TECHNIQUES FOR THE RECOGNITION OF VARIOUS VOICE ISSUES: NEUROLOGICAL, FUNCTIONAL, AND LARYNGEAL DISEASES

Ms. Tejasvini C L[*1], Dr. Padmaja Devi G[*2]

[*1]Department Of Electronic & Communication, Malnad College Of Engineering Hassan, Karnataka, India.

[*2]Associate Professor, Department Of Electronic & Communication, Malnad College Of Engineering Hassan, Karnataka, India.

ABSTRACT

This is basically based on evaluation of multiple speech disorders. The methodology discussed here is based on Artificial Neural Network with Machine Learning Algorithm. Other classifiers which are considered are Decision Tree and Random Forest which will give a better idea with the classification accuracy for voice impairment. In this paper few voice disorders considered are Laryngeal, Parkinson, Cleft lip and Normal. The dataset taken of patients speaks the vowels basically. Vowels are considered because they give certain variations in the voice which makes the prediction of disorder easy for the machine when trained and tested. Once the data is taken first, they are pre-processed i.e., the noise which are available in the samples are removed where there is very minimum loss of data. Once this is done feature extraction of the samples are carried out under frequency and time domain. Frequency domain extraction are done mainly for the voice or word extraction from the sample using the fundamental frequency. MFCC, LPC, WPD are the extraction done in Frequency domain whereas CA, Jitter, Shimmer and others are carried under Time domain. These methods utilized ranges its accuracy from 86.6% in the automatic detection of voice impairments. The study and results obtained from this suggests that neural networks algorithms give better accuracy then other classifiers utilized.

Keywords: Artificial Neural Network (Ann), Mel-Frequency Cepstral Analysis (Mfcc), Wavelet Packet Decomposition (Wpd), Cepstral Analysis (Ca), Jitter, Shimmer.

I. INTRODUCTION

A voice disease is characterized by means of the atypical production and/or absences of vocal exceptional, pitch, loudness, resonance, and/or period, that is irrelevant for a character's age and/or sex. The pathological problems like vocal nodules, edema, vocal fold paralysis, and neurological troubles like mind tumor, lesions, neural degeneration, mind damage may also affect the speech generating portion of the mind. About the hidden records of disorders of frightened system includes in voice. The efficiency of movement in addition to the potential to speak honestly is reduced via the issues. Tumors, mind damage, growth of lesions, neural degeneration within the cells of mind which controls the speech conversation will have an effect on the speech manufacturing.

The development of computer-aided techniques that enable objective voice assessment is one of the goals of pathological speech processing. Spectral and cepstral modelling, perturbation measurements, noise content measures, prosodic aspects, and nonlinear behaviour are all investigated in the literature [1]. In order to achieve high recognition rates, all of these measures are often combined in the same representation space; nevertheless, the interpretation of those results is not entirely obvious. This paper examines voice registers from six databases with recordings of sustained phonations of speakers with pathologies of three different origins: laryngeal, functional, and neurological, with the goal of improving the interpretation and analysis of different voice pathologies using different characterization methods. Patients with laryngeal pathologies (LP) experience throaty voice, wheezing, and irregular movement of the vocal cords as a result of the presence of polyps [2]. Hypernasality, on the other hand, is the most prevalent characteristic in the voice of individuals with cleft lip and palate and is one of the vocal disorders having functional roots (CLP). This condition causes the patient's voice to be nasalized excessively, as a result of the velum's improper regulation, resulting in inappropriate resonances in the vocal and nasal cavities [3]. Impaired laryngeal function has been clinically observed in these patients [4]. Regarding the neurological disorders, PD is one of the most common. The voice of PD patients is characterized among others, by excess of tremor, reduced loudness, monotonicity, hoarseness [5,6].

II. METHODOLOGY

The system design process follows in this manner. The required libraries are added then required folders are selected. The functions which are essential for training and testing purposes are called. The major implementation part of the recommended system consists of two essential source codes: training and testing, i.e. The data is split into two parts. Once after the data splitting the features generation takes place i.e., features are extracted. After feature extraction the ANN model is created. The data recorded according to the batch size specified in the training section, and the different accuracies are calculated and recorded. The Classifiers is assigned to display the values. The outcome is predicted, and a Confusion Matrix is presented. After, we determine whether the output is Laryngeal, Cleft Lip, Normal, or Parkinson.

Few Libraries utilized in this technique are:

- **SpeechPy library:** It includes speech preparation approaches, speech features, and key post-processing activities. It contains the most often used speech characteristics, such as MFCC and filter-bank energies, as well as the log-energy of filter-banks. The goal is to give academics a basic tool for extracting and processing speech features.
- **SciPy Library**: It is a free library used for scientific computing and technical computing.
- **Sklearn Library:** It includes classification, regression, clustering, and dimensionality reduction as well as other useful machine-learning and statistical modelling techniques.
- **Pywt library:** It is a wavelet transform library for Python that is open source. It merges a simple high-level interface with C and Cython performance at the bottom level. This is quite simple to use.

III. MODELING AND ANALYSIS

The proposed methodology consists of four phases. The dataset taken of patients speaks the vowels basically. Vowels are considered because they give certain variations in the voice which makes the prediction of disorder easy for the machine when trained and tested.

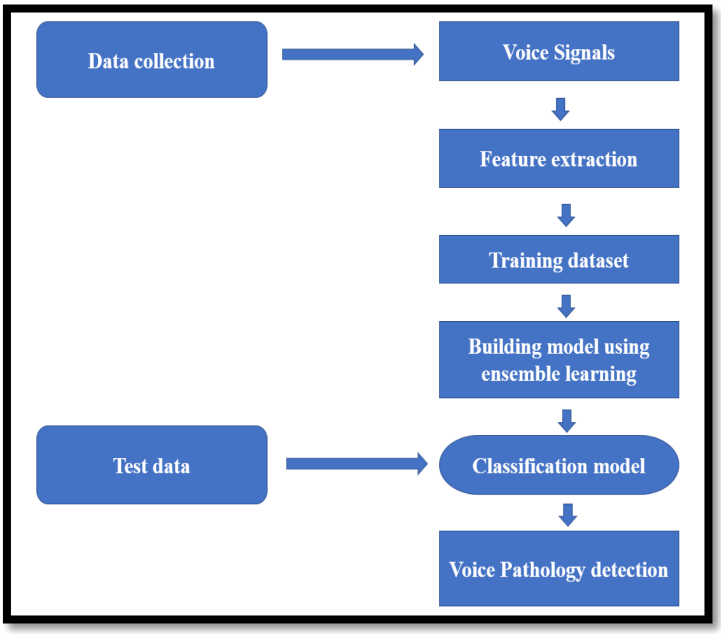

Fig 1: Steps involved

1. Data Collection

A session is described as a collection of recordings of vowels /a/, /i/, /u/ made at regular, high, low, and low-high-low pitch. The length of the audio clips with sustained vowels is a pair of seconds. All recordings are captured at 50kHz and have a 16-bit resolution. Files with sustained vowels and participants between the ages of 30 and 35 were used in these trials. Both male and female voices are taken into account. The data collection phase is the first step. These four diverse pathological voices, such as Cleft Lip, Laryngeal Pathologies, Parkinson, and Normal, are collected and recorded in a database.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

- **Cleft Lip:** 'Grupo de Procesamiento y Reconocimiento de Seales (GPRS) from the Universidad Nacional de Colombia, branch Manizales', compiled the database. This database contains the phonations of the five Spanish vowels uttered by kids with reconditioned-Cleft lip. A total of 127 recordings have been gathered.
- **Laryngeal pathologies**: The Voice & Speech Lab at the Massachusetts Eye and Ear Infirmary (MEEI) collected these 71 recordings. The audio recordings in this database were captured at various sampling rates. All recordings made during this investigation for self-addressed experiments were down-sampled to 25 kHz with a resolution of sixteen bits.
- **Parkinson's disease:** The PC-GITA database's recordings are taken into account. It includes 62 recordings of native Spanish and German speakers speaking, all of whom are roughly balanced in age and gender.
- **Normal:** The PC-GITA database's recordings are taken into account. It includes 79 recordings of native Spanish and German speakers, all of whom are nearly equal in age and gender.

2. Pre-Processing and Characterization

The data are examined using Hamming windows of different lengths based on an estimated parameter. Each extracted features are given with a descriptive of the length and time shift. Several characteristics are retrieved from the speech frames when the windowing procedure is completed.

a. Noise Measures: The existence of glottal sound in the signal throughout speech production is determined by the presence of glottal noise in the signal due to an inadequate closure of the vocal folds. To undertake a comprehensive description of the noise content in speech signals, a set of six parameters is determined. Hamming windows of 40 ms length and a 20 ms time shift are used to window the voice recordings. HNR, which is a ratio of the signal's harmonic energy to its noise content, is one of the measures shown in the set.

b. Periodicity and Stability of Voice: These features correspond to the capacity to maintain a steady airflow during continuous vowel production. Two distinct windowing lengths are utilized in this feature set. For one set of measurements, 40 ms windows with a 20 ms temporal shift are utilised. This category encompasses variations in pitch period amplitude (shimmer) and cycle-to-cycle pitch period fluctuation (jitter). For the second set of measurements, 150-millisecond windows with a 75-millisecond temporal shift are employed.

c. Spectral-Cepstral modeling: The purpose of modelling the speech spectrum in the spectral or cepstral domains is to evaluate the speaker's capacity to create periodic movements of the vocal folds, i.e., with a large number of harmonic components this feature is referred to as "spectral wealth." This work includes spectral and cepstral domain features with the goal of simulating variations in the speech spectrum, particularly around the first two formants (F1 and F2), where the majority of the signal's energy is focused. The characteristics are calculated in 40-millisecond windows with a 20-millisecond time shift.

d. Nonlinear Behavior: Nonlinear pressure-flow in the glottis, nonlinear stress-strain curves in vocal fold tissues, and nonlinearities in the vocal fold collision are all examples of nonlinear processes in voice signals. Nonlinear behaviour in the speech signal might be caused by compensatory movements in various muscles and limbs involved in the speech production process. This phenomenon occurs when a speaker knows he is speaking in an incorrect manner and attempts to remedy the "errors." Windowing with a duration of 55 ms and a temporal shift of 27.5 ms is used to estimate nonlinear attributes.

3. Feature Extraction

Feature extraction is accomplished by converting the speech waveform to a parametric representation that can be processed and analysed at a lower data rate in the future. This is known as front-end signal processing. It transforms the processed speech signal into a logical, compact representation that is more discriminatory and reliable than the original.

A. Frequency Domain Extraction

The Frequency Domain is a region in which signals are expressed in frequency rather than time. The extracted features are classified as global and local.

Global Features: The global characteristics are retrieved from the full spoken utterance, whereas the local characteristics are retrieved in frames from each voice sample.

a. Mel-Frequency Cepstral Coefficients (MFCC)

The MFCC computation is a mimic of the human hearing system that aims to artificially duplicate the ear's working principle under the assumption that the human ear works in the same way it does in nature.
The most significant advantage of Mel-Frequency Cepstral Coefficients is it employs "Mel Frequency Scaling," that is particularly relevant to the human sensory framework. With the use of an algorithmic rule, the coefficients produced a superb example of sign spectra with good knowledge compacting. A frame size of 20 milliseconds is employed, with a frame shift of 10 milliseconds. Mel-frequency is a Swedish word that means "melodic frequency."

b. Wavelet Packet Decomposition

The wavelet packet methodology could be viewed as an extension of wavelet decomposition, allowing for a broader range of signal evaluation alternatives and the easiest sign matching assessment. By translating the level of a signal from the time to the frequency domain, it provides degree. It's calculated by repeating filter-decimation methods, which reduces time resolution while increasing frequency resolution. The frequency packing containers are identical in size to the wavelet transform since the WPT separates not only the lower, but also the high frequency sub band. Each detail coefficient vector is separated into two pieces using a mechanism similar to wavelet packet decomposition's approximation vector splitting.

The wavelet transform divides the voice signal into sub-bands, with approximation components holding sign characteristics and high frequency components linked to noise and transmission disruption. While removing the high frequencies preserves the signal's characteristics, it may occasionally have useful properties.

c. Cepstral Analysis (CA)

It is one among the most effective approaches for determining whether a signal includes periodic components. The method has also been used to determine the pitch of a signal. The cepstrum analysis was used to determine the logarithm spectrum's power spectrum. Its original application was to detect echoes in seismic waves, where it outperformed the autocorrelation function due to its aversion to signal colour. Because the effects of vocal excitation (pitch) and vocal tract (formants) are cumulative in the logarithm of the power spectrum and hence easily separable, Cepstrum pitch determination is extremely successful.

B. Time Domain Extraction

Local Features: Jitter, shimmer, pitch, intensity, pulse, and harmonicity are some of the local features recovered with the ML tool in this work.

a. Jitter

From one sequence to the next, the speaker's voice frequency will shift. Frequency perturbation or vocal interference = random quantity variability. Vocal interference is the cause of husky, harsh, or scratchy voice quality in people with vocal disorders. Interference could be used as a metric for vocal stability, a parameter used to detect vocal flaws.

The cycle-to-cycle change of harmonic is measured by the noise measurements area unit noise (absolute). The average absolute difference between consecutive periods, divided by the common amount, is jitter (relative). The Relative Average Perturbation (rap) is defined as the average absolute difference between a period and the average of its two neighbours, divided by the average period. Jitter (ppq5) is the five-point amount Perturbation Quotient, which is calculated as the average absolute difference between a number and its four nearest neighbours, divided by the common number.

$$\boxed{jitter(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \times 100}$$

Where, Ti - the retrieved glottal period lengths and N - the number of extracted glottal periods.

b. Shimmer

“Shimmer” and “frequency perturbation” are the same thing. The term "shimmer" is similar to the term "amplitude." The degree of amplitude perturbation, also known as vocal shimmer, is a measure of voice stability. Excessive shimmer gives the impression of huskiness. Shimmer measures are measured in decibels and are expressed as the variation of the height to peak amplitude.

The average absolute difference between the amplitudes of consecutive periods, divided by the normal amplitude, is defined as "Shimmer" (relative).

$$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20 * \log\left(\frac{A_{i+1}}{A_i}\right)\right|$$

Where, Ai - retrieved peak-to-peak amplitude data and N - The number of retrieved fundamental frequency periods.

c. Pitch

Pitch is a feature that measures sound as highs and lows inside the perception of voice. Pitch is an initiative feature of sounds that is a frequency-related scale. It's divided into a number of frames that are evenly spaced in time. Pitch is decided inside the speech transmission to distinguish it from noise, with frequency being clear and steady.

d. Intensity

The intensity of sound pressure is measured in decibels (loudness). There are two possibilities when it comes to intensity. The discrepancy between the quantity of flow from the lungs and the amount of resistance from the vocal folds is one example.

e. Harmonicity

The existence of signals superimposed on the elemental signal, whose frequencies are whole number numbers of the fundamental frequency, is what harmony is. The presence of harmonicity in the voltage or current wave shape generates a distorted signal for voltage or current, as well as a non-sinusoidal signal that causes load problems or injury.

Extraction method was carried out and feature vectors were formed. Class label 0 is allocated to healthy speech signals, whereas class label 1 is allocated to pathological voice signals. To construct a training dataset, feature vectors were re-created. To generate models with varied properties and to design an efficient speech pathology detection model, several types of information are collected, and four separate training datasets are created

4. Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANN) mimics the anatomy of the brain. The human brain is simply a sponge that absorbs information from its surroundings. Some issues that unit of measurement on so many aspects the scope of modern computers unit of measurement typically unit of measurement only resolvable by energy efficient packages is a confirmed truth. In the case of machine resolution, this form of brain modelling also gives a less technical route. They essentially consist of a large number of simple method units connected to an efficient communication network. Each simple method unit is a real somatic cell that sends out a replacement signal or fires if it gets a strong signal from the associated unit of choice.

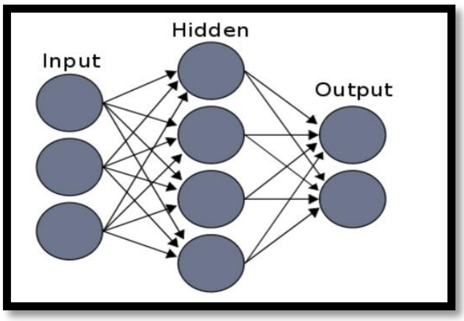

Fig 2: ANN Model

The back-propagation rule has now become popular in the new Multi-Layer Perceptrons class of layered feedforward networks. At least two layers of perceptrons are frequently present. The structure is made up of one input layer, one or more hidden layers, and output layers. The hidden layer performs an essential role as a feature extractor. It uses a nonlinear operation like sigmoid or radial-basis to generate sophisticated input

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

functions. The output of the output layer is based on the information it gets from the hidden layer to reduce classification error.

5. Classification

Gaussian Kernel Support Vector Machine

They are algorithms which are utilized for both classification and regression. In SVM, each data is plotted as a point in space with respect to features. The feature values of each are the particular value of coordinate. After which classification will be performed and found the hyper plane which will differentiates the class of two.

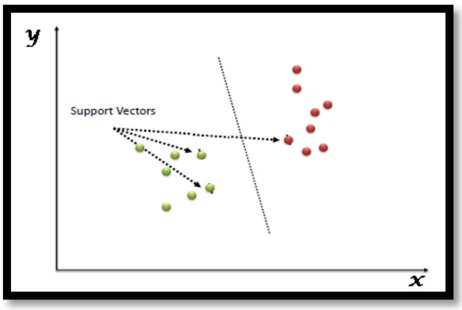

Fig 3: SVM MODEL

SVM Classifiers

This is a Binary Linear classification that is built for the purpose of minimising error. Its flexible Machine Learning Model can do both linear and non-linear classifications. It can also detect outliers and perform regression. It's ideal for datasets based on networks that run on the same system at the same time.

To distinguish between ill and healthy speakers, hyperplane is used. The SVM is trained, which means that one part of the data is chosen for training and the rest is used for testing. To calculate the system's performance and confidence interval, this operation is done several times.

The data is divided into 70 percent training and 30 percent testing parts since the distribution of the train and test subsets is often determined by the database's dimensions. The accuracy gained inside the test set was used to support the choice criterion, which might lead to slightly optimistic findings, but because there are only two parameters to optimise, the bias impact is negligible. The support vector machine is used here because of its proven performance in other research involving the automatic identification of abnormal speech signals.

6. Random Forest and Decision Tree

It is a type of supervised classification technique that employs a large number of decision trees, each of which is trained on a unique set of data. The final forecasts are based on the random forest, which is backed by computations that average each tree's forecast. A random forest is preferred over a single decision tree because it avoids the over-fitting problem, which can lead to wasteful results in conventional classification approaches.

The Random Forest algorithm differs from the Decision Tree algorithm in that the Random Forest algorithm discovers the base node and then randomly separates the feature nodes.

IV. RESULTS AND DISCUSSION

In the proposed research there are four voice disorders that have been considered i.e.,

- Laryngeal
- Cleft Lip
- Normal
- Parkinson

These datasets are downloaded and are trained in a manner where the output is designed to display in classes i.e., [0], [1], [2], [3] respectively. Once the trained samples are executed or run, it displays the feature extracted as shown in figure

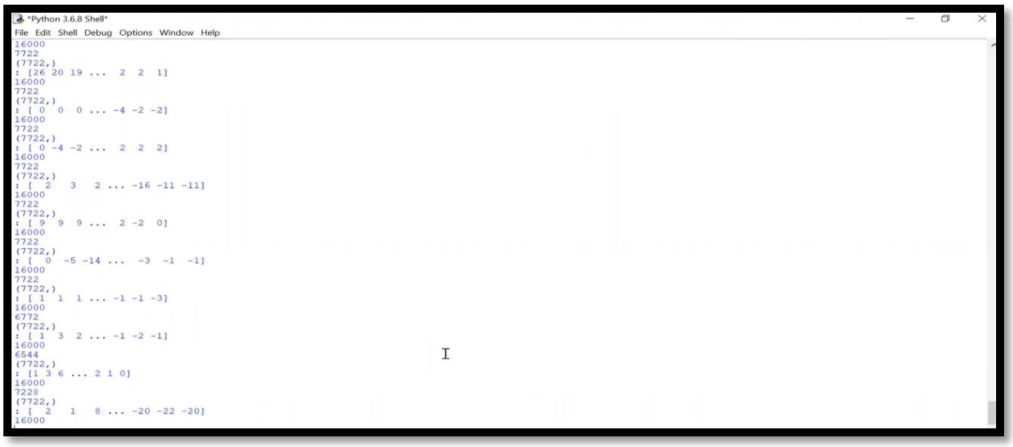

Figure 4: Training Process

In the above figure the pattern followed by the machine is first it shows the features then its corresponding label value also the sampling frequency and then number of features.

Once after the machine ends with reading the folder, it displays the classes where in brackets it reads total 271 samples and feature length will be 7722 as shown in the below figure.

Figure 5: Total Samples Extracted and Features Length

Then the samples start to iterate where this is the phase where training process starts and reduces the errors in the samples. This iteration is continued till the minimum value which means the samples are trained till it becomes zero theoretically. But in practical the machine goes till minimum. This iteration is as shown in the figure below

Figure 6: Iteration with Minimum Loss

Once the iteration is stopped the accuracy is displayed with the Confusion Matrix as shown.

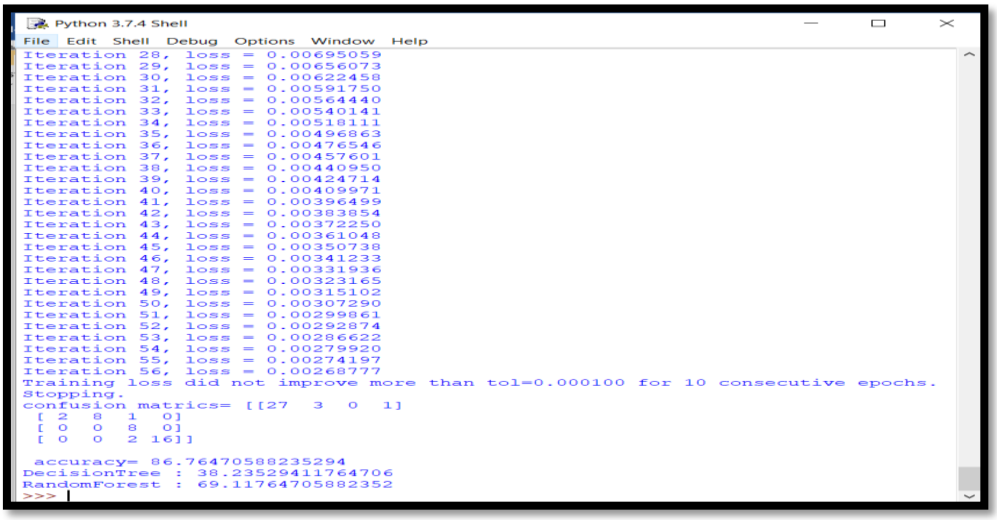

Figure 7: Accuracy with Confusion Matrix

The confusion matrix displayed actually conveys the data of total samples taken from each class and displays the number of true and false samples from the dataset. These values may vary. Once after this according to the iterations, learning curve i.e., learning rate is shown in graph format as shown below. This curve shows the error reduction which plots the Learning rate v/s iterations.

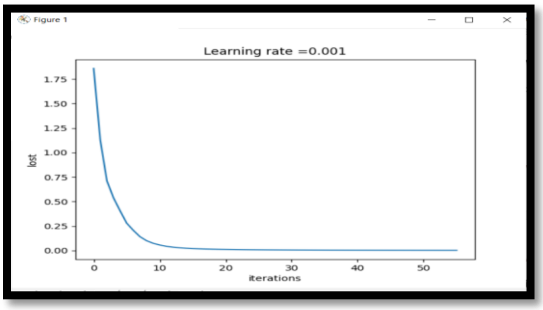

Figure 8: Learning Curve

Testing of samples

Once after the training, coming to testing of samples is successfully done where the sample given as input will be displayed with which voice disorder it belongs to and also its class number will be seen. The figure below shows the result for prediction for Laryngeal and Parkinson before which the figure seen is the selection of sample.

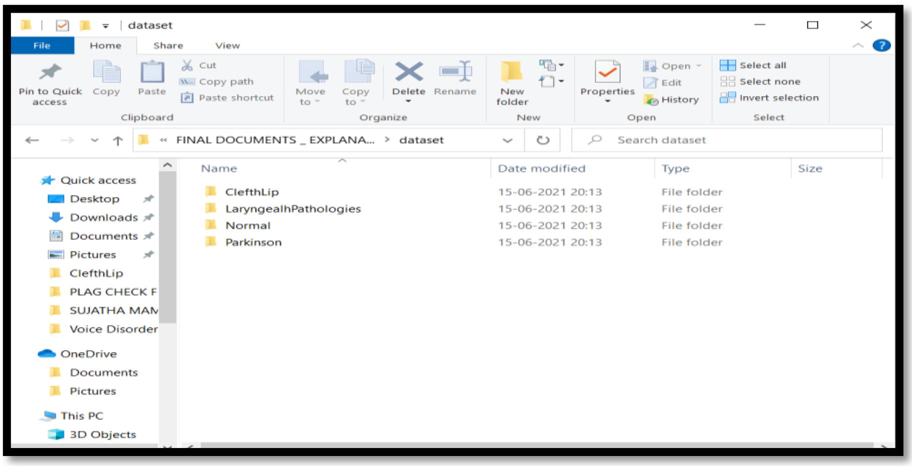

Figure 9: Selection of Sample for Prediction

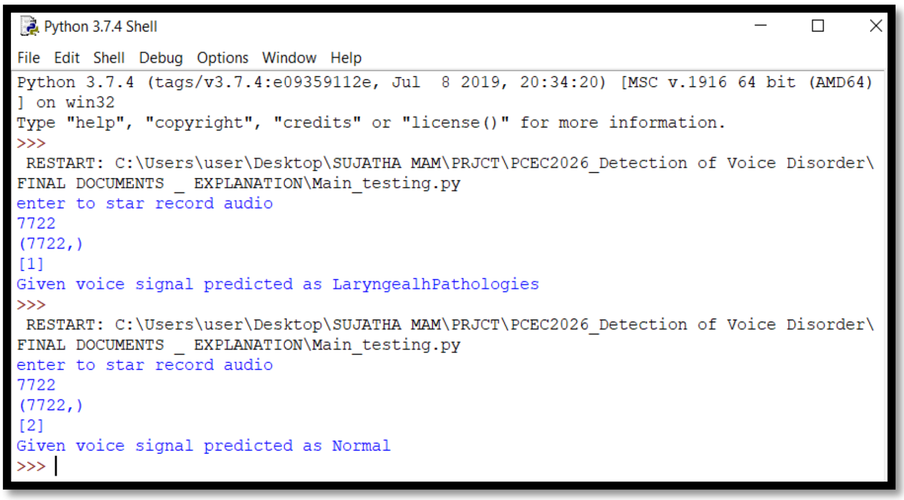

Figure 10: Prediction of Laryngeal and Parkinson

The table below shows the accuracies of different classifiers implemented in the proposed system

Table 1: Comparison of Classifiers

CLASSIFIER	ACCURACY (%)
ANN	86.764
Decision Tree	38.235
Random Forest	68.117

V. CONCLUSION

Voice Disorders are quickly expanding while they are commonly neglected. In order to overcome from this the early measures are to be implemented. In the proposed approach, the ANN, or artificial neural network technique, has been used and implemented. The development of an algorithms that more accurately distinguishes between pathological and healthy voices. The main goal is to evaluate the effectiveness of multiple Machine Learning approaches for detecting vocal pathologies. All of the analysis is done on a voice dataset. The result obtained are evaluated in terms of accuracy. ANN, Random Forest, and Decision Tree are utilized to identify speech pathology. Comparing the techniques, they show that ANN achieves the greatest accuracy in detecting vocal disorders, depending on the characteristics examined using feature selection methods.

VI. REFERENCES

[1] A. Bayestehtashk and M. Asgari and I. Shafran and J. McNames, “Fully automated assessment of the severity of Parkinson’s disease from speech,” Computer Speech & Language, vol. 29, no. 1, pp. 172–185, 2015.

[2] J. Hillenbrand and R. A. Houde, “Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech,” Journal of Speech and Hearing Research, vol. 39, no. 2, pp. 311–321, 1996.

[3] G. Henningsson, D. P. Kuehn, D. Sell, T. Sweeney, J. E. Trost-Cardamone, and T. L. Whitehill, “Universal parameters for reporting speech outcomes in individuals with cleft palate,” Cleft Palate-Craniofacial Journal, vol. 45, no. 1, pp. 1–17, 2008.

[4] L. L. D’Antonio, H. R. Muntz, M. A. Province, and J. L. Marsh, “Laryngeal/voice findings in patients with velopharyngeal dysfunction,” The Laryngoscope, vol. 98, no. 4, pp. 432–438, 1988.

[5] J. Logemann, H. Fisher, B. Boshes, and E. Blonsky, “Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients,” Journal of Speech and Hearing Disorders, vol. 43, pp. 47–57, 1978.

[6] A. K. Ho, R. Jansek, C. Marigliani, J. L. Bradshaw, and S. Gates, “Speech impairment in a large sample of people with Parkinson’s disease,” Behavioral Neurology, vol. 11, pp. 131–137, 1998.

[7] J. R. Orozco-Arroyave, F. Honig, J. D. Arias-Londo ¨ no, ˜ J. F. Vargas-Bonilla, and E. Noth, "Spectral and cepstral ¨ analyses for Parkinson's disease detection in Spanish vowels and words," Expert Systems, pp. 1–10, 2015, in press, 2015.

[8] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease," Journal of the Acoustical Society of America, vol. 129, pp. 350–367, 2011.

[9] C. M. Travieso-Gonzalez, J. B. Alonso-Hern ´ andez, J. R. ´ Orozco-Arroyave, J. Sole-Casals, and E. Gallego-Jutgl ´, "Automatic detection of laryngeal pathologies in running speech based on the HMM transformation of the nonlinear dynamics," Lecture Notes in Computer Science, vol. 7911, pp. 136–143, 2013.

[10] K. Umapathy, S. Krishnan, and V. Parsa, "Discrimination of pathological voices using a time-frequency approach," IEEE Transactions on Biomedical Engineering, vol. 52, no. 3, pp. 421–430, 2005.

[11] J.R. Orozco-Arroyave, J. D. Arias-Londono, J. F. Vargas- ˜ Bonilla, M. C. Gonzalez-R ´ ativa, and E. N ´ oth, "New ¨ Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," Proceedings of the 9th LREC, 2014

[12] T. Bocklet, S. Steidl, E. Noth, and S. Skodda, "Automatic ¨ Evaluation of Parkinson's Speech - Acoustic, Prosodic and Voice Related Cues," Proceedings of the 15th INTERSPEECH, pp. 1149–1153, 2013

[13] M. de Rijk, L. Launer, K. Berger, M. Breteler, J. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder, and A. Hofman, "Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group," Neurology, vol. 54, no. 11 Suppl 5, pp. S21–S23, 2000.

[14] K. Golding, Therapy Techniques for Cleft Palate Speech and Related Disorders. San Diego: Singular Thomson Learning, 2001.