iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35658

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

Disease Prediction using Voice Analysis

Ajay Dabas[1], Pratyush Avi[2]
[1, 2]*Student, Information Technology, Svkm's-Nmims, India*

Abstract: ***Researchers enlist Machine learning a viable solution to find vocal patterns. Our believe is voice can relay key information about personnel health and, our research proposes working towards medical diagnosis and disease risk through voice analysis. In next few years, to the advent of technology in medical science, we positively believe health conditions will be diagnosed using smartphones and another wearable technology. This paper aims to propose viable techniques to researchers that will work by recording short speech samples and analyzing for underlying diseases.***
For psychiatric disorders particularly, there are no blood samples and patients are often embarrassed to talk about facing mental health issues therefore, voice analysis can also be used over here to identify peculiar traits.
Different countries are collecting voice analysis samples to test their tone pitch, rhythm, rate, and volume for sign of prediction like PTSD and other brain injuries and depression. Using Machine learning algorithms will help identify vocal patterns of personnel suffering diseases or any other conditions by comparing data with healthy individuals' voice samples.
Keywords: Machine Learning, Supervised Learning, Voice Analysis, Classification Problem, Feature Selection

I. INTRODUCTION

The aim of this project is to enable disease prediction through voice analysis. We are trying to develop a technology that extracts various acoustic features from a speaker's voice, in real time, giving insights on personal health condition, well-being, and emotional understanding, to develop a voice-based technology with the potential to transform the way we monitor and diagnose mental and physical health [1]. We must make a functioning machine learning model that will be able to take the voice of an individual as input. This input will then be worked upon, and different parameters will be extracted, and these parameters will be mapped to detect the diseases of the individual based on the factors in their voice. Libraries like Librosa can be used for voice extraction, and features will then be predicted based on the voice. Our speech pattern can tell us the conditions of our various organs and help us identify any anomalies. It can also be used to detect the mood of an individual based on factors such as pitch, frequency, and amplitude. We can recognize different patterns in our speech using deep learning techniques and few preset benchmarks. This model can be very helpful and cost efficient if implemented at a large scale**.**

II. MODEL

In this section, we provide a comprehensive overview of the various tools which can be used by the users to evaluate, compare, and optimize pattern, plans and analysis.

1) *Data Collection* is to collect indigenous data and create a usable dataset with the relevant information. The aim is to prepare datasets that will aide in disease screening through the user's voice samples and basic information. The collected data should give us a clear idea of the input entities for the dataset to be used effectively. This step is about preparing the data in such a way that it best exposes the structure of the problem and the relationships between our input attributes with the output variable. This is also the home of any global configuration we might need to do. It is also the place where you might need to make a reduced sample of our dataset if it is too large to work with. Ideally, our dataset should be small enough to build a model or create visualization within a minute. To collect samples from volunteers willing to co-ordinate for the purpose.

Each tuple can contain the following:
a) A voice sample.
b) Additional voice sample in case the 1st gets corrupted.
c) Heart Rate (Systolic and Diastolic).
d) Gender of participant (Preferably Male/Female.)

The voice samples will preferably be recorded in a sound insulated room (conference room) to prevent any unwanted noise from mixing with the sample and since it can help us save time spent, cleaning the voice samples for any background noise present.

Additionally, the extra voice samples collected will go a long way to help us replace any corrupted/unusable voice samples originally collected and will act as a valuable backup for the same.
The heartbeat metrics will let us know any underlying issues which might be present (e.g., an abnormally high resting heart rate) that might variate the results, differ from accuracy in our study.
The gender metrics will help us label data samples male or female voice samples so that the pitch/frequency difference between voice samples (females have a higher pitch naturally) does not render the model inaccurate. We can calculate abnormal metrics for both the genders in this manner without risking accuracy of prediction from our expected model.
To have a fully operable dataset which is ready to be used to perform calculations upon and various operations in a .csv file format which entails values for everything accompanied by a .wav file which contains respective voice samples for training and testing.

2) *Data Pre-processing* is where pre-processing of data occurs. The difficulty we mostly face int the stage is that different algorithms make different assumptions about data and may require different transformation with respect to pre-processing. Further, when we follow all rules and prepare our data, sometimes algorithms can deliver better results without pre-processing but not better accuracy. Generally, trying to create many different views and transforms of our data, then examine of algorithms on each view of the dataset is encouraged for the purposes of this study. This will help winnow out data transforms that might be better at exposing the structure of our problem in general.

To streamline the dataset collected i.e., to:

a) rescale the data,
b) standardize the data,
c) to make it homogenous and more conformed,
d) normalize the data and if possible,
e) make the data binary in nature for easier calculation.

Data preprocessing is critical in a machine learning pipeline. It directly impacts success rate of the project. Data is said to be unclean if it is missing an attribute, attribute values, contains noise or outliers and duplicate or wrong data. Presence of any of these will degrade quality of the results. Data cleaning is a very fundamental yet necessary task that we must perform so that our model does not provide overly positive test results or breaks. Any columns in our dataset table that has a single value is basically useless for calculation since it shows zero variance. Columns which have extremely similar values are also treated the same since they too, would have negligible variance whereas we require unique values in our dataset to learn a pattern and train our model. Also, any rows that have similar values are removed since the extra row is useless to us in terms of data. The pandas function duplicated will report whether a given row is duplicated or not. All rows are marked as either False to indicate that it is not a duplicate or True to indicate that it is a duplicate. If there are duplicates, the first occurrence of the row is marked False (by default), as we might expect.
We have a clean, integrated and fully functional dataset which is divided into training set and test set in a 70:30 ratio i.e., 70% train data and 30% test data. The data is suited for use by the machine algorithms at this stage. It is successfully utilized by the algorithms for the purpose of training and testing.

3) *Feature Extraction* is another procedure that is liable to data transformation. Like data pre-processing, feature extraction procedures must be restricted to the data in the training dataset. The extraction utilizes a tool called the Librosa which allows the results of multiple feature selection and extraction procedures to be combined into dataset on which a model can be trained. Since there could be tons of data available and many features for each target variable to be calculated, it can lead to dilution by wasting time choosing features manually and inviting overfitting as a reason. To avoid these kinds of problems, it is necessary to apply either regularization or dimensionality reduction techniques (Feature Extraction). Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features). These new reduced set of features should then be able to summarize most of the information contained in the original set of features. We choose the Librosa library as it is one of the most widely used library choices for the purpose of music and audio analysis. It can be efficiently used to find personal and key characteristics from audio [2]. It has functions that are predefined in the library that performs different operations and gives a visually available result. It can give us time v/s sound charts, amplitude charts, sampling rate, and different specialized charts.

Some of Feature Extraction processes are:

a) *Mel-Scaled Spectrogram:* It represents an acoustic time-frequency representation of sound. It is a way to visually represent a signal's loudness, or amplitude, as it varies over time at different frequencies. In 1937, Stevens, Volkmann, and Newmann proposed a unit of pitch such that equal distances in pitch sounded equally distant to the listener. This is called the Mel **scale. So** basically, A Mel spectrogram is a spectrogram where the frequencies are converted to the mel scale.

b) Zero Crossing Rate: The zero-crossing rate indicates the number of times that a signal crosses the horizontal axis (zero). The zero-crossing count is an indicator of the frequency at which the energy is concentrated in the signal spectrum. Estimates of spectral properties can be obtained using a representation based on the short time average zero-crossing rate.

c) Spectral Centroid: The spectral centroid indicates where the center of mass of the spectrum is located. It is calculated as the weighted mean of the frequencies present in the signal, determined using Fourier transform, with their magnitudes as the weights to be multiplied. It is calculated as a ratio. To put it simply, it tells us the center of gravity of the spectrum.

d) Spectral Roll-off: Spectral roll-off is the frequency below which a specified percentage of the total spectral energy, e.g., 90%, lies.

e) MFCC — Mel-Frequency Cepstral Coefficients: This feature is one of the most important method to extract a feature of an audio signal and is used whenever working on audio signals. The Mel frequency cepstral coefficients (MFCCs) of a signal are a small set of features (usually about 15–20) which concisely describe the overall shape of a spectral envelope. By printing the shape of MFCCs we can gauge how many mfccs are calculated on how many frames.

III. ALGORITHM PROCESSING

This step is entails finding subset of machine learning algorithms that are good at exploiting the structure of our sample data. The metrics that are chosen to evaluate machine learning algorithms are very important. Choice of metrics influences how the performance of machine learning algorithms is measured and compared. They influence how we will weigh the importance of different characteristics in the results and our ultimate choice of which algorithms to choose.

The process of choosing algorithms to work or make it functional on chosen dataset is set in motion. Having already extracted voice samples in our data collection procedure we take other relevant data of individual patients', now proceeding to step wise algorithm processing.

1) *Classification Algorithm:* We are using classification algorithm because it is a type of supervised learning algorithm which is used when output has finite and discrete value. It helps us predict a class for an input variable. As we are following supervised technique for Disease predication, Classification is best suitable algorithm to be used in to store and label dataset. Therefore, we are labelling our dataset for further function. Classification of data is further undergone through algorithm process:

a) *Classification Accuracy:* Classification accuracy is the number of correct predictions made as a ratio of all predictions made. This is the most common evaluation metric for classification purpose. It is only suitable when there are an equal number of observations in each class and that all predictions of samples of record and their observation. Now we can have the proportion that can be reported. This can be converted into a percentage by multiplying the value by 100 to give an accuracy result of approximately digit % accurate in our model.

b) *Logarithmic Loss:* It is a performance metric for evaluating the predictions of probabilities of sample and record to a given class. The scalar probability between 0 and 1 has been given as a measure of confidence for a prediction by an algorithm. Predictions that are correct or incorrect are rewarded or punished proportionally to the confidence of the prediction. So over here prediction of sample will be done on bases of probability.

c) *Confusion Matrix:* The matrix is a presentation of the accuracy of a model with two or more classes. The table presents predictions on the x-axis and true outcomes on the y-axis. The cells of the table are the number of predictions made by a machine learning algorithm. In our sample of records, we will predict 0 or 1 and each prediction may have been a 0 or 1. And so on.

d) *Classification Result:* The scikit-learn library provides us convenience report when working on classification problems to give us a quick idea of the accuracy of a model using several measures. The classification report function displays the precision, recall, F1-score, and support for each class. The sample which we are demonstrating the report on the binary classification problem.

Now we have initialized the classifier to be used, we will now train the classifier in scikit-learn to fit (X, Y) method to fit the model or training for the given data X and train label Y. Now the given unlabelled observation X, the prediction(X) returns the predicted label Y and finally we classify the model.

- *Decision Tree:* It is defined as giving a data of attributes together with its classes, the decision tree produces a sequence of rules that can be used to classify the data. It is simple to understand and visualise, requires little data preparation, and can handle both numerical and categorical data. Tree models where the target variable can take a discrete set of values are called **classification trees**. We will be using the few libraries in our model like gmodels - To display the result of prediction in cross table format. Then dplyr - for Data manipulation package and the C50 - for Decision trees and rule-based models for pattern recognition. This function we are using to predict values is based on linear model object. Parameters passed to this function will be object to variable obtained from earlier step and new data for the original dataset, excluding the column gender. The main aim to consider in our model that decision tree helps us to sort the data in systematic way in the form of branch and node. For example, in our model we could provide age along with record so that they can be considered into factors during prediction of sample and relate them with other samples according to prediction.We must avoid over-fitting decision tree model decision. Overfitting is one of the major problems for our model in machine learning. If model is overfitted it will be poorly generalized to voice samples and recorded behaviour. To avoid decision tree from overfitting we remove the branches that make use of features having low importance**.** This method we call as Pruning or post-pruning. This way we will reduce the complexity of tree, and hence improves predictive accuracy by the reduction of overfitting. To reduce disturbance in accuracy result, avoid overfitting.
- *Support Vector Machine (SVM):* SVM are classifier and pattern recognition and classification problems [6]. It is hoped that the proposed method can help to detect the speech from the data set of disease. For Example, we will use our model for prediction diabetes with behaviour record such as sugar level and blood pressure monitoring. SVMs have a distinctly different modelling strategy in the detection of voice impairments problem, compared to other methods found in the literature. Noise parameters has been applied for the detection of voice impairments with good performance.
- *Naive Bayes:* Algorithm based on Bayes theorem with the assumption of independence between every pair of features. This algorithm requires a small amount of training data to estimate the necessary parameters. Naive Bayes classifiers are extremely fast compared to different other methods. Naive Bayes Classifier is mostly used to classify the audio signal into different emotions which will help us to measure certain behaviour of each individual sample in the model. Voice signal is random signal so we must predict the future sample and Naive Bayes Classifier is totally probability-based classifier therefore in voice analysis for in-depth analysis, we will be using Naive Bayes classifier. In the voice signal for recognition of signal classifier require large number of datasets. The advantage of Naive Bayes classifier is that it recognizes the signal with minimum dataset and our model consist of small dataset for accurate result.

We can use CART in our model where it is defined as Classification and Regression Trees (CART or just decision trees). Construct a binary tree from the training data. Split points are chosen greedily by evaluating each attribute and each value of each attribute in the training data to minimize the cost function. To split the dataset and make the model cost efficient we can use CART.

2) Regression: We can use regression analysis to make predictions and determine whether they are both unbiased and precise. Regression model is defined to predict the output values based on input value from the dataset in the system. The algorithm builds a model on the features of training data and using the model to predict the value for new data. We can use regression for prediction. Using regression to make predictions does not necessarily involve predicting for the future. Instead, we can predict the record sample given specific values of the behaviour pattern. For our example, we will use one sample at a time to predict the behaviour pattern. We have measured both variables at the same point in time. In our model we will use the regression model to predict body fat percentage based on body mass index (BMI). I have collected these data for a study with our fellow friends of our class. The variables we measured include height, weight, and body fat. We have calculated the BMI using the height and weight measurements. Measurements of body fat percentage are among the best. We want to use BMI to predict body fat percentage. If we can use our BMI to predict our body fat percentage, that provides valuable information more easily and cheaply in medical technology. We have a valid regression model that appears to produce unbiased predictions and can predict new observations nearly as well as it predicts the data used to fit the model. We are trying to make possible to use the regression equation and calculate the predicted values ourselves. However, we will use statistical software to do this for model. Not only is this approach easier and more accurate, but we will also have it calculate the prediction intervals so we can assess the precision.

Now we will undergo different types of Regression to reduce error while working on the model they are:

a) *Mean Absolute Error:* MAE is the sum of the absolute differences between predictions and actual values. This gives an idea of how wrong the predictions can be. The measure gives an idea of the magnitude of the error, but no idea of the direction for example if we are over or under predicting.

b) *Mean Squared Error:* MSE is defined as mean absolute error in that it provides a gross idea of the magnitude of error. Taking the square root of the mean squared error converts the units back to the original units of the output variable and can be meaningful for description and presentation.

c) *R Squared Metric:* This provides an indication of fit of a set of predictions to the actual values. In statistical terms this measure is called the coefficient of determination. This is a value between 0 and 1 for no-fit and perfect fit, respectively.

The above Regression matrix helps us to reduce error in our model for smooth processing.

Now, some of the Regression model we will be using in our disease prediction model are:

- Simple linear regression is a statistical method that allows or enable users to summarise and study relationships between two continuous variables. Linear regression is a linear model wherein a model that assumes a linear relationship between the input variables (x) and the single output variable (y). This algorithm can be used to analysis voice signal in terms of X and Y input according to specific behaviour pattern to be tested.
- Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model measured as the sum squared value of the coefficient values. This model we can use with Linear Regression.

IV. ACCURACY CHECK

Precision and accuracy check are termed use to describe method that measure, predict, or estimate. In all this cases there are some parameters called true value to provide measure value so that the model should be true as possible. So, the ultimate objective is to minimize the error and go for highest percentage for accuracy and precision check. Therefore, precision and accuracy are ways of describing the error between value among themselves.

To check the accuracy of our model and determine whether it can be used at an industrial/commercial level or can be classified as a successful model based on various accuracy scores. We must check each algorithm's accuracy to see how the model performs in response to the test dataset and how correctly can it predict the new data.

The most used accuracy metric is used in our model i.e., the confusion matrix. It can be also found in classification report.

A. Classification Report

To calculate the accuracy and precision of the model that we have curated, we designed a reporting format for all algorithms. Each of these algorithms have been run through this result displaying their accuracy, precision, F1-Score, and confusion matrix. The report significantly reduces time to individually display metrics and rather display them as a collective. The report is extremely useful to judge performance of all algorithms used. The report is generated for training data as well as testing results.

```
def print_score(clf, X_train, y_train, X_test, y_test, train=True):
    if train:
        pred = clf.predict(X_train)
        print("Train Result:\n================================================")
        print(f"Accuracy Score: {accuracy_score(y_train, pred) * 100:.2f}%")
        print("------------------------------------------------")
        print("Classification Report:", end='')
        print(f"\tPrecision Score: {precision_score(y_train, pred) * 100:.2f}%")
        print(f"\t\t\tRecall Score: {recall_score(y_train, pred) * 100:.2f}%")
        print(f"\t\t\tF1 score: {f1_score(y_train, pred) * 100:.2f}%")
        print("------------------------------------------------")
        print(f"Confusion Matrix: \n {confusion_matrix(y_train, pred)}\n\n")

    elif train==False:
        pred = clf.predict(X_test)
        print("Test Result:\n================================================")
        print(f"Accuracy Score: {accuracy_score(y_test, pred) * 100:.2f}%")
        print("------------------------------------------------")
        print("Classification Report:", end='')
        print(f"\tPrecision Score: {precision_score(y_test, pred) * 100:.2f}%")
        print(f"\t\t\tRecall Score: {recall_score(y_test, pred) * 100:.2f}%")
        print(f"\t\t\tF1 score: {f1_score(y_test, pred) * 100:.2f}%")
        print("------------------------------------------------")
        print(f"Confusion Matrix: \n {confusion_matrix(y_test, pred)}\n")
```

Figure 1 : Code Snippet for Classification Report

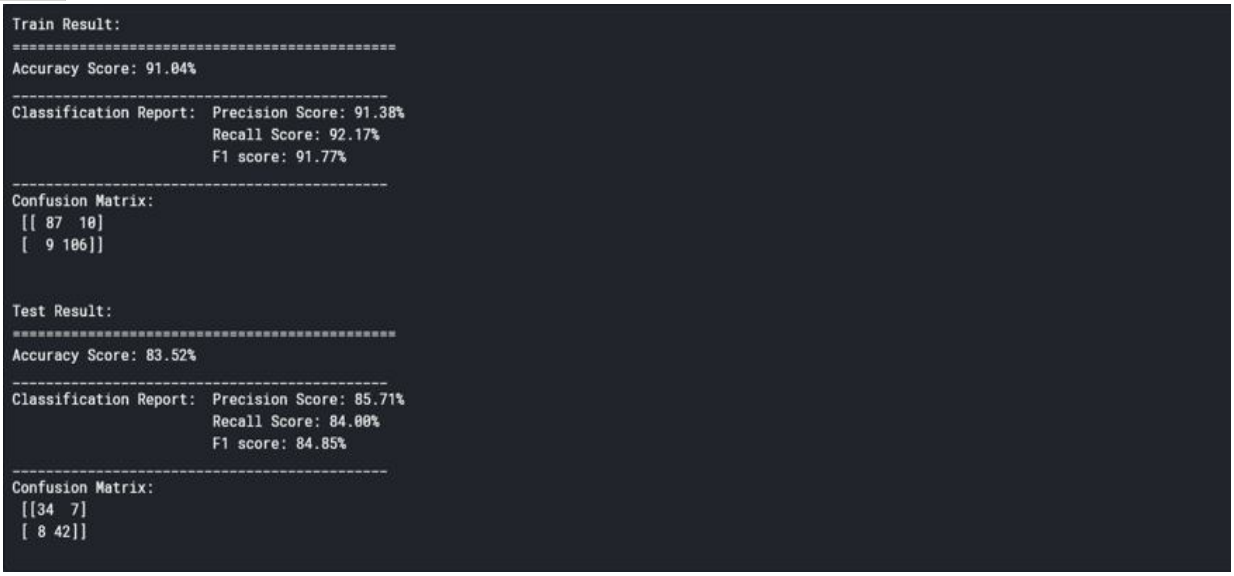


Figure 2 : Classification Report Example

B. Confusion Matrix

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.

1) The target variable has two values: **Positive** or **Negative.**
2) The columns represent the **actual values** of the target variable.
3) The rows represent the **predicted values** of the target variable.

- True Positive (TP) - predicted value matches the actual value i.e., positive was predicted as positive by the model.
- True Negative (TN) - predicted value matches the actual value i.e., negative was predicted as negative by the model.
- False Positive (FP) - Type 1 error - predicted value was falsely predicted i.e., negative value is predicted positive by the model.
- False Negative (FN) - Type 2 error - predicted value was falsely predicted i.e., positive value was predicted as negative value by the model.

C. Precision

Precision talks about how precise/accurate your model is out of those predicted positive, how many of them are actual positive. Precision is a good measure to determine when the costs of False Positive is high.

Precision = (True Positive)/ (True Positive + False Positive)

D. Recall

Recall calculates how many of the Actual Positives our model capture through labeling it as Positive (True Positive). Applying the same understanding, we know that Recall shall be the model metric we use to select our best model when there is a high cost associated with False Negative.

Recall= (True Positive)/ (True Positive + False Negative)

E. F1 Score

F1 Score is a measure used to see if we need to seek a balance between Precision and Recall and there is an uneven class distribution (large number of Actual Negatives).

F1 Score= 2 * ((Precision*Recall)/ (Precision + Recall))

V. CONCLUSIONS

This paper proposes a high accuracy, patient-oriented, economical, and less time-consuming prediction model for detecting disease in patients or symptoms at an early stage. Although, this elementary area of medical engineering, which deals with a variety of procedures for disease-related voice diagnosis, is now attracting the attention of many researchers and health care giants. This also proves that it is extremely necessary to prevent mass outbreak of diseases by prediction an early onset and mass testing of patients in a hassle-free way [7]. The proposed model strives to do exactly that. It will also help millions of individuals in remote locations to have at least basic access to diagnosis and preventive healthcare which can be implemented via the medium of smartphones. It can help increase life expectancy by screening diseases in their early onset and thus inducing an early diagnosis. It is justified by the fact that it has the capacity to save many precious human lives. If an entire locality is screened for asthma for example, then we will know that there is something wrong with the air in the locale and uproot the cause. Thus, we conclude with a proposed general model that aids in the advancement of voice diagnosis systems. As it seems now, the process of evolution of medical technology has already begun and the way it is nurturing the humankind, looks very promising for the future generations.

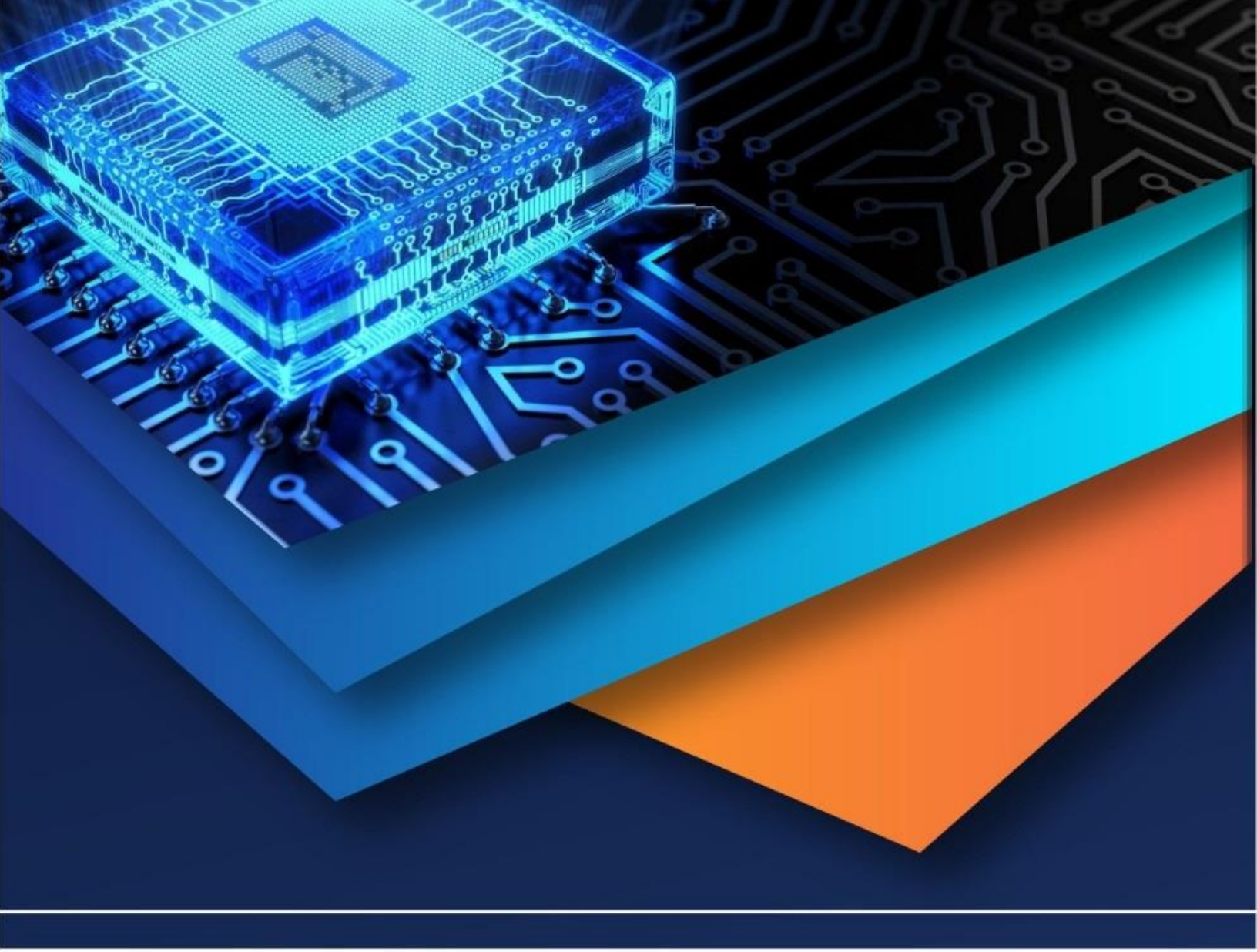
REFERENCES

[1] Wroge, T.J., Özkanca, Y., Demiroglu, C., Si, D., Atkins, D.C. and Ghomi, R.H., 2018, December. Parkinson's disease diagnosis using machine learning and voice. In 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB) (pp. 1-7). IEEE.
[2] Zhang, L., Chen, X., Vakil, A., Byott, A. and Ghomi, R.H., 2019. DigiVoice: Voice biomarker featurization and analysis pipeline. arXiv preprint arXiv:1906.07222.
[3] Hao, J. and Ho, T.K., 2019. Machine learning made easy: A review of scikit-learn package in Python programming language. Journal of Educational and Behavioral Statistics, 44(3), pp.348-361.
[4] Chen, P.H.C., Liu, Y. and Peng, L., 2019. How to develop machine learning models for healthcare. Nature materials, 18(5), pp.410-414.
[5] Bhavsar, K.A., Singla, J., Al-Otaibi, Y.D., Song, O.Y., Zikria, Y.B. and Bashir, A.K., 2021. Medical diagnosis using machine learning: a statistical review. Computers, Materials and Continua, 67(1), pp.107-125.
[6] Bzdok, D., Krzywinski, M. and Altman, N., 2018. Machine learning: supervised methods.
[7] Chen, I.Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K. and Ghassemi, M., 2020. Ethical Machine Learning in Healthcare. Annual Review of Biomedical Data Science, 4.

BIOGRAPHY

Ajay Dabas is a 3rd year undergraduate student pursuing Bachelor Of Technology with majors in Information Technology. Having done the courses and curriculum pertaining to Machine learning, he has done research and authored a paper in the domain of Machine Learning and has come up with a generalized study of disease prediction through voice analysis.

Pratyush Avi is a 3rd year undergraduate student of Bachelor Of Technology with majors in Information Technology. Having done the courses and curriculum pertaining to Machine Learning, he has done research and has authored the paper on voice analysis through disease prediction using Machine learning with python focusing on supervised Learning algorithm.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY