

Early detection of vocal disorders such as laryngeal cancer and dysphonia using voice analysis and machine learning

Authors: Arhaan Garg¹; Reetu Jain²; Syed Abou Iltaf Hussain³

Affiliation: Grade 12 Step By Step School Noida¹; Supervisor, On My Own Technology Pvt. Ltd., Mumbai^{2,3}

Email: gargarhaan13@gmail.com¹; reetu.jain@onmyowntechnology.com²;
syed.hussai@onmyowntechnology.com³

Abstract - Many serious disorders with our throat, such as laryngeal cancer, laryngitis, muscle tension dysphonia, vocal cord paralysis, and so on, are detected after the patient has become critically ill. These disorders can also be life-threatening, as I witnessed with my uncle. Looking at one of the most painful cancers, laryngeal cancer, I wanted to work on a remedy. This occurrence was crucial in directing my attention to this field of study. The majority of these disorders can be discovered early since the voice begins to change due to vocal cord deformations at an early stage. Smoking, drinking, bad eating habits, career, and other factors are all key contributors to these problems. The change in voice is typically the first sign of all of these disorders. People, on the other hand, have a tendency to disregard the very first symptom, which leads them deep into the problem. Voice irregularities, such as variations in frequency, may potentially be too deceiving to the human ear to be taken seriously. Voice disorders such as dysphonia and laryngeal cancer can be detected early using artificial intelligence and machine learning. I worked with Santosh Hospital to collect data and do background study on vocal problems and irregularities. Throughout the procedure, I collected 100+ minutes of audio data from individuals with laryngeal cancer while also researching approaches for detecting voice problems such as laryngoscopy. The project's goal is to distinguish between the voices of a healthy patient and a patient with a vocal cord disorder. A voice analysis comparison between a healthy patient and a patient with a vocal issue was used for this objective. 40 human voice parameters such as frequency, pitch, and zero crossing rate were retrieved using MFCCs and methods such as the

discrete cosine transformation and the mel filter bank. A wrapper was used to pick the most important features in determining if the patient has a vocal problem or not. After that, the logistic regression model was used to train a machine learning model to determine if the audio sample was disordered or healthy. The instrument has an incredibly high accuracy of 88%, making it extremely efficient. This is a technology that assists patients at an early stage in order to keep therapy simple and cure cancer and other critical conditions faster. It also lessens stress on doctors and lowers medical costs while decreasing the effect of sedatives on patients. The technique is very simple to use and available in all places where competent doctors and proper equipment to detect such major voice problems are lacking.

Keywords- *Voice Disorder, MFCC, Voice analysis, Machine learning, Dysphonia Detection, Laryngeal Cancer Detection*

I. INTRODUCTION

The voice box (larynx), which is made up of cartilage, muscle, and mucous membranes, is situated close to the base of the tongue and the top of the trachea. At the beginning of the windpipe are two pliable bands of muscle tissue known as the vocal cords. The vocal cords make sound as they vibrate. Your voice chords are closer together due to air moving through your larynx, which causes this vibration. The vocal cords also assist in closing the voice box, preventing food or liquid from being inhaled after swallowing. A person may have a vocal disorder if there is a problem with their voice's pitch, volume, tone, or other characteristics. One will not

be able to speak normally if the vocal chords get inflamed, develop growths, or become paralyzed.

Laryngitis, muscle tension dysphonia, neurological voice disorders including spasmodic dysphonia, polyps, nodules, or cysts on the vocal cords (non-cancerous lesions), precancerous and cancerous lesions, and vocal cord paralysis are a few of the most prevalent forms of voice problems.

Laryngeal cancer is one of the most severe and common malignancies, and those who have it endure a great deal of discomfort while living in constant fear of not being able to find a solution. This occurrence was crucial in directing one's attention to this field of study and finding a potential solution to it would be a great relief to patients. The majority of these disorders can be discovered early since the voice begins to change due to vocal cord deformations at an early stage. Smoking, drinking, bad eating habits, career, and other factors are all key contributors to these problems. The change in voice is typically the first sign of all of these disorders. People, on the other hand, have a tendency to disregard the very first symptom, which leads them deep into the problem. Voice irregularities, such as variations in frequency, may potentially be too deceiving to the human ear to be taken seriously. Machine learning has the potential to detect vocal cord disorder at an early stage using frequency patterns.

Among the signs of voice abnormalities are having a trembling sound, feeling a rough or harsh (hoarseness) voice that is strained or choppy, shaky, whispery, or breathy and the pitch of the voice is too high or low, or it changes.

Some possible factors include:

1. **Growths.** On occasion, extra tissue may develop on the vocal cords. The cords cannot work properly as a result. Growths include things like fluid-filled sacs called cysts, wart-like lumps called papillomas, and callus-like bumps called nodules. Lesions are defined as areas of scar tissue or patches of damaged tissue. Other growths include polyps, which are tiny blisters, and granulomas, which are tiny pockets of chronic inflammation. Growth may be brought on by sickness, injury, disease, or voice
2. **Inflammation and swelling.** There are several different causes of vocal cord edema and

inflammation. These include verbal abuse, smoking, drinking, surgery, respiratory illnesses or allergies, GERD (acid reflux), some medications, and chemical exposure.

3. **Nerve problems.** Nerves that control the voice chords may malfunction due to certain medical conditions. Examples include Huntington disease, Parkinson disease, myasthenia gravis, multiple sclerosis, and Amyotrophic lateral sclerosis (ALS). Surgical procedures or protracted laryngitis may result in nerve damage (laryngitis).

4. **Hormones.** Thyroid hormone, female and male hormones, and growth hormone deficiency are all potential causes of voice anomalies.

5. **Misuse of the voice.** Misuse of one's voice. When speaking, using too much tension might strain the vocal cords. The voice may suffer as a result of problems with the throat muscles. Vocal abuse can also result in voice problems. Voice abuse is defined as anything that strains or injures the vocal cords. Vocal abuse includes excessive talking, shouting, and yelling. Vocal abuse also includes smoking and frequently clearing one's throat. Abuse of the vocal chords can cause nodes and polyps, which are calluses or blisters, to grow on the vocal cords. These change how voices sound. In some cases, vocal abuse might result in vocal chord rupture. This results in the cord bleeding, which might lead to voice loss. The bleeding of the vocal cords needs to be addressed as soon as possible.

A. Motivation and Novelties

The present study aims for the early detection of vocal disorders such as laryngeal cancer and dysphonia. The chances of curing cancer become higher with the early detection of the cancer. Cancer can be only cured if detected at an early stage. Aiming to use all resources at disposal, a novel tool that helps in the early detection and helps save lives alongside reducing the costs of detection. From the overall literature reviewed for this study, it is noted that the application of machine learning (ML) model in integration with frequency pattern for detecting and diagnosing laryngeal cancer and dysphonia is very little. Hence, in this paper Logistic Regression (LR) is employed with close conjunction with the Mel-Frequency Cepstral Coefficients (MFCC) to achieve the aim of the paper.

II. LITERATURE REVIEW

In paper [1] this regard, the research presents a non-invasive voice illness diagnosis method for laryngeal cancer patients. Fifty-five laryngeal and fifty-five healthy cases of the sustained vowel /a/ were recorded. Seven nonlinear parameters are retrieved along with biologically induced 39 Mel-Frequency Cepstral Coefficients due to the nonlinearity of the vocal chords (MFCC). Laryngeal documentation measuring 110 by 46 pixels. The wrapper technique is used to strengthen the discriminating power of current work and to improve function selection. To accomplish the type, a customised Support vector machine (SVM) with grid seek and random forest is employed (RF). The new experiment showed improved accuracy of 80% for random forest and 76.56% for SVM. The introduction of non-linear functions and the forward selection of capabilities have both significantly enhanced the overall performance of the current device.

In paper [2], the author suggested using the idea of human voice production to automatically identify premalignant lesions. The vocal fold is closely linked to premalignant lesions including leukoplakia, erythroplakia, keratosis, and others, hence features extracted from the glottal fold waveform can be pertinent and important. The proposed technique recovers the vocal fold waveform from recorded utterances without any intrusion. The main goal is to separate pertinent data from the raw signal (glottal waveform). Such a signal is not, however, easily accessible. As a result, we begin by extracting the glottal waveform from recorded speech using Iterative Adaptive Inverse Filtering (IAIF). The Saarbrücken Voice Database and the Massachusetts Eye and Ear Infirmary (MEEI) database both have databases with sustained vowel samples (SVD). Glottal flow impulses are used as relevant instances to develop pertinent attributes. The most important and pertinent features are chosen after a thorough analysis of the obtained features using statistical techniques like boxplot and Principal Component Analysis (PCA). The Support Vector Machine Tool is used as a classification method to distinguish between normal and premalignant tumours. According to this study, premalignant lesions can be identified with respectable and acceptable sensitivity, specificity,

precision, and accuracy. These algorithms can help with the earlier diagnosis of laryngeal cancer.

In paper [3], the author suggests that medical applications for machine learning (ML) algorithms are well-being states based on evaluating the different characteristics that have a significant influence on illness. One of the few illnesses affecting people for whom doctors are still looking for the ideal treatment, cancer is also unanticipated. Cancer is a complex illness, and different treatments are used for different types and stages. The term "throat cancer" refers to a tumour that invades the tonsils, throat, or voice box (larynx) (pharynx). It is highly recommended to identify throat cancer and start therapy right away. To predict throat cancer, especially for supervised learning classification algorithms, deep learning (DL) image processing techniques and machine learning (ML) approaches are used. This report examined the ML and DL-based research activities that have been conducted to categorize throat cancer.

In paper [4], the authors introduce a novel marker, the dysphonia detection index, in this work, which might be integrated into a mobile health solution and help with the assessment of voice abnormalities. To evaluate the general state of the voice and determine whether a vocal issue occurs, four acoustic characteristics are merged into a single marker. The relationship between these characteristics was examined using a model tree regression approach, and the threshold value to discriminate between a pathological and a healthy voice was estimated using a Youden analysis. To evaluate the suggested index's dependability, its accuracy, sensitivity, and specificity have been precisely categorised. We used a dataset of 2003 voices from the Massachusetts Eye and Ear Infirmary, the Saarbrücken Voice, and the VOice ICar fEDerico II databases to assess the effectiveness of our proposed index. Our technique surpassed other algorithms in terms of performance, with an accuracy of 82.2 percent, sensitivity and specificity of 82 percent and 82.6 percent, respectively.

In paper [5], the neural network approach of Kohonen's self-organizing map was utilized to recognize the spectrum patterns of dysphonia. A team of speech pathologists assessed the speech samples, which included 17 men and 18 women speaking Finnish words with long [a:]. The positions

of the [a:] samples on a self-organized spectral feature map were then compared to the judgments made about the speech samples. The map revealed a statistically significant distinction between normal and dysphonic speech spectra. The energy ratio at 1-2 kHz and 7-9 kHz, which supported the disparity, was the most obvious spectral component.

In publication [6], they previously described the phenotypes of hypokinetic dysarthria, stuttering, breathy voice, strained voice, and spastic dysarthria in Parkinson's disease (PD) patients receiving STN-DBS. Changes throughout time, however, remain a mystery. 32 Parkinson's disease patients were assessed before and up to a year after surgery for this research (PD-DBS). Additionally, eleven Parkinson's patients were assessed who were on medication (PD-Med). The functions of speech, voice, movement, and cognition were assessed. At the beginning, all groups reported similar rates of hypokinetic dysarthria (50 percent vs. 45 percent), breathy speech (66 percent vs. 73 percent), and strained voice (3 percent vs. 9 percent) (63 percent of PD-DBS vs. 82 percent of PD-Med). At one year compared to baseline, only the PD-DBS group had a marginally significant decline in speech intelligibility ($p = 0.001$) and dysphonia grade ($p = 0.001$). Only strained voice (28%) and spastic dysarthria (44%) did not appear during the follow-up in either PD-DBS or PD-Med, although stuttering (9% vs. 18%) and breathiness (13% vs. 9%) occurred. The majority of the patients' breathy and strained voices, as well as their spastic dysarthria, improved after the stimulation was ceased. These results suggest that strained voice and spastic dysarthria are the most frequent speech and vocal issues brought on by DBS, and that STN-DBS may exacerbate stuttering and breathy voice. It could be easier to spot speech and voice impairment early on with a greater knowledge of these illnesses, which might lead to more successful therapies.

III. PRELIMINARY CONCEPTS AND METHODOLOGY

In this section of the paper a brief description about the ML techniques adopted for building the detection system is given.

A. Machine learning (ML)

Software systems may anticipate outcomes more correctly with the use of machine learning (ML), a type of artificial intelligence (AI), without needing

to be explicitly told to do so. Machine learning is a crucial component of data science, a rapidly increasing field. In order to provide classifications or predictions and uncover crucial insights in data mining projects, algorithms are trained using statistical approaches. Ideally, the choices taken as a result of these insights affect important growth metrics in applications and businesses. As big data continues to grow and improve, data scientists will become increasingly in demand. They will be required to help identify the most important business queries and the data required to answer them.

Logistic regression

This type of statistical model, often known as a logit model, is extensively used in classification and predictive analytics. Logistic regression determines the probability that an event, such as voting or not voting, will occur based on a collection of independent factors. Since the outcome is a probability, the range of the dependent variable is 0 to 1. The odds, or likelihood of success divided by probability of failure, are converted using the logit formula in logistic regression. This logistic function, often known as the log odds or the natural logarithm of odds, is represented by the following formulae.

This type of statistical model, often known as a logit model, is extensively used in classification and predictive analytics. Logistic regression determines the probability that an event, such as voting or not voting, will occur based on a collection of independent factors. Since the outcome is a probability, the range of the dependent variable is 0 to 1. The odds, or likelihood of success divided by probability of failure, are converted using the logit formula in logistic regression. This logistic function, often known as the log odds or the natural logarithm of odds, is represented by the following formulae.

Once the model has been calculated, it is advised to evaluate the model's goodness of fit, or how well it predicts the dependent variable. A popular method for assessing model fit is the Hosmer-Lemeshow test. This analytics approach may be used in medicine to predict the chance of a certain group of people contracting a disease or becoming unwell. Healthcare facilities can set up preventative care for patients who are at a higher risk of contracting a certain illness.

B. MFCC

For this, we use MFCCs, or Mel frequency cepstral coefficients. They come from the cepstral representation of an audio sample (a nonlinear "spectrum-of-a-spectrum"). The frequency bands of the mel-frequency cepstrum (MFC) are evenly spaced on the mel scale, which is different from the cepstrum in that it more closely resembles the response of the human auditory system than the linearly-spaced frequency bands used in the normal spectrum. Frequency warping can enhance the depiction of sound. To get the MFCCs, we employ the subsequent process. The following is a common derivation of MFCCs:

1. Perform a signal's Fourier transform on a windowed extract.
2. Use triangle overlapping windows or, as an alternative, cosine overlapping windows to map the powers of the spectrum acquired in step one onto the mel scale.
3. Calculate the power logs for each of the mel frequencies.
4. Treat the list of mel log powers' discrete cosine transform as a signal.
5. The amplitudes of the resultant spectrum are the MFCCs.

Procedure of extracting MFCC

Preemphasis

Pre-emphasis increases the amount of energy at the higher frequency. When we look at the frequency domain of the audio signal for voiced segments like vowels, the energy at a higher frequency is much smaller than the energy at a lower frequency. By improving phone detection accuracy, increasing energy at higher frequencies will improve the model's performance. Pre-emphasis is carried out using the first order high-pass filter, as seen below. The frequency domain of the vowel "aa" in the audio stream is displayed below, both before and after the pre emphasis.

Windowing

The MFCC method's objective is to extract elements from the audio stream that may be used to identify phones in speech. The audio signal will be split into separate segments, each with a width of 25 ms and

the signal spaced at 10 ms, as shown in the illustration below, because there will be multiple phones in the audio transmission. With four phones and an average speech rate of three words per second, a person may speak 36 states per second, or 28 milliseconds per state, which is within the range of our 25 millisecond window. 39 features will be taken from each segment. Furthermore, the dramatic reduction in amplitude at the edges will result in noise in the high-frequency domain if we immediately slice the signal off at the edges while breaking it. Therefore, to chop the signal, we will employ Hamming/Hanning windows rather than a rectangular window, which will prevent noise in the high-frequency region.

DFT (Discrete Fourier Transform):

We will use the dft transform to translate the signal from the time domain to the frequency domain. Analyzing audio signals in the frequency domain is simpler than in the time domain.

Bank Mel-Filter:

Machines do not hear sound the same way that our ears do. Our hearing is more precise at lower frequencies than at higher ones. Therefore, even if there is a 100Hz difference between sounds at 1500 Hz and 1600 Hz, we can readily discriminate between them when we hear noises at 200 Hz and 300 Hz. On the other hand, the machine's resolution remains the same at all frequencies. The model's performance has been shown to be enhanced by replicating the human hearing property during the feature extraction stage. So, in order to translate the true frequency into a frequency that people can sense, we will use the mel scale. Below is a formula for mapping.

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

Application of log:

Humans are less sensitive to changes in audio signal energy at higher vs lower energy levels. The log function has a similar trait in that it has a bigger gradient at low input values but a lower gradient at high input values. So, to simulate the human hearing system, we apply a log on the Mel-filter output.

IDFT:

The result of the previous phase is reverse transformed in this stage. We must first understand how people create sound in order to understand why we must do an inverse transform. The sound is produced by the glottis, a valve that regulates airflow into and out of the respiratory airways. The air in the glottis vibrates, which creates the sound. The vibrations will take the form of harmonics, with the fundamental frequency being the lowest frequency generated and all other frequencies being multiples of the fundamental frequency. The vocal cavity will receive the ensuing vibrations. The vocal cavity selectively amplifies and dampens frequencies based on the placement of the tongue and other articulators. Each sound will have a certain tongue and other articulator's location. A cepstrum is the inverse of the logarithm of the signal's magnitude.

The fundamental frequency, which is located at the figure's rightmost peak, will offer information on the pitch, while the frequencies to the right will do the same for the phones. The fundamental frequency won't be used because it doesn't have any phone-related information.

$$Energy = \sum_{t=t_1}^{t_2} (x^2[t]) \quad (2)$$

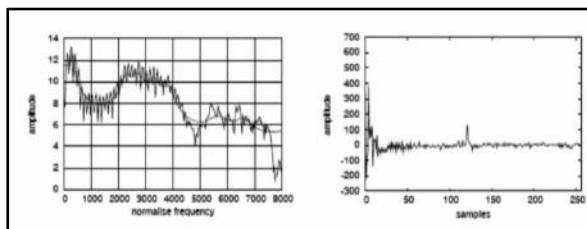


Figure 1: MFCC converted by IDFT

The MFCC model uses the first 12 coefficients of the signal following the completion of the idft operations. In addition to the 12 coefficients, it will also make use of the signal sample's energy as a feature. It will help with phone identification. The energy formula for the sample is shown below.

Dynamic Features:

The MFCC technique will consider the first and second order derivatives of the remaining 26 characteristics in addition to these 13 features. These coefficients between audio signal samples are used to create derivatives, which help explain how the transition happens. A model that determines whether

the patient has a voice issue or not will be given 39 characteristics produced by the MFCC approach from each audio signal sample.

C. METHODOLOGY

Voice features are extracted using librosa from an audio file for a healthy patient and one with vocal disorders. Further, the best feature is chosen to classify the disorders. We then split the feature array and created a dataset and classified each according to the diagnosis and age. Then the best feature was chosen and logistic regression was applied. Further, we encoded the data using feature mapping. Our next step was to scale the features and then split the dataset into test and train so our model could be accurately validated. We kept 80% as training data and 20% test data. Then we used feature scaling to normalize the range of features in the data. Our next step was to use logistic regression to train the model accurately. After successfully training the model, we checked the accuracy of the model and found it to be efficient with an accuracy of 87.5%.

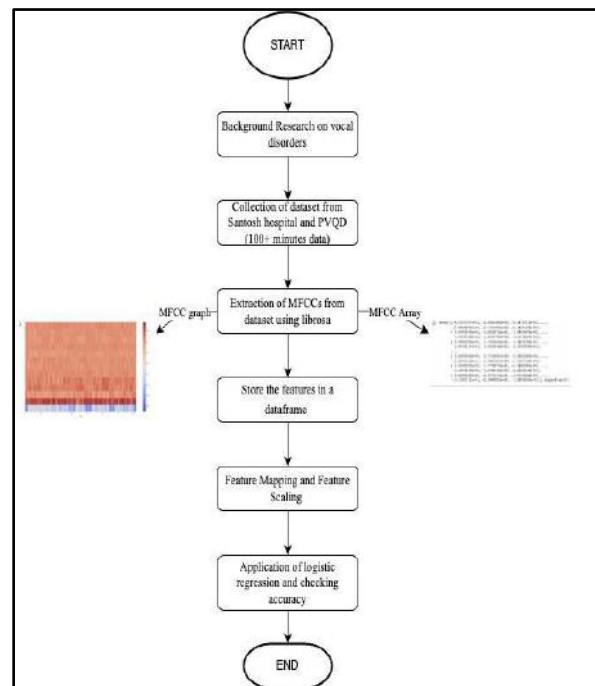


Figure 2: Methodology flowchart

This was the process used to design the research study, and it was conducted by the above method.

IV. RESULTS AND DISCUSSIONS

The proposed methodology is coded in python 3.9 using the Visual Code Studio and runs on the Macbook Air M1 processor.

A. Results and validation

The confusion matrix is as follows:

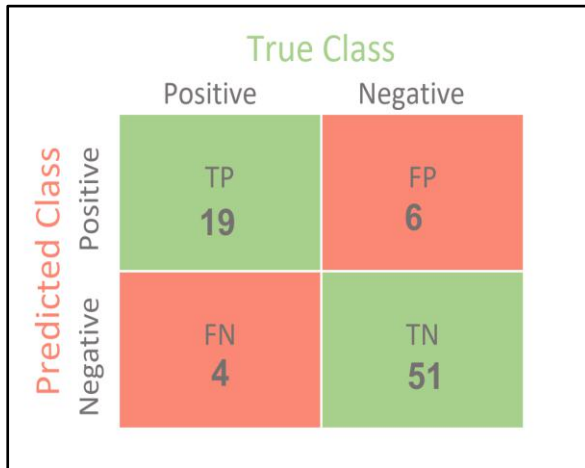


Figure 3: The confusion matrix

The "TN" in figure 3 stands for True Negative and indicates the number of correctly recognised negatively classified cases. Similar to this, "TP" signifies the quantity of accurately detected positive instances and stands for True Positive. The abbreviation "FP" stands for "real positive instances that were wrongly classified as negative," while "FN" stands for "genuine positive cases that were incorrectly classified as negative." The following is the categorization report, which shows the F1 score, accuracy, and precision.

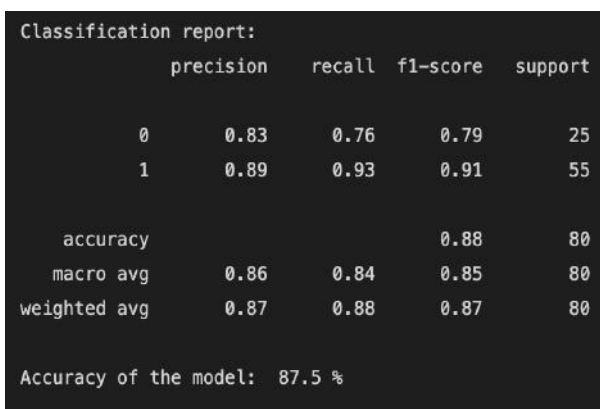


Figure 4: The classification report

The Receiver operating characteristics (ROC) curve is shown in figure 5.

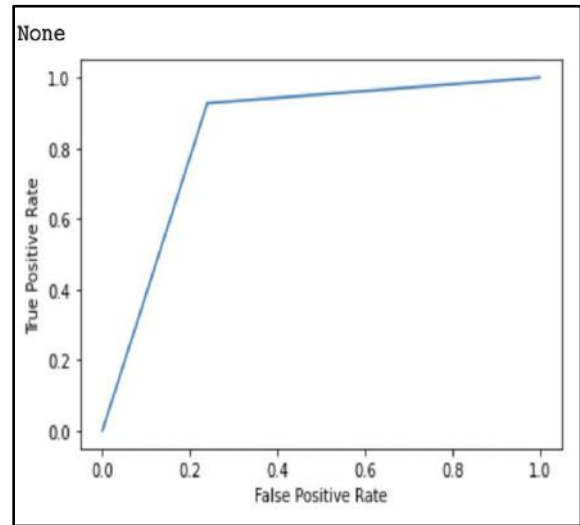


Figure 5: The ROC curve

B. Discussions

The prediction results of a classification task are summarised in a confusion matrix. The number of accurate and incorrect predictions for each class is expressed using count values. This is the key to the confusion matrix. One of the most often employed measures in categorization is accuracy. A set of categorization prediction results is referred to as a confusion matrix. The accuracy of a model is determined using the formula below (via a confusion matrix).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The accuracy of our machine learning model, which is shown above, is 87.5%, which is indicative of a larger percentage of accurate predictions of whether a voice sample is pathological or healthy. The two most popular metrics that account for class imbalance are precision and recall. Additionally, they serve as the basis for the F1 score. Precision measures the percentage of predictions that are accurate and lies within all of the positive predictions. Recall is contained within everything that is genuinely positive. The harmonic mean of recall and precision is used to calculate the F1 score.

$$F1\ score = \frac{Recall * Precision}{Recall + Precision}$$

The AUC-ROC curve is a performance measure for classification problems at various threshold levels. ROC is a probability curve, and AUC stands for the level or measurement of separability. It demonstrates how effectively the model can distinguish between classes. An AUC near to 1 indicates a high level of separability, which is a sign of a good model. Our model can discriminate between positive and negative classes 90% of the time with an AUC of 0.9.

V. CONCLUSION AND FUTURE SCOPE

This method will allow people to discover whether they have a vocal disease, such as laryngeal cancer, at an early stage, preventing significant problems from developing. These concerns are typically disregarded in the early stages since changes in voice, such as frequency, pitch, and breath, are not detectable by humans. As a result, individuals are ignorant and tend to ignore these difficulties, thus complicating the situation. Certain circumstances are highly dangerous and cannot be remedied. In certain regions of the world, such illness detection tools do not even exist, and the number of specialists in the field is also relatively restricted. This would be available to everyone and would help save many lives. It will allow the population to receive treatment on time, making it beneficial in saving lives. Many lives could be spared around the world if early detection is used. It would also drastically cut the cost of checkups because it would be a publicly accessible application. Additionally, it would ease stress on doctors, and patients would not have to undergo rigorous detection techniques such as laryngoscopy, which can be uncomfortable and require the use of sedatives. This app would employ a straightforward methodology, requiring only that the patient's voice be recorded in order to determine whether he or she has a vocal issue. It would be accomplished by a simple strategy including the usage of a basic application. This would be done using the machine learning model, which has a high accuracy of 87.5%. The model is capable of detecting all kinds of voice disorders including laryngeal cancer. The main advantage of this technology is that it does not exist and has not been worked on previously, therefore we will present people with a new perspective on typical health-related concerns.

One of our primary goals would be to expand the dataset and improve model accuracy. We would also focus on additional uses, such as establishing prediction models that are effective at forecasting what type of person in a specific job or doing certain activities could suffer from based on their histories. We will add breathing rate to our model, which will be useful in detecting other voice-related disorders as well. Other vocal cord-related issues could be diagnosed using the extensive frequency data we have. Major studies on vocal cord issues could be conducted.

We will also work on the mobile application and its user interface to make it more user-friendly and interactive. Then we'd need to raise greater knowledge about the app and its benefits so that more people may use it and benefit on a broad scale.

ACKNOWLEDGEMENT

1. We would like to acknowledge Dr. Abhay Singh ENT specialist at Santosh Hospital for providing us background knowledge about voice disorders and the treatment procedures.
2. We would acknowledge Dr. Tripta Bhagat, chancellor at Santosh medical college for helping us in fetching data from patients suffering from Dysphonia, laryngitis and other diseases.

REFERENCES

- [1] Gour, G. B., Udayashankara, V., Badakh, D. K., & Kulkarni, Y. A. (2020). Voice-Disorder Identification of Laryngeal Cancer Patients. *International Journal of Advanced Computer Science and Applications*, 11(11).
- [2] Aicha, A. B. (2018). Noninvasive detection of potentially precancerous lesions of vocal fold based on glottal wave signal and SVM approaches. *Procedia Computer Science*, 126, 586-595.
- [3] Akshara, R., & Latchoumi, T. P. (2021). A Survey: Identification of Throat Cancer by Machine Learning. *Annals of the Romanian Society for Cell Biology*, 6616-6622
- [4] Verde, L., De Pietro, G., Alrashoud, M., Ghoneim, A., Al-Mutib, K. N., & Sannino, G. (2019). Dysphonia detection index (DDI): A new multi-parametric marker to evaluate voice quality. *IEEE Access*, 7, 55689-55697.
- [5] Leinonen, L., Hiltunen, T., Kangas, J., Juvas, A., & Rihkanen, H. (1993). Detection of dysphonia by

- pattern recognition of speech spectra. *Scandinavian Journal of Logopedics and Phoniatics*, 18(4), 159-167
- [6] Tsuboi, T., Watanabe, H., Tanaka, Y., Ohdake, R., Hattori, M., Kawabata, K., ... & Sobue, G. (2017). Early detection of speech and voice disorders in Parkinson's disease patients treated with subthalamic nucleus deep brain stimulation: a 1-year follow-up study. *Journal of Neural Transmission*, 124(12), 1547-1556
- [7] Pham, Minh & Lin, Jing & Zhang, Yanjia. (2018). Diagnosing Voice Disorder with Machine Learning. 5263-5266. 10.1109/BigData.2018.8622250.
- [8] Mittal, Vikas and R. K. Sharma. "Deep Learning Approach for Voice Pathology Detection and Classification." *IJHISI* vol.16, no.4 2021: pp.1-30. <http://doi.org/10.4018/IJHISI.20211001.0a28>
- [9] Hu H, Chang S, Wang C, Li K, Cho H, Chen Y, Lu C, Tsai T, Lee O, Deep Learning Application for Vocal Fold Disease Prediction Through Voice Recognition: Preliminary Development Study, *J Med Internet Res* 2021;23(6):e25247 URL: <https://www.jmir.org/2021/6/e25247> DOI: 10.2196/25247
- [10] Won Ki Cho, Seung-Ho Choi, Comparison of Convolutional Neural Network Models for Determination of Vocal Fold Normality in Laryngoscopic Images, *Journal of Voice*, 2020,ISSN 0892-1997, <https://doi.org/10.1016/j.jvoice.2020.08.003>. (<https://www.sciencedirect.com/science/article/pii/S0892199720302927>)
- [11] Jonathan Reid, Preet Parmar, Tyler Lund, Daniel K. Aalto, Caroline C. Jeffery, Development of a machine-learning based voice disorder screening tool, *American Journal of Otolaryngology*, Volume 43, Issue 2, 2022, 103327, ISSN 0196-0709, <https://doi.org/10.1016/j.amjoto.2021.103327>. (<https://www.sciencedirect.com/science/article/pii/S0196070921004282>)
- [12] H. Wu, J. Soraghan, A. Lowit and G. Di Caterina, "Convolutional Neural Networks for Pathological Voice Detection," *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 1-4, doi: 10.1109/EMBC.2018.8513222.
- [13] B. Halpern, J. Fritsch, Enno Hermann, R. V. Son, O. Scharenborg, M. Magimai-Doss, "An Objective Evaluation Framework for Pathological Speech Synthesis," *ITG Conference On Speech Communication* 2021,
- [14] Kim, H.; Jeon, J.; Han, Y.J.; Joo, Y.; Lee, J.; Lee, S.; Im, S. Convolutional Neural Network Classifies Pathological Voice Change in Laryngeal Cancer with High Accuracy. *J. Clin. Med.* 2020, 9, 3415. <https://doi.org/10.3390/jcm9113415>