* Model performance matrix for SVM

4.1.2 Decision Tree Algorithm (DT)

This model works on information theory; it measures the impurity which is the heterogeneity of data set. The algorithm built a decision tree by iterative dichotomize induction process.

Results: Overall model accuracy: **57.1%.** Exhibits different merit ratios for different emotions as shown below



0.5714285714285714

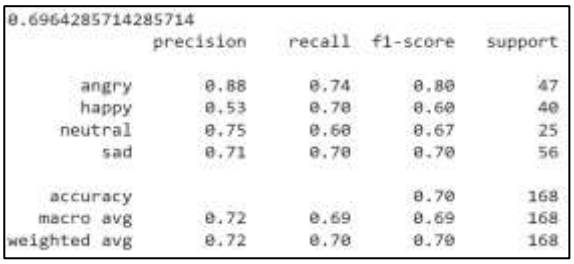
	precision	recall	f1-score	support
angry	0.71	0.72	0.72	47
happy	0.44	0.45	0.44	40
neutral	0.48	0.48	0.48	25
sad	0.59	0.57	0.58	56
accuracy			0.57	168
macro avg	0.55	0.56	0.56	168
weighted avg	0.57	0.57	0.57	168

Figure 9: Model performance matrix for Decision Tree

4.1.3 Random Forest (RF)

This Machine learning uses ensemble learning in which multiple models are built and respective predictive decisions are averaged out to improve the accuracy of the prediction. In fact, it is an extension of a decision tree with subsets in training data.

Results: Overall model accuracy: 69.64%. Exhibits different accuracy for different classes of emotions as shown below



0.6964285714285714

	precision	recall	f1-score	support
angry	0.88	0.74	0.80	47
happy	0.53	0.70	0.60	40
neutral	0.75	0.60	0.67	25
sad	0.71	0.70	0.70	56
accuracy			0.70	168
macro avg	0.72	0.69	0.69	168
weighted avg	0.72	0.70	0.70	168

Figure 10**:** Model performance matrix for Random forest

Results for Model Evaluation of Non-deep Learning Algorithms:

The given graph summarizes the accuracy of each individual emotion being observed by every machine learning model.

Doi : https://doi.org/10.54216/JCHCI.070202
Received: October 28, 2023 Revised: January 24, 2024 Accepted: March 21, 2024

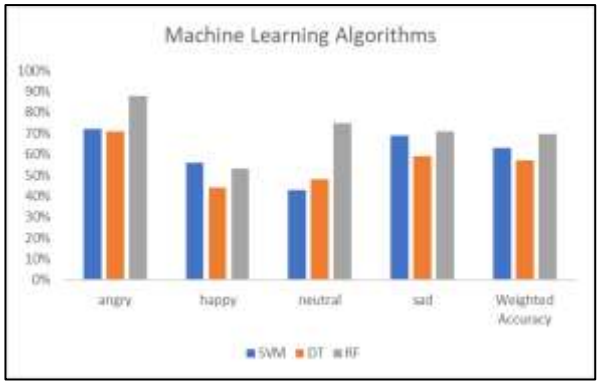


Figure 11: Models relative performance

Observation: From figure, it is clear as the accuracy of various algorithm differ with respect to the type of emotions but overall weighted accuracy indicates the Random Forest gives the best results.

4.2 Deep Learning based Models

4.2.1 ANN Classifier

Artificial Neural Network (ANN): This modelling is resembling of human brain; which is composed of "neurons with activation functions". This network of neurons is organized in the number of layers which are major categories as Input layer, Hidden layer and output layer.

Optimizing Model Performance through Hyperparameter Tuning

After the basic model was ready and the model started giving initial results, the hyperparameters were changed in order to increase the overall performance of the model. This was achieved by changing the number of hidden layers, increasing the number of neurons in each layer, and changing the learning rate algorithms, type of optimizers. Results shows improvisation of the accuracy of the classifier from 63% to 80%. GridSearchCV function is deployed to find best setting of hidden layer size.

Figure 12: Code snippet for accuracy improvement

Five models have been implemented for which Accuracies are as follows:

Doi : https://doi.org/10.54216/JCHCI.070202
Received: October 28, 2023 Revised: January 24, 2024 Accepted: March 21, 2024

Table 1: Comparison if Accuracy for all Models

Classifier Model Type	Accuracy
ANN	**80.21 %**
SVM	62.50 %
RANDOM FOREST	69.64 %
DECISION TREE	57.01 %

5. Summary and Conclusions

Automatic Emotion detection in any conversation at run time can be very useful application in many contexts like recruitment interview, customer support etc. Hence development of such system with practically reasonable accuracy is the objective for this research. Any machine learning model development begins with appropriate dataset. The RAVDESS dataset used in this research work is quite comprehensive and very well composed for development of machine learning based models.

Next task is to convert data set into vital features, which characterises the important distinct aspects of the emotions, which are class identification for our machine learning based classifiers. Hence, in this research major efforts are made to extract maximum features, which are appropriately capturing the emotions type.

Once we are equipped with good features from dataset for the given problem statement next objective is to develop most suitable and accurate performing model. Researchers work around with different algorithms and implemented with python codes using available library functions. Every model is then evaluated for various merit parameters and compare to identify most suitable model. The results reflect as ANN with appropriate hyper parameter setting gives best results and hence it is recommended to deploy for real life applications.

Future improvisation can be done with more variety of datasets and experimenting with various signal features and model parameters to get best performance.

References

[1] Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni, “Speech based Emotion Recognition using Machine Learning”, Institute Of Electrical And Electronics Engineers, Mar. 2019.

[2] Peng Shi, "Speech Emotion Recognition Based on Deep Belief Network", Institute Of Electrical And Electronics Engineers, March 2018.

[3] Ajith Krishna R,Ankit Kumar,Vijay K. "An Automated Optimize Utilization of Water and Crop Monitoring in Agriculture Using IoT." Journal of Cognitive Human-Computer Interaction, Vol. 1, No. 1, 2021 ,PP. 37-45.

[4] J. Uma Maheswari, A. Akila, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN", Institute of Electrical and Electronics Engineers, Feb 2019.

[5] Sri Raksha R. Gupta, M.S. Likitha, A. Upendra Raju and K. Hasitha “Speech Based Human Emotion Recognition Using MFCC”, Institute Of Electrical And Electronics Engineers, March 2017.

[6] Parvesh K,Tharun C,Prakash M. "Apparel Recommendation Engine Using Inverse Document Frequency and Weighted Average Word2vec." Journal of Cognitive Human-Computer Interaction, Vol. 1, No. 2, 2021 ,PP. 46-56.

[7] Tian Kexin, Huang Yongming, Zhang Guobao, Zhang Lin, "Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition", Institute Of Electrical And Electronics Engineers, Nov. 2019.

[8] Ye Sim Ülgen Sonmez, Asaf Varol, "New Trends in Speech Emotion Recognition", Institute of Electrical and Electronics Engineers, June 2019.

[9] T. Giannakopoulos, A. Pikrakis and S. Theodoridis, "A dimensional approach to emotion recognition of speech from movies", IEEE Int'l Conf. Acoustics Speech and Signal Processing (ICASSP), 2009.

Doi : https://doi.org/10.54216/JCHCI.070202
Received: October 28, 2023 Revised: January 24, 2024 Accepted: March 21, 2024

[10] Vijay K. "Collaborating The Textual Reviews Of The Merchandise and Foretelling The Rating Supported Social Sentiment." Journal of Cognitive Human-Computer Interaction, Vol. 1, No. 2, 2021 ,PP. 63 - 72.

[11] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized TV", *IEEE Signal Proc. Magazine*, vol. 23, no. 2, pp. 90-100, 2006.

[12] L. Lu, D. Liu and H.J. Zhang, "Automatic mood detection and tracking of music audio signals", IEEE Trans. on audio speech and language processing, vol. 14, no. 1, pp. 5-18, 2006.

[13] A. Nogueiras, A. Moreno, A. Bonafonte and J.B. Marino, "Speech emotion recognition using hidden Markov models", INTERSPEECH, 2001.

[14] R. Venkatesan ,Althaaf Shaik ,Suraj Kumar,Vipul Guria ,Abhishek Raj. "Intelligent Smart Dustbin System using Internet of Things (IoT) for Health Care." Journal of Cognitive Human-Computer Interaction, Vol. 1, No. 2, 2021 ,PP. 73 - 80.

[15] R. Plutchik and H. Kellerman, "Emotion: theory research and experience" in Theories of emotion, Academic Press, vol. 1, 1980.

[16] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals", IEEE Trans. on Multimedia, vol. 10, no. 5, pp. 936-946, 2008.

[17] P. Kavitha,R. Subha Shini,R. Priya. "An Implementation Of Statistical Feature Algorithms For The Detection Of Brain Tumor." Journal of Cognitive Human-Computer Interaction, Vol. 1, No. 2, 2021 ,PP. 57 - 62.

[18] Ling Cen, Fei Wu, Zhu Liang Yu, Fengye Hu, "A Real-Time Speech Emotion Recognition System and its Application in Online Learning", Emotions, Technology, Design and Learning, ScienceDirect, 2016

[19] Mehmet Cen Sezgin,Bilg Gunsel & Gunes Kurt, Perceptual audio features for emotion detection, EURASIP Journal on Audio,Speech and Music Processing, 2012,Article number 16,Springer.

[20] Renu Taneja;Aman Bhatia;Javesh Monga;Purva Marwaha, Emotion detection of audio files,2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE Publication.

Doi : https://doi.org/10.54216/JCHCI.070202
Received: October 28, 2023 Revised: January 24, 2024 Accepted: March 21, 2024