IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Speech Emotion Recognition using Machine Learning Algorithms

[1]T.Sai Samhith, [2]G.Nishika, [3]M.Prayuktha, [4]M.Bharat Chandra, [5]Dr.Sunil Bhutada, [6]G.Prasadu

[1,2,3,4]B.Tech Student, [5]Professor, [6]Assistant Professor
Department of Information Technology
Sreenidhi Institute of Science and Technology, Hyderabad, India

Abstract: A speech is a verbal action that includes expressing feelings through a person's words and sentences. People use different languages to express their emotions. A person has several emotions in his speech. He tries to tell them while speaking in the form of a speech. We took the emotions into the light: angry, sad, happy, disgust, neutral, surprise, angry, and fear. This paper is about recognizing the emotions of a person from a speech. To identify emotions, We used machine learning algorithms. We considered the classifiers to include random forest, extra trees, gradient boosting, decision tree, light gradient boosting classifiers. We took some datasets, trained them using the classifiers mentioned above, and got the results.

***Index Terms*– Speech, Emotion Recognition, Mel Spectrogram, MFCC, Multi Layer Perceptron Classifier, Random Forest Classifier, Decision Tree Classifier, Light Gradient Boosting Classifier, Extra Trees Classifier.**

I. INTRODUCTION

Speech is a form of expressing our emotions, ideas, and thoughts to people through speaking. Speeches allow us to communicate with people, establish connections with them. It is equally important for the listeners to know what the speech is about and understand the speaker's emotions. A human being can quickly identify the speaker's feelings by listening to his speech carefully and give feedback accordingly.

Multiple accurate speech reinforcements that are practiced on a regular basis on voice-based information. The performance of voice in utilization is significant. The study in a modern report foresees that by 2022, about 12% of all user applications would adequately perform based on voice instructions alone. These voice communications could be bi-directional or mono-directional, and in both illustrations, it is imperative to discern the speech signal. Self-driving cars are 1 such reinforcement that regulates several of its purposes utilizing voice-based management.

The voice-based smart machines are obtaining demand in a large-scale range of utilization. In a voice-based system, a computer agent must thoroughly comprehend the human's speech percept to pick up the commands given to it accurately. This study is called Speech Processing and consists of three segments: Speaker Identification Part, Speech Recognition Part, and Speech Emotion Detection Part.

Speech Emotion Detection is claiming to achieve between the additional parts due to its complexity. Moreover, the representation of a sensible computer system lacks the system to impersonate human response. A fascinating character individual to humans is the capability to reconstruct communications based on the emotional nature of the addresser and the witness. Speech emotion detection built as a classification obstacle determined using numerous MLalgorithms. This report emphasizes more on the Speech Emotion Detection system. Recognizing the emotional nature of the user appears with a significant lead in this application.

Recognizing emergency circumstances in which the user may be incapable of presenting a voice command, the emotion communicated within the user's expression of the decision can adapt to several emergency characteristics of the vehicle. A much more straightforward utilization of speech emotion detection can be observed in call stations, in which automatic voice proposals can be effectively assigned to customer service representatives for additional investigation.

II. LITERATURE SURVEY

In the Paper [1], The authors have implied two alternatives Gaussian mixture models and the other being temporal complexity by using hidden Markov models, these have evolved to be a suitable standard technique for speech processing. They have taken German and English datasets, which were capable to produce the required results. Although same training and testing methods were used, the two solutions differed significantly and there were not any correlations.
In the Paper [2], The authors principally concentrate on three main issues, the features used to characterize different emotions, various classification techniques and required design criteria for speech databases. They have even presented the facts related to combining acoustic, linguistic and other distinctive information. They were successful in developing a few new features as well.
In the Paper [3], The authors have researched and represented thirty-two speech databases in this paper. They have been in terms of their language, number of participants, number of emotions and their purpose. Linear as well as non-linear classification models have been used for this purpose.
In the Paper [4], The authors have presented a emotion recognition system using one to one class in one neural networks. They have used a database of phoneme balanced words, which is speaker and context independent.
In the Paper [5], The authors have proposed Modulation spectral features for efficient results. The features are obtained using both the auditory filter bank and modulation filter bank, which captures acoustic frequency along with temporal modulation frequency components which also conveys the missing information which is not available in conventional short-term spectral features.
In the Paper [6], The authors have fused together facial expressions and speech with other modalities for better overall results. Decision level and feature level integrations are performed in this paper. The results proved that the system based on facial expressions gave more accurate results, than the system based on speech emotions.
In the Paper [7], The authors selected various base features like pitch, MFCC, log energy and added acceleration of pitch as feature stream. These streams were assumed to be one- dimensional signals. This system used SVM and QDA classifications and was able to recognize five emotions.
In the Paper [8], The authors have introduced emotional probability distribution for each speech signal using deep neural networks. These probability distributions are divided into segment levels which are constructed into utterance level features and these are fed into extreme learning machines.
In the Paper [9], The authors have mainly focused on harmony features of speech, in this paper. Hence, the authors have proposed a Fourier parameter model to identify different emotional states, using German database, Chinese language database and Chinese elderly emotional database along with MFCC.
In the Paper [10], The authors have proposed a solution to feature extraction by combining CNN with Long Short-Term Memory networks, using RECOLA database.The authors have clearly proved that this approach towards speech emotion recognition has clearly outperformed existing approaches.
In the Paper [11], This paper mainly focuses on the framework for real time recognition of SSI. The authors also shed light on various other aspects like synchronization and coherent treatment of signals, variability, uncertainty, ambiguity of signals and fusing of multimodal data.
In the Paper [12], The authors have compared the performance of CRNN to CNN. These models were trained using Torgo dataset, which consists of impaired and unimpaired data. The results concluded that recurrent neural networks with a convolution layer would improve the performance for DSR.

III. METHODOLOGY

Existing System:
Given that the possible computational sources were confined, and only a small database of emotionally identified speech examples was available, the initial aim was to be more inclined towards a computationally productive approach that would work with a bit of training data set. These restrictions are pretty standard and may be addressed by the appliance of pre-trained networks and transfer learning. Since the bulk of existing pre-trained networks are designed for image classification, the SER problem had to be re-defined as a picture classification task to use these networks to speech. To realize this, labelled speech samples were buffered into short-time blocks. Every block was measured in a spectrogram array of spectral amplitude and later converted its format into RGB, then passed into a pre-trained Convolutional Neural Network Model. After comprehensive training, now the Convolutional Neural Network Model can be able to judge different emotions. Now Every speech is going through a comprehensive process of Speech to Visualized Picture conversion. Within the experiments presented here, the SER performance was tested using two different sampling frequencies (sixteen and eight kHz) and, therefore, the μ-low companding procedure.

Proposed System:
The speech emotion detection system is performed as a Machine Learning (ML) model. The steps of operation are similar to any other ML project, with supplementary fine- tuning systems to make the model function adequately. The fundamental action is data collection, which is of prime importance. The model being generated will acquire from the data contributed to it and all the conclusions and decisions that a progressed model will produce is supervised data. The secondary action, called as feature engineering, is a combination of various machine learning assignments that are performed over the gathered data. These systems approach the various data description and data quality problems. The third step is often explored the essence of an ML project where an algorithmic based prototype is generated. This model uses an ML algorithm to determine about the data and instruct itself to react to any new data it is exhibited to. The ultimate step is to estimate the functioning of the built model. Very frequently, developers replicate the steps of generating a model and estimating it to analyze the performance of various algorithms. Measuring outcomes help to choose the suitable ML algorithm most appropriate to the predicament.

Dataset

English Language Dataset, namely the Torronto Emotional Speech Set (TESS) was taken into consideration. This is a dataset which consists of 200 target words and were spoken by two women, one younger and the other older, and the phrases were recorded in order to portray the following seven emotions happy, sad, angry, disgust, fear, surprise and neutral state. There are 2800 files in total, in this dataset. The phrase spoken by them was "Say the word ___". Both women have thresholds within normal range and the audio quality is very high. The authors are Kate Dupuis and M. Kathleen Pichora Fuller.\

IV. SYSTEM DESIGN

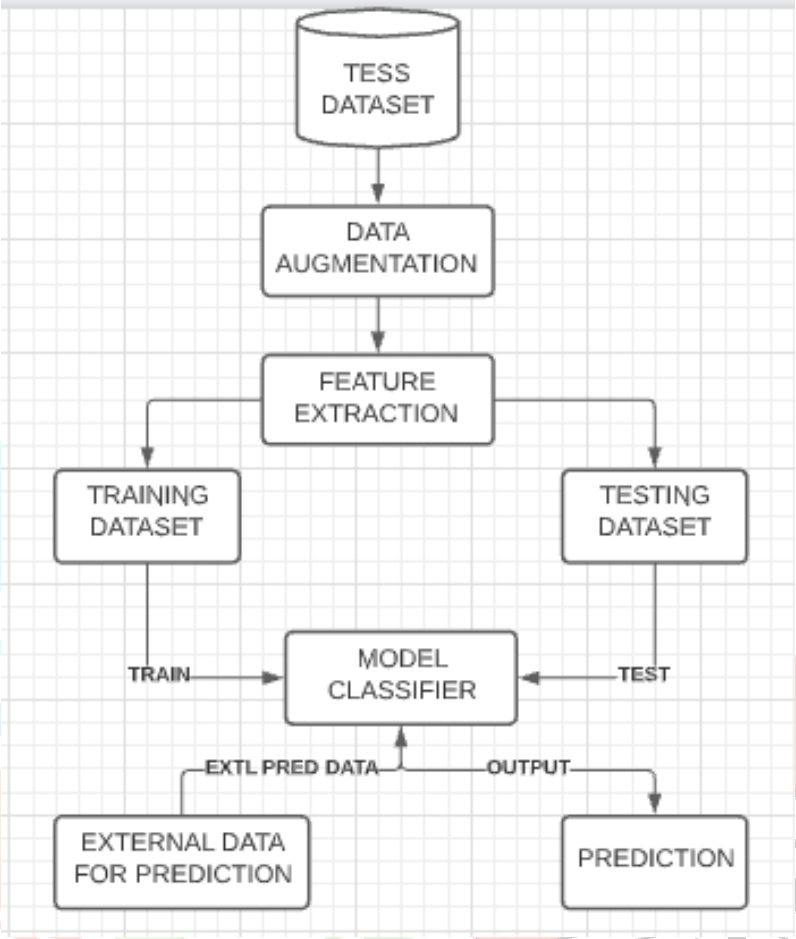

Figure 1) Architecture of Emotion Recognition from TESS Dataset Using various Machine Learning Algorithms

MODULES:

Data Set & Data Visualization:

In this Speech Emotion Recognition Project, Audio File is taken from the TESS Dataset, and that will be uploaded in .wav file format before the file upload process is validated, which relates to the file format and empty file input, and will be connected directly to python files where the output generated in the form of Emotional Labels. Data visualisation provides information about the given audio data in graphic and pictorial form. Here, initial dataset is divided into its emotional labels, and then the whole data is pictured into spectrogram graph and wave plot diagrams(like in fig 3,4). Spectrogram may be a visual representation of the spectrum of frequencies of a sign because it varies with time. Wave-plot is employed to plot waveform of amplitude vs time where the primary axis is an amplitude and the second axis is time.

Figure 2) Bargraph mentions the no of emotions for each label vs Label

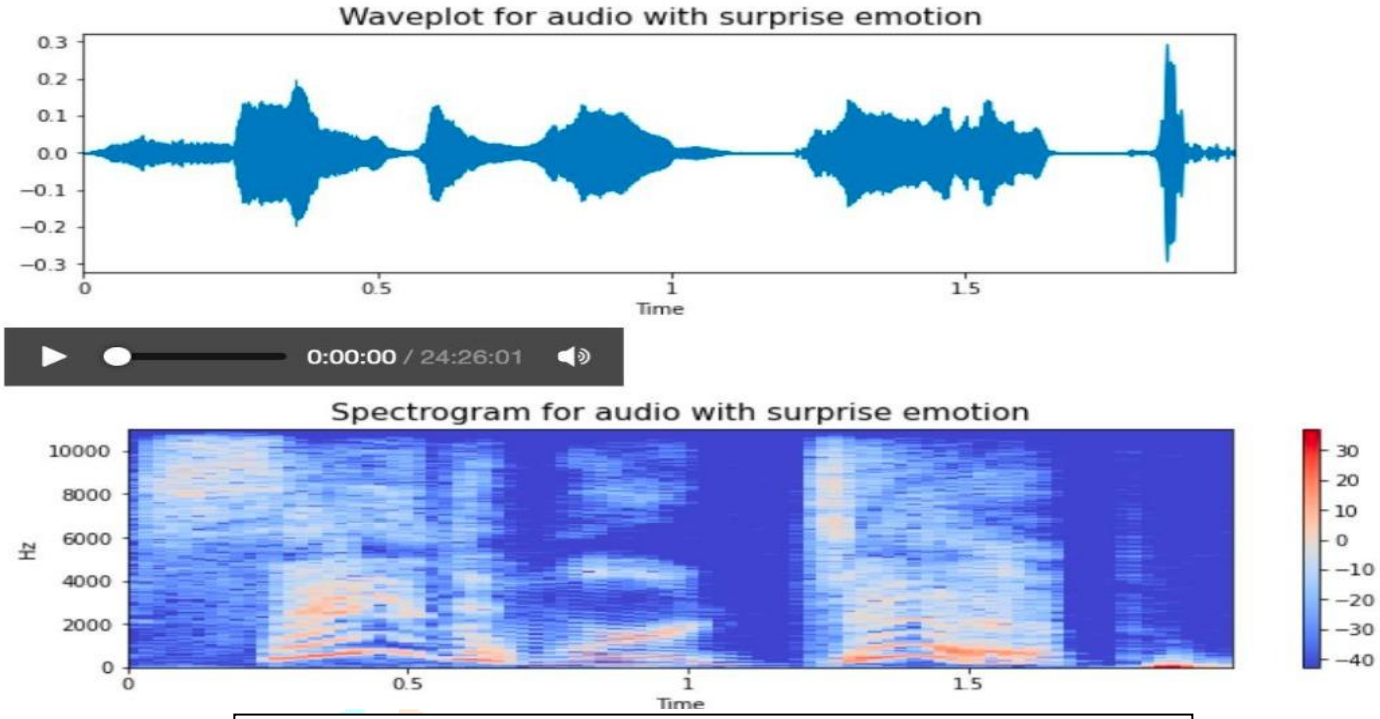

Figure 3) Waveplotand Spectrogram Diagrams for Surprise Emotion

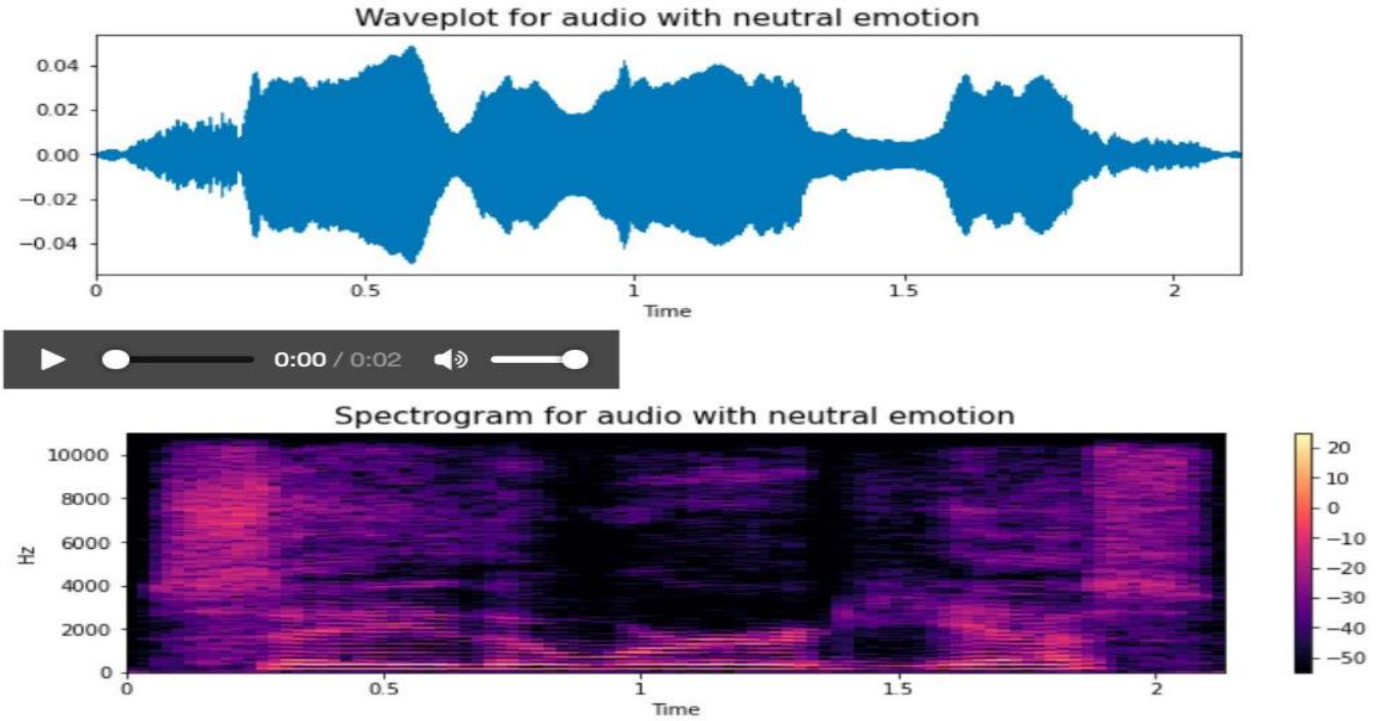

Figure 4) Waveplotand Spectrogram Diagrams for Neutral Emotion

Data Augmentation:

In this mainly focuses on disturbing data where initial taken will be more tuned without noise, But in a real-life scenario, that is not the case where we can have more noisy parts in the original recorded audio file. So, we thought of adding some more data by just taking the given data and run through two augmentation techniques like adding extra noise to the data and normal data values at the same time for every data we collected by keeping the same emotional label to it.

FeaturesSelection:

In this Project mainly comprises features like Zero Crossing rate, Chroma Shift, Root Mean Square Value, Mel Spectrogram and MFCC(Mel Frequency Cepstral Coefficient). These are few predominantly used audio features for emotional-related audio content, acoustic recognition, and information retrieval in the music category; these extract the crux of given audio. The zero-crossing rate is when a significant change from +ve to 0 to -ve or from -ve to 0 to +ve. Mel Spectrograms are spectrograms that visualise sounds on the Mel scale as against the frequency domain. The Mel Scale must be a logarithmic transformation of a signal's frequency.

V. DIFFERENT ALGORITHMS USED FOR TRAINING MODEL

Multi Layer Perceptron Classifier (MLPC):

A multilayer perceptron (MLP) is a group of artificial neural networks (ANN). MLP is utilized vaguely, seldom loosely to each feedforward ANN, seldom rigidly related to systems constituted of repetitious layers of perceptron's (with threshold activation). Multilayer perceptron's are seldom commonly introduced to as "vanilla" neural networks, peculiarly while they have a mainly single hidden layer. Multilayer perceptron's present a supervised classification technique for multi-band passive optical remote sensing data. Multilayer perceptrons (MLP) offer robust classifiers that may produce superior performance corresponded with additional classifiers, but MLP Classifiers are frequently scrutinized for the abundance of unconstrained parameters. Moreover, complexities with MLPs include prolonged training intervals and local minima.

Light Gradient Boosting Machine(LGBM) :

Light Gradient Boosted Machine, or LightGBM for brief, is an open-source python library which gives an adequate and effective implementation of the gradient boosting algorithm. LightGBM prolongs the gradient boosting algorithm by appending various automatic feature selections, including boosting examples including more strong gradients. Thus, LightGBM can happen in a moving speedup of training and enhanced predictive performance. For LightGBM, Gradient-based one-side sampling (GOSS) is managed to distinguish the views adopted for computing the division. GOSS retains those situations with extensive gradients, including only irregularly abandons those occurrences with slight gradients to preserve the efficiency of information gain estimation. This procedure can drive to a further particular gain estimation than consistently random sampling, with the equivalent target sampling rate, particularly when the value of information gain has a widespread range.

Gradient Boosting Classifier (GB) :

Gradient boosting is an ensemble classifier determined to execute properly in the setting where the abundance of variables surpasses the number of samples like high-dimensional data. However, it has not been estimated for the prediction of unique events. It is illustrated that Gradient boosting experiences from rare critical events bias, accurately classifying a small proportion of specimens from the rare species barely. The bias can be excluded by utilizing subsampling in unification with a suitable quantity of shrinkage but solely for a particular amount of boosting repetitions and for binomial loss function. It is shown that the amount of boosting iterations where the unique events bias is excluded but cannot be evaluated efficiently from the training data when the data size is tiny. Consequently, various improvements for the unique events bias of Gradient boosting are recommended and estimated by using fabricated and real high-dimensional data.

Extra Trees Classifier (ET):

Extra Trees Classifier is a process of ensemble learning that aggregates the consequences of numerous de-correlated decision trees (DTs) incorporated inside a "forest" to exhibit classification effects. It is pretty analogous in performance to the random forest and alters essentially from this in envisioning the DTs inside the forest. That DT is organized from the primary training set essentially in Extra Trees Forest. Consequently, about each testing node, an erratic collection of k-features is assigned to an individual tree from the feature set on which each DT must choose the crucial features to divide the statistics based on specific numerical parameters (Gini Index).

This arbitrary adoption of peculiarities leads to diverse, de-correlated DTs being developed. During the construction of the forest, for every feature, the normalized total reduction in the analytical standards adopted in the split feature judgment (Gini Index if the Gini Index is used in the construction of the forest) is determined to achieve feature preference utilizing this identical forest arrangement. This attribute is termed the feature's Gini significance. For the adoption of features, every feature is regulated in deteriorating order as per the Gini Value of each feature, and the user chooses the best k features based on preference.

Random Forest Classifier (RF) :

The philosophy behind classifier ensembles is based on superimposed the primary assumption that a collection of classifiers do achieve more reliable classifications than a single classifier does. Breiman was the one who proposed a unique and assuring classifier (2001) called random forest. Random Forest is a tree-based ML algorithm that leverages the potential of various decision trees for delivering decisions. Forest of arbitrarily generated decision trees. Every node in the decision tree operates on an arbitrary subset of characteristics to measure the output. The random forest then merges the number of unique decision trees to produce the ultimate output.

Decision Tree Classifier (DT):

The decision tree models are perceived to be the most beneficial in the area of data mining, data science, and machine learning, considering they achieve consistent accuracy and decision tree are comparatively inexpensive to estimate. Most utmost decision-tree classifiers present classification in two phases: Tree Building Part and Tree Pruning Part. In tree building part, the decision tree model is constructed by recursively dividing the training data set based on a sectionally optimal pattern until all or most of the works pertaining to several of the hindrances sustain the same class label. To enhance generalization of a decision tree, tree pruning is applied to prune the leaves and branches accountable for analysis of individual or relatively more minor data vectors. The decision tree classifier (DTC) is an example of the desirable ways to multistage judgment making. The basic idea included in several multistage approaches is to divide up a complex judgment into a combination of individual, more simplistic decisions, anticipating the final solution achieved; this approach would agree with the proposed aspired solution.

Difference between Random Forest Classifier (RF) and Decision Tree Classifier (DT) :

The decision tree as is indicated by the titles "Tree" and "Forest," a Random Forest is a combination of Decision Trees. A decision tree is constructed on a whole dataset, utilizing every feature of interest, whereas a random forest randomly decides rows and features to establish various decision trees and then averages the results.

Adopting a decision tree model on a provided training dataset, the accuracy improves with more divisions. Nevertheless, one can undoubtedly overfit the data and do not understand when someone has crisscrossed the line unless cross-validation (on training data set) is done. On the other hand, the benefit of a simplistic decision tree is model is simple to perform, and anyone who is working should know whatever variable also whatever value of that variable is utilized to split the data and foretell a result.

Random Forest has a more prolonged training period than a single decision tree. It would be great if one acknowledged this because as we progress the abundance of trees in a random forest, the period is taken to train them also improves. Notwithstanding vulnerability and yoke on a particular assortment of features, decision trees are advantageous because they are more manageable to evaluate and quicker to train.

VI. EVALUATION & RESULTS

As we all know that its challenging task to predict the emotions of the speaker in the real time. After Performing Training to the Models, few tests were carried out in order to evaluate the accuracy and performance of each model. Here, few testing techniques like Accuracy, F1 Score, MCC, Recall, Precision, Kappa are used to determine the performance of each and every model mentioned in Table 1.

Model	Accuracy	F1 Score	Kappa	MCC
Light Gradient Boosting Machine	0.9714	0.9714	0.9666	0.9667
Random Forest Classifier	0.9585	0.9585	0.9516	0.9517
Extra Trees Classifier	0.9479	0.9478	0.9392	0.9393
Gradient Boosting Classifier	0.9448	0.9449	0.9355	0.9357
Multi Layer Perceptron Classifier	0.924844	0.925	0.912259	0.912966
Decision Tree Classifier	0.8831	0.883	0.8636	0.8638

Table 1) Performance of Different Algorithms

Accuracy:

The Accuracy is generally calculated with Test data in our project, where Accuracy relates to number of correct classifications made to total number of predicted classifications made as mentioned in Figure 5.

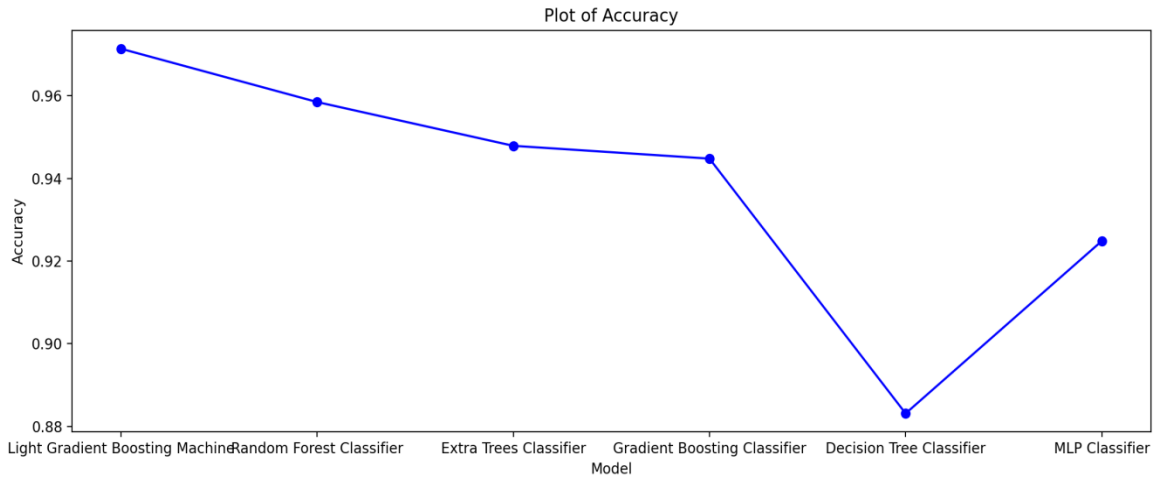

Figure 5) Plot showing the Accuracy values for Each Model

F1 Score :

The F1 Score is needed when you want to seek a balance between Precision and Recall.F1-score, is a measure of a model's accuracy on a dataset, F1 score as mentioned in Figure 6.

Unusual benefits of F1-score: Very minute precision or recall will appear in more bad overall score. Consequently, it improves balance the two metrics. If you accept your positive class as the 1 with several specimens, F1-score can improve support the metric crossed positive/negative samples.

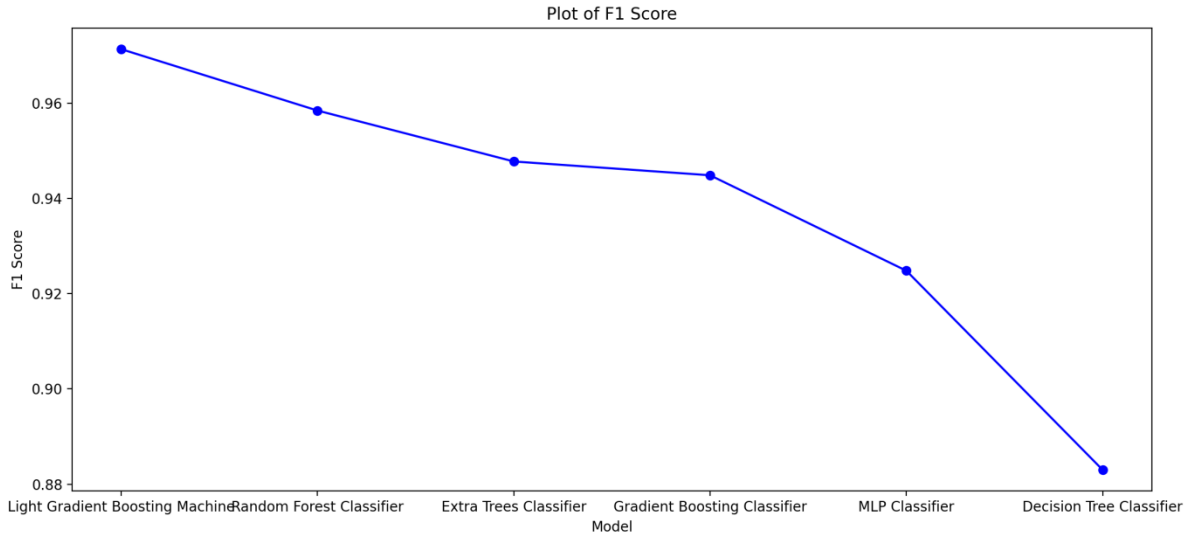

Figure 6) Plot showing the F1 Score values for Each Model

Cohen's Kappa Co-efficient:

The Cohen's Kappa is a statistic that is utilized to estimate inter-rater dependability for qualitative things. It is ordinarily estimated to be a extra robust pattern than uncomplicated percent adjustment computation, as κ demands into description the probability of the compromise transpiring by uncertainty. Kappa values are as mentioned in Figure 7.

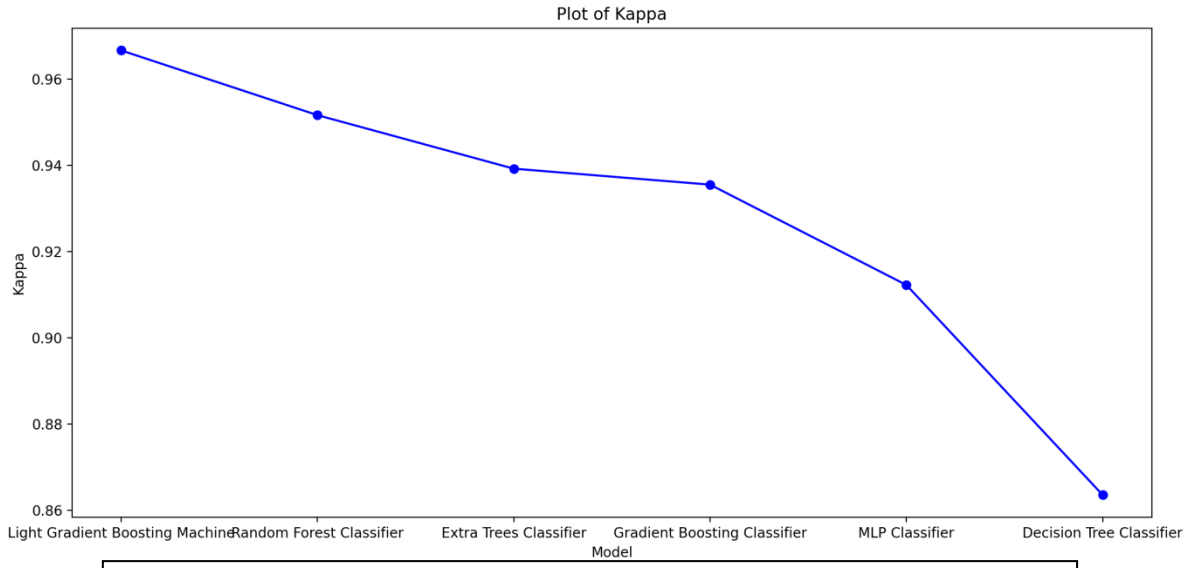

Figure 7) Plot showing the Kappa values for Each Model

Matthews Correlation Co-efficient:

The coefficient demands in description true and false positives and negatives, including frequently observed as a proportionate measure that can be utilized even if that classes are of quite varied sizes. In Postulate, the MCC is a correlation coefficient linking the perceived and predicted two classifications; it delivers a value inserted −1 and +1. A coefficient of +1 describes a perfect prognostication, 0 no adequately than arbitrary prediction and −1 indicates cumulative disagreement among prediction and observation. Nonetheless, if MCC agrees neither −1, 0, or +1, it is not a particular pointer of how comparable a predictor is to arbitrary inference because MCC is conditioned on the dataset. MCC values are as mentioned in Figure 8.

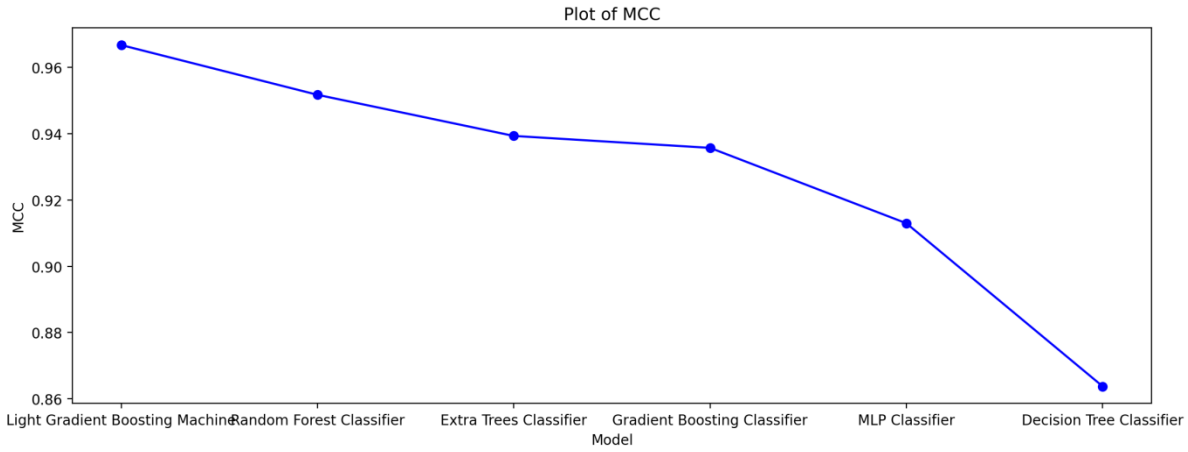

Figure 8) Plot showing the MCC values for Each Model

VII. Conclusion

The TESS dataset that we considered is fine-tuned. Since it has noiseless data, it was easy for us to classify and feature the data. The classifier with utmost accuracy, AUC, F1 score, kappa, MCC is Random Forest Classifier. The classifier with the least accuracy and all the other terms is the decision tree classifier. We also mentioned why the decision tree classifier has low precision and the random forest classifier has high accuracy. The other classifiers that we analyzed, i.e., extra trees classifier, light gradient boosting machine, multi perceptron classifier, gradient boosting classifier, have the mid values in accuracy and other values. In conclusion, the random forest classifier is the most accurate algorithm and can be used in real-life scenarios to detect a person's emotions through his speech.

VIII. Future Scope

The dataset that we examined is a small one. So the algorithms that we used won't be practical. There might be chances of overfitting and underfitting. When we extend the dataset and dig the deep neural networks, they are high chances for these algorithms to be efficient. Talking about the features that we considered, if we take few more features into account, we can get high accuracy. We should also consider many more neural networks for the model to be accurate. Two or more datasets can also be combined and used as input data to make the model accurate and efficient. This model that we created only has noiseless data. In the real world, the data has null values, missing data, and outliers. If we can work on the preprocessing, the model can be used in real-life scenarios.

References

[1] B. Schuller, G. Rigoll and M. Lang, "Hidden Markov model-based speech emotion recognition," 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., 2003, pp. II-1, doi: 10.1109/ICASSP.2003.1202279.

[2] Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition, Volume 44, Issue 3, 2011, Pages 572-587, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2010.09.020..

[3] Koolagudi, S.G., Rao, K.S. Emotion recognition from speech: a review. Int J Speech Technol 15, 99–117 (2012).https://doi.org/10.1007/s10772-011-9125-1.

[4] Nicholson, J., Takahashi, K. &Nakatsu, R. Emotion Recognition in Speech Using Neural Networks. NCA 9, 290–296 (2000). https://doi.org/10.1007/s005210070006.

[5] Siqing Wu, Tiago H. Falk, Wai-Yip Chan, Automatic speech emotion recognition using modulation spectral features, Speech Communication, Volume 53, Issue 5, 2011, Pages 768-785, ISSN 0167-6393,https://doi.org/10.1016/j.specom.2010.08.013.

[6] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of the 6th international conference on Multimodal interfaces (ICMI '04). Association for Computing Machinery, New York, NY, USA, 205–211. DOI:https://doi.org/10.1145/1027933.1027968

[7] Kwon, Oh-Wook / Chan, Kwokleung / Hao, Jiucang / Lee, Te-Won (2003): "Emotion recognition by speech signals", In EUROSPEECH-2003, 125-128.

[8] Han, Kun / Yu, Dong / Tashev, Ivan (2014): "Speech emotion recognition using deep neural network and extreme learning machine", In INTERSPEECH-2014, 223-227.

[9] K. Wang, N. An, B. N. Li, Y. Zhang and L. Li, "Speech Emotion Recognition Using Fourier Parameters," in IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 69-75, 1 Jan.-March 2015, doi: 10.1109/TAFFC.2015.2392101.

[10] G.Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5200-5204, doi: 10.1109/ICASSP.2016.7472669.

[11] Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André. 2013. The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In Proceedings of the 21st ACM international conference on Multimedia. Association for Computing Machinery, New York, NY, USA, 831–834. DOI:https://doi.org/10.1145/2502081.2502223

[12]Albaqshi, Hussain, and AlaaSagheer. "Dysarthric Speech Recognition using Convolutional Recurrent Neural Networks.",doi: http://www.inass.org/2020/2020123134.pdf