Human Emotion Detection From Speech

Natasha Lalwani
Drexel University
Philadelphia, PA, USA
nl498@drexel.edu

Nishchala Barde
Drexel University
Philadelphia, PA, USA
nb932@drexel.edu

***Abstract*—Over the past few years a lot of research has gone into human emotion detection from speech or Speech Emotion Recognition (SER). Numerous systems have been designed and tested in order to detect the speaker's emotion. In this paper, we explore different Machine Learning models like KNN Classifier, Support Vector Machines, Logistic Regression Classifier, Random Forest Classifier, in order to classify seven basic emotions; neutral, calm, happy, sad, angry, fearful, disgust, surprised (as seen in $fig.1$). Dataset used is RAVDESS. We made use of $librosa$ to extract various features such as MFCCs, chromagram, melspectrogram, spectral contrast and tonnetz. The idea of ensemble learning was also implemented with an attempt to increase the accuracy.**

***Index Terms*—KNN Classifier, Support Vector Machines, Logistic Regression Classifier, Random Forest Classifier, Feature Extraction, Ensemble Learning**

I. Introduction

Recognition of human emotion has always intrigued data scientists. Emotions play an essential role in our daily conversations. A lot of research in recent years aimed at developing reliable emotion identification systems based on a variety of data sources, including audio and video. One of the easiest methods to deduce a person's emotional condition is to look at their facial expressions. Another option is to use speech as a modality. Emotion analysis using solely audio data is difficult due to the lack of visible visual information of human faces. Speech is a complex signal that comprises information about the message, the speaker, the language used, and the emotions expressed. Speech signals can usually be collected more easily and inexpensively than many other biological signals (such as electrocardiogram) and hence it has been gaining popularity. Speech can express a variety of emotions. Emotional speech recognition is a system that determines a person's emotional state based on his or her voice; speech even for humans, can be deceiving when it comes to judging the speaker's emotion [1].

When working with audio files, especially those including human recordings and speeches, Mel Frequency Cepstral Coefficients (MFCC) is a critical component to be considered. Any sound generated by humans is determined by the shape of their vocal tract (including tongue, teeth, etc.). If this shape can be determined correctly, any sound produced can be accurately represented. The envelope of the speech signal's time power spectrum represents the vocal tract, and MFCC accurately represents this envelope [2]. Other features like chromagram, melspectrogram, spectral contrast, tonnetz are extracted as well from the audio so that the emotion can be determined efficiently. In order to be able to do this, there one such package in python known as Librosa. Not only does it serve as a necessary building elements for the creation of music information retrieval systems but also aids in the visualization of audio signals as well as feature extraction utilizing various signal processing techniques.

Application: People's emotions can be detected in a variety of settings, including robot interfaces, audio surveillance, web-based E-learning, commercial applications, clinical studies, entertainment, banking, call centers, cardboard systems, computer games, and so on. Information regarding students' emotional states can be used to focus on improving teaching quality in classroom orchestration or E-learning. A teacher, for example, can utilize SER to choose which subjects can be taught and must be able to build emotional management skills in the classroom [3].

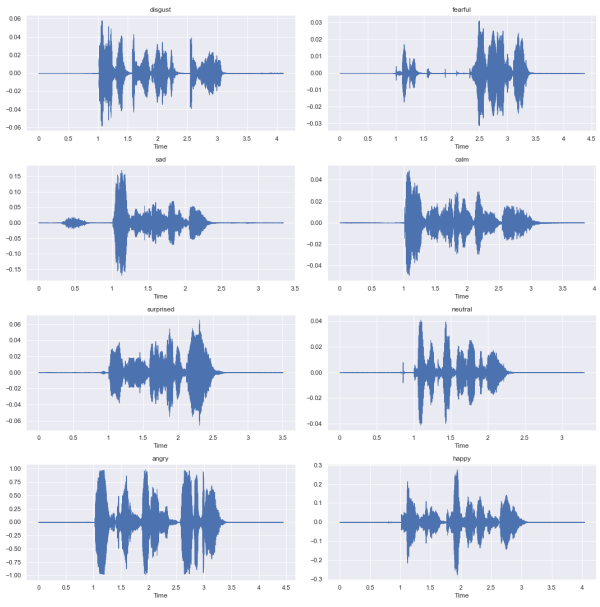

Fig. 1. Waveform of each emotion (one audio file per emotion)

II. Related Work

This section primarily covers work related to Human Emotion Detection from Speech. Frank Dellaert et al. [4] created their own dataset consisting of 1000 utterences from various speakers. Four emotions were detected from their dataset; happy, sad, anger, fear. 17 selected features from 5 groups were used and three methods were implemented- MLB Classifier, KR and KNN. Maximum accuracy was obtained from KNN. M.K. Sarker et al. [5] recognized four human emotions by employing four machine learning techniques along with majority voting techniques over features of Berlin dataset and EMA dataset. Majority voting technique resulted in better accuracy as compared to the machine learning models. Lingli Yu et al. [6] implemented an improved version of SVM, feature driven hierarchical SVM, in order to speech emotion recognition performance. Dataset used by them was Chinese speaker-dependent and Berlin speaker-independent speech corpus. 50 sentences for each emotion was chosen (30 training and 20 testing). They further proposed ways to improve the experiment. Yu C. et al. [7] adopted a method based on back propagation neural network. They worked on a Chinese emotional database (CASIA). Initially a set of five traditional features (E FOD, F0 mean, MFCC mean, ZCR FOD, LPC mean) are used for speech emotion recognition. Then, based on feature selection, two new characteristics of speech emotion (MFCCF0 and APSAM) were added to existing set of five features. Addition of new characteristics increased the accuracy. Md. Zia Uddin et al. [8] proposed a novel approach for emotion recognition from audio speech signals where MFCC features were tried with discriminant analysis and neural structured learning based on neural graph learning and adversarial learning. The proposed approach was compared with traditional approaches where it produces better results. S. Basu et al. [9] presented recent works pertaining to affect detection using speech and issues related to affect detection. Significance of various models like KNN, GMM, HMM, MLP, RNN, was also presented along with a feature extraction technique name Mel Frequency Cepstral Coefficiants. J. Umamaheswari et al. [10] designed an emotion recogntion system using a hybrid of PRNN and KNN. They identified six emotions: neutral, anger, happiness, sadness, surprise and fear. They implemeted a cascaded system of MFCC and GLCM for feature extraction along with a Wiener filter for filtering the noise in speech. The hybrid model produced great results as compared to other conventional methods. S. Lalitha et al. [11] recognized seven emotions using pitch and prosody features. Berlin emotional database was classified using Support Vector Machine (SVM) classifier for which a recognition rate of about 81% was obtained. M. W. Bhatti et al. [12] identifies and extracted potential prosodic features from speech data. They proposed a systematic feature selection strategy involving the use of Sequential Forward Selection (SFS) with a General Regression Neural Network (GRNN) in conjunction with a consistency-based selection method. The selected features were fed as input to a Modular Neural Network (MNN), which prduced satisfying results. E. Ramdinmawii et al. [13] conduct analysis of emotion states using features, namely, instantaneous fundamental frequency using Zero Frequency Filtering, Formant frequencies (F1, F2, F3), signal energy, and dominant frequencies. In order to cross validate the results, two databases (German and Telugu Emotion Databases) we used. Significant difference was observed. A. Iqbal et al. [14] extracted 34 audio features including MFCCs, energy, spectral entropy, etc. from RAVDESS and SAVEE database. Gradient Boosting models were used to classify emotions in their system. Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) are two more classifiers that were used to assess their accuracy on audio recordings. E. Lieskovska et al. [15] compared various forms of recurrent networks, such as LSTM, LSTM with peephole connections, GRU. They worked on IEMOCAP database and presented accuracy for the same. LSTM produced better results as compared to LSTM with peephole connections and GRU. A. Sonawane et al. [16] performed extensive experiment on happy, anger, sad, disgust, surprise and neutral emotion sound database. MFCC was used for extraction of features and multiple support vector machines was used as a classifier. Further it was revealed that non-linear kernel SVM achieved greater accuracy than linear SVM. G. Deshmukh et al. [17] classified three emotions using 3 features vectors: Pitch, MFCC and Short Term Energy, which were then sent to an SVM classifier model. Q. Zhang et al. [18] looked at how the quantities and statistical values of speech features affect emotion recognition accuracy. They used Gaussian Mixture Model (GMM) to extract two useful features from speech signals: Mel Frequency Cepstral Coefficients (MFCCs) and Auto Correlation Function Coefficients (ACFC), and undertook experiments with the Berlin emotional database using a GMM supervector created by the values of MFCCs, delta MFCCs, and ACFC, and six proposed emotions: anger, disgust, fear, happy, neutral, and sad.

III. Methodology

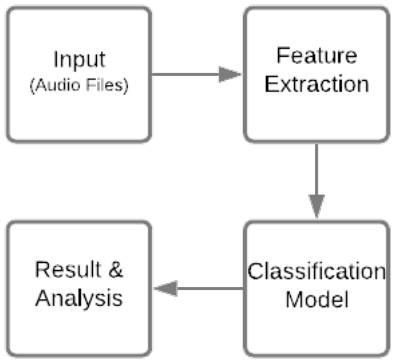

Fig. 2. Block Diagram illustrating basic outline of human emotion detection from speech

A. Feature Extraction

The extraction of features is a critical step in evaluating and discovering relationships between various objects. Because the audio data provided by the models cannot be directly comprehended by the models, feature extraction is utilized to turn it into a format that can be understood. It's a method that explains the majority of the facts in a straightforward manner. Classification, prediction, and recommendation algorithms all require feature extraction [19]. Librosa, a python package, is used to extract features. The extracted feature for our work are MFCCs, chroma, melspectrogram, spectral contrast and tonal centroids (tonnetz). We also consider mid-term features (calculates mean and standard deviation over short-term features.)

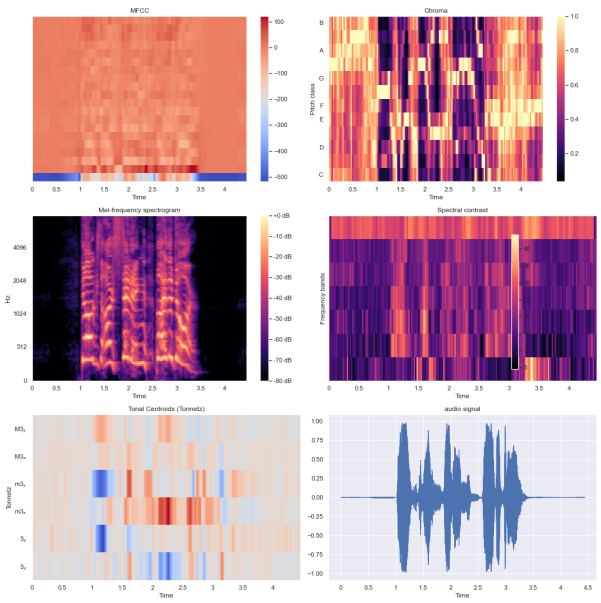

Fig. 3. Extracted Features for a single audio file.

B. Normalizing Features

Normalization is the process of transforming data so that it is either dimensionless or has similar distributions. In any machine learning application or model fitting, normalization is a critical step in data pre-processing as the accuracy hugely depends on it. Normalization provides each variable equal weights/importance, ensuring that no single variable biases model performance in one direction simply because it is larger. The normalized data is then used to train the system.

C. Machine Learning Models

1) Support Vector Machine (SVM): Support Vector Machine, or SVM, is a linear model that can be used to solve classification and regression issues. It can solve both linear and nonlinear problems and is useful for a wide range of applications. SVM is a basic concept: The method divides the data into classes by creating a line or hyperplane. SVM technique is used to find the points from both classes that are closest to the line. These points are knows as support vectors. The distance between the line and the support vectors is now computed. The distance calculated is known as margin. Our goal is to maximize the margin. The ideal hyperplane is the one for which the margin is the maximum. As a result, SVM seeks to create a decision boundary with as much separation between the two classes as possible. Important parameters of SVM are C and $Gamma$. C is in charge of the trade-off between a smooth decision border and correctly classifying training points. A large C number indicates that more correct training points will be received. Gamma specifies the extent to which a single training example has an impact. If it has a low value, it signifies that each point has a long reach, whereas if it has a high value, it suggests that each point has a short reach. We trained the model using optimal values.

2) Logistic Regression: It's a predictive technique similar to Linear Regression but with a difference that it uses independent factors to predict the dependent variable, but the dependent variable must be categorical.It is a statistical model that uses logistic function in order to model the conditional probability. It can be used for both, binary classification as well as multi-class classification. For binary classification issues, which are problems with two class values, such as predictions like "this or that," "yes or no,", "A or B," logistic regression is one of the most widely used machine learning techniques. The goal of logistic regression is to estimate event probabilities, which includes establishing a relationship between variables and the likelihood of specific outcomes.

3) K-Nearest Neighbors : It is a simple, easy-to-implement technique that may be used to address both classification and regression issues. However, in the industry, it is mostly utilized to solve classification and prediction problems. KNN is a lazy learning algorithm because it does not have a dedicated training phase and instead uses all of the data for training and classification. Because it makes no assumptions about the underlying data, KNN is a non-parametric learning algorithm [23]. This algorithm predicts the values of new data points based on 'feature similarity', which implies that the new data point will be assigned a value depending on how closely it matches the points in the training set. The 'k' in the KNN algorithm is based on feature similarity. The process of selecting the appropriate value for K is known as parameter tuning, and it is critical for improved accuracy.

4) Random Forest Classifier: Random forest is a supervised learning technique that can be used to classify and predict data. However, it is mostly employed to solve

classification problems. A forest, as we all know, is made up of trees, and more trees equals a more healthy forest. Similarly, the random forest method constructs decision trees from data samples, extracts predictions from each, and then votes on the best option. It's an ensemble method that's superior than a single decision tree because it averages the results to reduce over-fitting. Random forests perform better than a single decision tree for a wide range of data items. The variance of a random forest is lower than that of a single decision tree. Random forests are incredibly adaptable and have a high level of accuracy. The random forest algorithm does not require data scaling. Even after supplying data without scaling, it maintains a high level of accuracy. Even when a major amount of the data is missing, the Random Forest algorithms maintain high accuracy [24].

D. Ensemble Learning Model

Ensemble techniques are a type of machine learning methodology that integrates numerous base models to create a single best-fit predictive model. There are 2 types of ensemble methods: Bagging and Boosting. The word "bagging" comes from the fact that it combines Bootstrapping and Aggregation into a single ensemble model. Multiple bootstrapped sub-samples are taken from a sample of data. On each of the bootstrapped sub-samples, a Decision Tree is created. An algorithm is used to aggregate across the Decision Trees to build the most efficient predictor once each sub-sample Decision Tree has been formed. Boosting is an ensemble learning strategy for reducing training errors by combining a group of weak learners into a strong learner. In boosting, a random sample of data is chosen, fitted with a model, and then trained sequentially—that is, each model attempts to compensate for the shortcomings of the one before it. The weak rules from each classifier are joined with each iteration to generate a single, strong prediction rule. A Voting Classifier is a machine learning model that learns from an ensemble of many models and predicts an output (class) based on the highest probability of the output being the chosen class. It simply sums up the results of each classifier that has been submitted into Voting Classifier and predicts the output class based on the highest majority of votes. Rather than constructing separate dedicated models and determining their accuracy, we create a single model that is trained by multiple models and predicts output based on the cumulative majority of voting for each output class. Two types of voting are supported by Voting Classifier: Hard Voting and Soft Voting. The predicted output class in hard voting is the class that receives the greatest number of votes, i.e. the class that had the highest likelihood of being predicted by each of the classifiers. The output class in soft voting is the prediction based on the average probability given to that class [26].

IV. Results

For each of the following models, we've printed the confusion matrix as well as classification report:

1) Support Vector Machine
2) Logistic Regression
3) KNN Classifier
4) Random Forest
5) Ensemble Learning Model

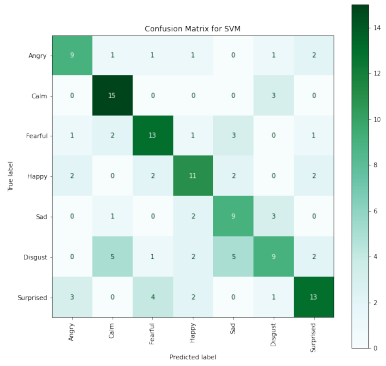

Fig. 4. SVM Confusion Matrix

	precision	recall	f1-score	support
0	0.60	0.60	0.60	15
1	0.62	0.83	0.71	18
2	0.62	0.62	0.62	21
3	0.58	0.58	0.58	19
4	0.47	0.60	0.53	15
5	0.53	0.38	0.44	24
6	0.65	0.57	0.60	23
accuracy			0.59	135
macro avg	0.58	0.60	0.58	135
weighted avg	0.59	0.59	0.58	135

Fig. 5. SVM Classification Report

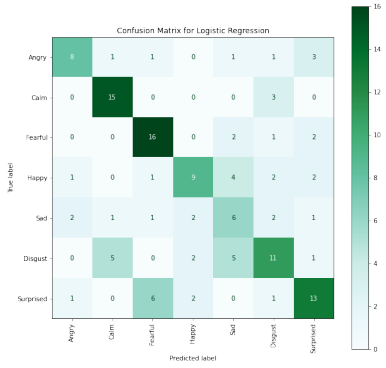

Fig. 6. Logistic Regression Confusion Matrix

	precision	recall	f1-score	support
0	0.67	0.53	0.59	15
1	0.68	0.83	0.75	18
2	0.64	0.76	0.70	21
3	0.60	0.47	0.53	19
4	0.33	0.40	0.36	15
5	0.52	0.46	0.49	24
6	0.59	0.57	0.58	23
accuracy			0.58	135
macro avg	0.58	0.58	0.57	135
weighted avg	0.58	0.58	0.57	135

Fig. 7. Logistic Regression Classification Report

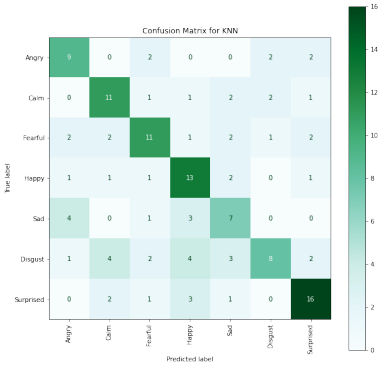

Fig. 8. KNN Confusion Matrix

	precision	recall	f1-score	support
0	0.53	0.60	0.56	15
1	0.55	0.61	0.58	18
2	0.58	0.52	0.55	21
3	0.52	0.68	0.59	19
4	0.41	0.47	0.44	15
5	0.62	0.33	0.43	24
6	0.67	0.70	0.68	23
accuracy			0.56	135
macro avg	0.55	0.56	0.55	135
weighted avg	0.56	0.56	0.55	135

Fig. 9. KNN Classification Report

V. Conclusion

The objective of this project was to compare different machine learning models in terms of classifying the human speech from given audio data set. We also explored the concept of ensemble learning to see if the accuracy improved as compared to the traditional models. On comparing individual models highest accuracy was obtained from SVM model; 59%. For our ensemble learning model, we combined the results of SVM, Logistic Regression model and KNN model, and got an accuracy of 60%. As seen, there wasn't much improvement in the accuracy. Better results can be obtained by conducting a more through hyper-parameter tuning, by implementing a more complex model, etc.

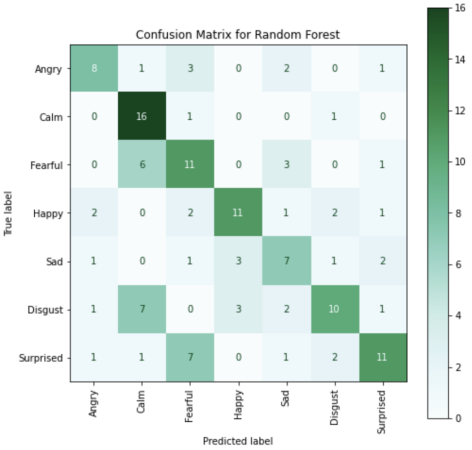

Fig. 10. Random Forest Confusion Matrix

	precision	recall	f1-score	support
0	0.62	0.53	0.57	15
1	0.52	0.89	0.65	18
2	0.44	0.52	0.48	21
3	0.65	0.58	0.61	19
4	0.44	0.47	0.45	15
5	0.62	0.42	0.50	24
6	0.65	0.48	0.55	23
accuracy			0.55	135
macro avg	0.56	0.56	0.55	135
weighted avg	0.57	0.55	0.54	135

Fig. 11. Random Forest Classification Report

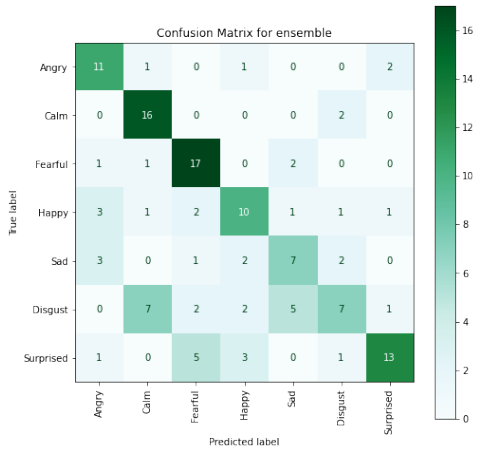

Fig. 12. Ensemble Confusion Matrix

VI. Future Scope

The impact of features on enhancing or lowering accuracy is significant. More features could be extracted before training the model. Deep learning is currently the most used machine

	precision	recall	f1-score	support
0	0.58	0.73	0.65	15
1	0.62	0.89	0.73	18
2	0.63	0.81	0.71	21
3	0.56	0.53	0.54	19
4	0.47	0.47	0.47	15
5	0.54	0.29	0.38	24
6	0.76	0.57	0.65	23
accuracy			0.60	135
macro avg	0.59	0.61	0.59	135
weighted avg	0.60	0.60	0.58	135

Fig. 13. Ensemble Classification Report

learning algorithm. The next stage will be to use deep learning to construct a more accurate system of emotion recognition from speech.

REFERENCES

[1] S. Lalitha, A. Madhavan, B. Bhushan and S. Saketh, "Speech emotion recognition," 2014 International Conference on Advances in Electronics Computers and Communications, 2014, pp. 1-4, doi: 10.1109/ICAECC.2014.7002390.

[2] https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/

[3] https://www.intechopen.com/chapters/65993

[4] F. Dellaert, T. Polzin and A. Waibel, "Recognizing emotion in speech," Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, 1996, pp. 1970-1973 vol.3, doi: 10.1109/ICSLP.1996.608022.

[5] M. K. Sarker, K. M. R. Alam and M. Arifuzzaman, "Emotion recognition from speech based on relevant feature and majority voting," 2014 International Conference on Informatics, Electronics and Vision (ICIEV), 2014, pp. 1-5, doi: 10.1109/ICIEV.2014.6850685.

[6] Lingli Yu, Binglu Wu, Tao Gong; September 2–6, 2013. "A hierarchical support vector machine based on feature-driven method for speech emotion recognition." Proceedings of the ECAL 2013: The Twelfth European Conference on Artificial Life. ECAL 2013: The Twelfth European Conference on Artificial Life. Sicily, Italy. (pp. pp. 901-908). ASME.

[7] Yu, C. , Xie, L. and Hu, W. (2016) Feature Optimization of Speech Emotion Recognition. Journal of Biomedical Science and Engineering, 9, 37-43. doi: 10.4236/jbise.2016.910B005.

[8] Md. Zia Uddin, Erik G. Nilsson, Emotion recognition using speech and neural structured learning to facilitate edge intelligence, Engineering Applications of Artificial Intelligence, Volume 94, 2020, 103775, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2020.103775.

[9] S. Basu, J. Chakraborty, A. Bag and M. Aftabuddin, "A review on emotion recognition using speech," 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), 2017, pp. 109-114, doi: 10.1109/ICICCT.2017.7975169.

[10] J. Umamaheswari and A. Akila, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 177-183, doi: 10.1109/COMITCon.2019.8862221.

[11] S. Lalitha, A. Madhavan, B. Bhushan and S. Saketh, "Speech emotion recognition," 2014 International Conference on Advances in Electronics Computers and Communications, 2014, pp. 1-4, doi: 10.1109/ICAECC.2014.7002390.

[12] M. W. Bhatti, Yongjin Wang and Ling Guan, "A neural network approach for human emotion recognition in speech," 2004 IEEE International Symposium on Circuits and Systems (ISCAS), 2004, pp. II-181, doi: 10.1109/ISCAS.2004.1329238.

[13] E. Ramdinmawii, A. Mohanta and V. K. Mittal, "Emotion recognition from speech signal," TENCON 2017 - 2017 IEEE Region 10 Conference, 2017, pp. 1562-1567, doi: 10.1109/TENCON.2017.8228105.

[14] A. Iqbal and K. Barua, "A Real-time Emotion Recognition from Speech using Gradient Boosting," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1-5, doi: 10.1109/ECACE.2019.8679271.

[15] E. Lieskovska, M. Jakubec and R. Jarina, "Speech Emotion Recognition Overview and Experimental Results," 2020 18th International Conference on Emerging eLearning Technologies and Applications (ICETA), 2020, pp. 388-393, doi: 10.1109/ICETA51985.2020.9379218.

[16] A. Sonawane, M. U. Inamdar and K. B. Bhangale, "Sound based human emotion recognition using MFCC and multiple SVM," 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), 2017, pp. 1-4, doi: 10.1109/ICOMICON.2017.8279046.

[17] G. Deshmukh, A. Gaonkar, G. Golwalkar and S. Kulkarni, "Speech based Emotion Recognition using Machine Learning," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 812-817, doi: 10.1109/ICCMC.2019.8819858.

[18] Q. Zhang, N. An, K. Wang, F. Ren and L. Li, "Speech emotion recognition using combination of features," 2013 Fourth International Conference on Intelligent Control and Information Processing (ICICIP), 2013, pp. 523-528, doi: 10.1109/ICICIP.2013.6568131.

[19] https://towardsdatascience.com/extract-features-of-music-75a3f9bc265d

[20] https://towardsdatascience.com/data-normalization-in-machine-learning-395fdec69d02

[21] https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

[22] https://towardsdatascience.com/quick-and-easy-explanation-of-logistics-regression-709df5cc3f1e

[23] https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm

[24] https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm

[25] https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f

[26] https://www.geeksforgeeks.org/ml-voting-classifier-using-sklearn/

[27] P. Shen, Z. Changjun and X. Chen, "Automatic Speech Emotion Recognition using Support Vector Machine," Proceedings of 2011 International Conference on Electronic and Mechanical Engineering and Information Technology, 2011, pp. 621-625, doi: 10.1109/EMEIT.2011.6023178.

[28] https://iopscience.iop.org/article/10.1088/1742-6596/1921/1/012017/pdf

[29] https://librosa.org/doc/main/tutorial.html

[30] https://medium.com/@anonyomous.ut.grad.student/building-an-audio-classifier-f7c4603aa989

[31] https://github.com/markadivalerio/audio-classifier-project

[32] https://github.com/rbarzegar93/Speech-Emotion-Recognition

[33] https://jovian.ai/kuntal-das/emotional-speech-classification2d-resnet

[34] https://www.datacamp.com/community/tutorials/convolutional-neural-networks-python

[35] https://towardsdatascience.com/ensemble-learning-using-scikit-learn-85c4531ff86a