Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 7
July 2023

Early Parkinson's Disease Identification from Voice Recordings using Machine learning

***Jaya Singh[1], Dr. Ranjana Rajnish [1], Dr. Deepak Kumar Singh[1],**

[1]Amity University Uttar Pradesh, Lucknow

[2] Amity University Uttar Pradesh, Lucknow

[3] DK Infosolutions Pvt.Ltd

Abstract-In many experimental studies by applying supervised and non-supervised learning methods it is being observed that the human voice progressively worsens with time. Parkinson's Disease (PD) mostly affects the elderly more than 60 years old or more. PD is neurological disease that affects movements of the body and also causes tremors long time walking difficulty. And it is found by many researchers that voice disintegration is one of the earliest symptoms of PD diagnosis in which machine learning can be used to improve the accuracy of the diagnosis so that it can be detected early stage. Parkinson's Disease have been very attractive to researchers so there is a lot of voice dataset available on UCI Machine Learning repository. We used Parkinson's voice dataset which includes short sentences, numbers, words, and vowels which is compiled from PD patients and healthy subjects as well. Few researchers assert that concurrently capturing data from all participants using metrics for central tendency and dispersion is an effective method for creating PD prediction models. However, they failed to take into account that, in comparison to healthy people, PD patients have trouble pronouncing a small number of alphabets. Therefore, we will categorise each alphabet independently under this framework.. The final result would be majority vote from all the classifiers.

Keywords-Machine learning, disease prediction, parkinson's disease.

1 Introduction

James Parkinson, a physician, described Parkinson's disease (PD) as "shaking palsy" [1]. After Alzheimer's disease, it is the second most prevalent neurological disorders worldwide. There is no known cure for PD, which is an extremely progressive condition in which your brain gradually deteriorates over time.There are many symptoms of PD like walking difficulty, voice trembling, difficulty in writing and many more.

Speech disorder is accepted as the most common symptom of PD [7]. Many researchers have claim that about 90% people faces speech difficulty who are suffering from PD.One challenge in the search of the treament that can cure PD is that it can only be recognised once the symptoms are developed. And after the symptoms are developed, the disease progression starts which makes it difficult to cure.[4,5] Early detection and observation of the disease is very necessary to save the patient from losing their life.

Voice characteristics associated with Parkinson's disease patients include hoarseness, low volume, breathiness, high pitch, monotonocity, and occasionally the inability to sustain loudness on command[12]. Acoustic aspects of speech are packed with complicated information. Running speech testing and sustained phonation are two different types of speech. The patient is asked to speak a lengthy sentence in free-flowing speech that could disclose potential impaiment symptoms of a vocal condition. During prolonged phonation, the subject is asked to speak one vowel for as long as they can while maintaining the same pitch.

İn this paper we have proposed a framework which applies on each vocal samples and distinguishes how well are they performing with PD patients and also with the patients who does not have PD.

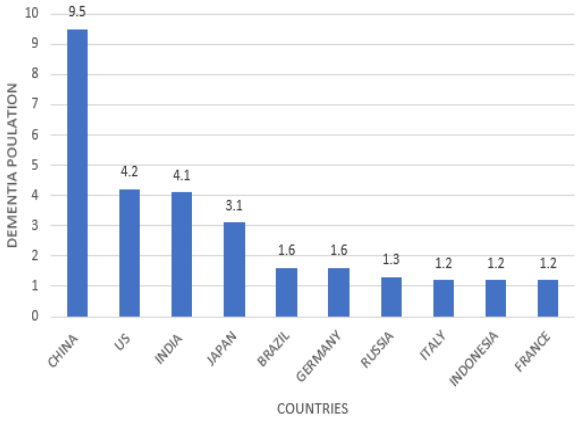


Fig. 1. Growth rate of neurodegenerative disorders in the world

1.1 Machine Learning

Machine Learning is a

1.2 Parkinson's Disease Dataset

The Parkinson's dataset contains biological voice measurements from 31 individuals, 23 of whom have the condition (PD). Each row in the table

corresponds to one of the 195 voice recordings from these people, and each column in the table designates a certain vocal measure ("name" column). The primary objective of the data is to identify between healthy people and people with PD, as shown by the "status" column, which is set to 0 for healthy and 1 for PD. The data is in ASCII CSV format. Each voice recording is represented by a single instance in a row of the CSV file. Each patient has about six recordings, and the first column of each tape lists the patient's name.

Patients with Parkinson's Disease now have relatively few diagnostic options. Blood tests and x-rays are used to diagnose various diseases, however in the case of PD, neither of these tests was helpful to the patient. Even though imaging technologies like DaTscan can help to shed light on the patient's health, they are too expensive for most people to afford and ineffective for making diagnoses. [7] Therefore, scientists and researchers are searching for different PD diagnosis options. Even without undertaking clinical tests, machine learning models are one of the most well-liked and effective methods for producing insightful comments.

Sustained phonation and running speech tests are the two vocal assessments that are generally regarded as the finest. In sustained phonation, individuals are instructed to speak one vowel at a time, while in running speech, they are required to recite the standard sentences so that we may hear any potential voice disorder symptoms.

In the past, many research have been conducted using datasets compiled from only a few different types of vocal exams, but we have drawn attention to a dataset that contains many recordings. The primary contribution of this work is the suggestion of a distinct classifier for each vowel separately rather than using a single classifier for all vocalists and the identification of the less effective vocal test using our suggested method.

2. Materials and Methods

We must follow a few procedures in order to effectively complete this experiment and develop our machine learning model for the early prediction of PD. We must first divide the dataset into several subsets so that our classifiers may be used on them. To achieve the optimum outcome for the provided dataset, the proposed model is afterwards used with cross-validation and different classifier algorithms. In the dataset, there are two different types of data: Parkinson's patients and healthy individuals. Parkinson's patients are classified as having the disease based on their health condition. Healthy subjects are classified as not having the disease.

Feature number	Feature name	Group
1	Jitter(local)	
2	Jitter(local,absolute)	Frequency
3	Jitter(rap)	Parameter
4	Jitter(ppq)	
5	Jitter(ddp)	
6	Number of pulses	
7	Number of periods	Pulse
8	Mean period	Parameters
9	Standard deviation of period	
10	Shimmer(local)	
11	Shimmer(local,dB)	
12	Shimmer(apq3)	Amplitude
13	Shimmer(apq5)	Parameters
14	Shimmer(apq11)	
15	Shimmer(dda)	
16	Fraction of Locally	
17	unvoiced Frames	Voicing
18	Number of voice breaks	Parameters
	Degree of voice breaks	
19	Median pitch	
20	Mean pitch	
21	Standard deviation	Pitch
22	Minimum pitch	Parameters
23	Maximum pitch	

24	Autocorrelation	
25	Noise-to-harmonic	Harmonicity
26	Harmonic-to-noise	Parameters

Table.1. Time frequency based featuresaccording to the given dataset of Parkinson's disease

2.1 Data seperation

In this phase, we separate the dataset into a subset that only contains voice note of the same category. For instance, all vocal sounds of type "a" belong to one subset, but all vocal sounds of type "o" belong to another. 40 samples from each subgroup are included in the dataset we are using to test our suggested model. Separation is necessary because combining all of a person's samples reduces the discriminating power of more descriptive tests and has a detrimental impact on the categorization outcomes.

2.2 Filter-based feature selection

The following stage required us to choose experiment columns that might be relevant to the issue. For ranking the attributes with greater weights for disease prediction, we choose the approach. The remaining information was then further organised for the following procedure, which involves putting it through one of the filter-based selection algorithms to determine which columns could be more pertinent to our machine learning model. In order to determine the weights of the characteristics that were strongly correlated, we utilised the Pearson's correlation algorithm using the SPSS software. After applying a filter-based feature selection, the table 2 displays the

2.3 MCFS AND A-MCFS

After applying the filter based feature selection method, we analysed that some of the features of the vocal tests were missing or not relevant enough for method. These vocal tests can be considered as unsuccessful vocal tests. So we have to do something so as to discriminate its effect on the final outcome of the result. There were two approaches which are chosen for the dealing of unsuccessful vocal tests. The first approch would be multiple classifier feature selection (MCFS) in which only prevalent features of the vocal test is used for the anlysis and the the other one is adjustable Multiple classifier Feature selection(A-MCFS). Table.2 shows the different features of the vocal tests and also which does not have any relevant feature figure 2 shows the frequency of each selected features.

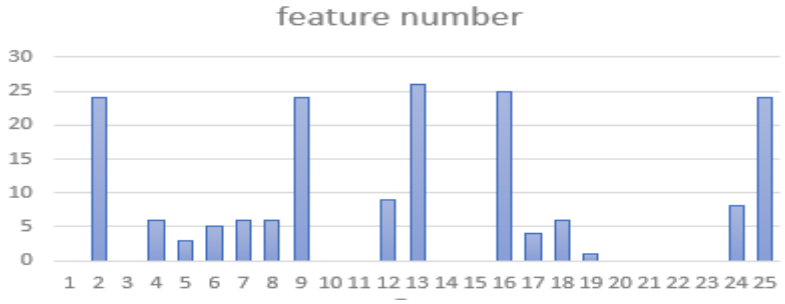


Figure.1. Frequency of each selected feature

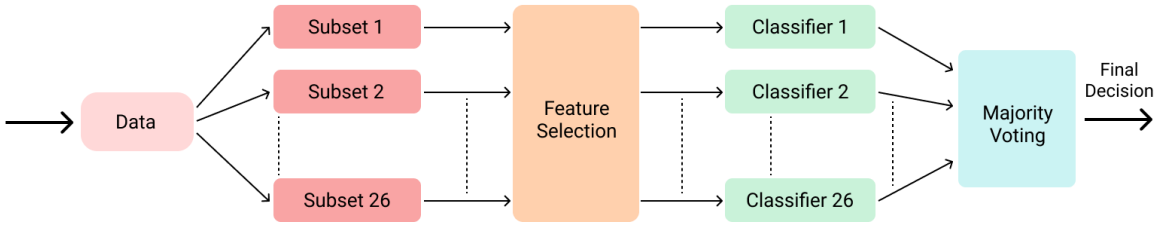
ID	VOCAL TEST	RELATED FEATURES
1	Vowel "a"	None
2	Vowel "o"	24
3	Vowel "u"	None
4	Number 1	1,2,3,4,5,24
5	Number 2	2,9,10
6	Number 3	17,19,23,25, 26
7	Number 4	1,2,3,4,5,10
8	Number 5	24
9	Number 6	None

10	Number 7	None
11	Number 8	9
12	Number 9	26
13	Number 10	None
14	Short sentence 1	None
15	Short sentence 2	25,26
16	Short sentence 3	4,10,25,26
17	Short sentence 4	1,2,3,4,5,10,26
18	Word 1	2
19	Word 2	None
20	Word 3	17, 19, 23, 25
21	Word 4	None
22	Word 5	None
23	Word 6	None
24	Word 7	None
25	Word 8	1,2,3,4,5,6,17,19,23,25
26	Word 9	24

Table. 2. Each vocal test and related features applying through filter-based feature selection

2.4 Classification and Majority voting

We must create a classifier for each subset after performing feature selection on each subset. There are roughly 26 subsets, so we must build 26 classifiers, each of which will determine the subset's class label. Which class a person belongs to is determined by the classifiers' majority vote. Each classifier determines whether or not the patient has PD. Additionally, the subject's status will be indicated as "1" if PD is predicted, or "0" if no signs of PD are present.



An Illustration of Proposed Method

2.5 Evaluating the Model

The evaluation of the model to demonstrate the success of the suggested strategy was the final and most important stage of our investigation. Accuracy, precision, AUC, and F1score are the parameters for this method's evaluation, and how these metrics work are as follows.

Accuracy

The most precise performance metric is accuracy, which may be used to quantify any performance score. One would assume that our

model would perform better the more precise we are. Yes, precision is a useful metric—but only if the values of the measures' negative and positive components are nearly equal. As a result, you should include other variables to evaluate the effectiveness of your model. Our model's result was 0.997, which indicates that it is about 99%

accurate.$Accuracy = \frac{correctpredictions}{totalpredictions}$

Precision

Precision is simply the accurate forecasting of positive results up to the highest number of results anticipated in any experiment. Any experiment's accuracy is reliant on its high positive rate. The model has 100% precision if there were no false positives. Our precision was 0.902, which is rather good.

$$Precision = \frac{truepositive}{falsepositive + truepositive}$$

F1 Score

The harmonic mean of Precision and Recall provides a better indicator of cases that were erroneously classified. It is also known as the balanced F-Score or the F Measure. The scoring system is as follows:

2*((precision*recall)/(precision+recall))

If you have an unequal class distribution, it performs better than Accuracy since it accounts for both false positives and negatives. F1 Score ranges from 0 to 1, with 1 being the worst.

3 Results

Once the data has been split, the given dataset is next refined by using the z-score normalisation procedure to each subgroup. Three classifiers—k-NN, SVM, and naive bayes—were applied to the preprocessed data. The k-NN classifier employed Euclidean distance and the values of k were 1 and 3, in contrast to the SVM classifier, which used linear and radical basis kernals. The results are shown in Table 4.

Vocal test ID	k-NN (k=1)	k-NN (k=3)	SVM (linear kernal)	SVM (RBF kernal)	Naive Bayes	Mean Accuracy
1	42.5	56	47.5	25	52.5	36.5
2	57.5	35	62.6	62.5	60	65
3	40	67	26.8	50	50	46.7
4	67.5.	48	70	65	45	64.6
5	67..5	47	60	67.5	65	60
6	62.5	70	62.5	72.5	25	64
7	52.5	57.5	67.5	67.5	54.7	55.9
8	57.5	67.6	65	67.5	45.4	58.4
9	47.5	60	60	50	35.6	67.3
10	62.5	62.5	55	55	55	56.9
11	42.5	62.5	75	72.5	65	52.5
12	50	60	65	65	70	54.1
13	40	45	52	60	45	65.9
14	52.5	45	50	60	72	64.4
15	57.5	60	72.5	72.5	72	64.1
16	50	55	72.5	72.5	65	59.7
17	60	57.5	72.5	60	22.7	37.8
18	45	45	67.5	67.5	65	57.5
19	40	37.5	63.5	40	60	45
20	52	55	42.5	62.5	55	53
21	47	40	63.6	55	37.7	56.4
22	47.7	35	40	60	60	37.8
23	42.5	55	53	55	65	63.7
24	35	35	35	42.6	55	44.3
25	62.5	62.5	57.8	67.5	50	55.6
26	55	62.5	55	57.4	45.5	50.4

Table.3. Results obtained from each classifier of different vocal tests

6 Conclusion

Several significant findings that can be incorporated into our work. First off, contrary to what we previously believed, it is possible to detect Parkinson's disease from voice patterns. Our model was successful since the data set was assembled from dependable sources and was previously cleansed. The suggested model aims to determine the results of each vocal test separately to prevent experimenter confusion. Researchers and other healthcare professionals can utilise this suggested model to make early diagnoses of Parkinson's disease. The approach suggested in this research can be applied to various models for a greater understanding of voice production in addition to Parkinson's disease patients. Numerous researchers are engaged in their work.

References

[1] G. R. Vásquez-Morales, S. M. Martínez-Monterrubio, P. Moreno-Ger, And J. A. Recio-García "Explainable Prediction of Chronic Renal Disease in the Colombian Population Using Neural Networks and Case-Based Reasoning" vol.7, 2019

[2] J. Ma, Y. Qiao, G. Hu, Y. Huang, A. K. Sangaiah, C. Zhang, Y. Wang, And R. Zhang "De-Anonymizing Social Networks With Random Forest Classifier" vol.6, 2018

[3] F. Zhao And Q. Tang "A KNN Learning Algorithm for Collusion-Resistant Spectrum Auction in Small Cell Networks" vol.6, 2018

[4] M. Li And K. Liu " Causality-Based Attribute Weighting via Information Flow and Genetic Algorithm for Naive Bayes Classifier" vol.7, 2019

[5] Z. Sun, K. Hu, T. Hu, J. Liu, And K. Zhu " Fast Multi-Label Low-Rank Linearized SVM Classification Algorithm Based on Approximate Extreme Points" vol.6, 2018

[6] L. Li, Y. Zhang, W. Chen, S. K. Bose, M. Zukerman , And G. Shen " Naïve Bayes Classifier-Assisted Least Loaded Routing for Circuit-Switched Networks" vol.7, 2019

[7] A. J. Stimpson And M. L. Cummings "Assessing Intervention Timing in Computer-Based

[8] Education Using Machine Learning Algorithms" vol.2, 2014

[9] M. Chen, Y. Hao, K. Hwang,L. Wang, And L. Wang "Disease Prediction by Machine Learning Over

[10] Big Data From Healthcare Communities" vol.5, 2017

[11] N. Zhi, B. K. Jaeger, A. Gouldstone, R. Sipahi, and S. Frank " Toward Monitoring Parkinson's Through Analysis of Static Handwriting Samples: A Quantitative Analytical Framework" vol.21, NO.2, 2017

[12] A. Agarwal, S. Chandrayan and S. S. Sahu " Prediction of Parkinson's Disease using Speech signal with Extreme Learning Machine" IEEE, 2019