Prognosis of Parkinson's Disease using Multidimensional Voice Program (MDVP) AnalysisData and Machine Learning

Ibrahim Javeed Khan
Dept. of Info. Science and Engineering
B.M.S College of Engineering
Bangalore, India

Roopa R
Dept. of Info. Science and Engineering
B.M.S College of Engineering
Bangalore, India

Nandan Hegde
Dept. of Info. Science and Engineering
B.M.S College of Engineering
Bangalore, India

Gururaja H S
Dept. of Info. Science and Engineering
B.M.S. College of Engineering
Bangalore, India

Rakshak Kunchum
DataWeave Pvt. Ltd.
Bangalore, India

***Abstract*—Moving forward in the field of applied medical sciences and healthcare, the focal point must be the disorders/diseases which develop with age and its early detection. Such diseases inflict pain on the patients, making prediction an important factor in healthcare. Parkinson's ailment (PD) is one ofthe brain's most commonly diagnosed neuro-degenerative illness.As the disease grows people with Parkinson's disease may have difficulties in walking as well as in speaking. So it is of utmost importance to detect PD early on. Symptoms are shown at a laterage making recovery almost impossible. Efficient and expansive models are required to diagnose PD at its onset stage. An issue of this magnitude requires large-scale automation of accurate and reliable diagnosis of Parkinson's.**

The objective of this work is to build an effective Machine Learning model. The proposed model predicts and diagnoses PD at a very early stage based on the results from Multidimensional Voice Program (MDVP) analysis. Also, our work focuses on enhancing the accuracy using Standard Vector Machines (Standard Vector Classifiers) in comparison with previous models andalgorithms.

***Index Terms*—Parkinson's, PD, Machine Learning, SVC**

I. INTRODUCTION

The Parkinson's Disease (PD) is a central nervous system illness that commonly results in tremors. The damage andloss of Nerve cells in the brain causes dopamine to drop, causing muscle twisting, slow movement and loss of balance. People may observe early waking, difficulty speaking, poor balance, jaw stiffness etc. There is no particular cure for Parkinson's and treatment options differ which include surgery and medications. Though Parkinson's is considered to be not fatal, the complexity of the disease could be dangerous. Parkinson's has been rated as the 14th cause of death in the United States by the Centers for Disease Control andPrevention (CDC) [1]. The stats tell that nearly 10 lakhs people in the US are living with PD and 60000 US citizens are diagnosed with PD each year. Scientists infer that about 9%to 16% of people with PD may have a change in one or more genes leading to the growth of the disease. PD is expected to rise to 1.2 million by the year 2030 [1]. Men are more likelyto have Parkinson's than women. Sadly medications alone costan average of 20 lakh rupees a year [1]. Notable figures with Parkinson's are Billy Connolly, Neil Diamond, Muhammad Ali, Michael J.Fox and many more [2].

Parkinson's disease does not have a cure, nevertheless medications, surgeries and other treatments can help in relieving some of the symptoms. Occupational therapy, speech-languagetherapy and physical therapy can help with balancing and walking problems. People with Parkinson's disease who exhibit symptoms like stiffness or tremor are given levodopa as atherapy. Surgeries such as deep brain stimulation, pallidotomyare also options available for the patients.

In the last few years, numerous data-driven methods have been used to improve the detection of Parkinson's [3]. The approach of expert systems and machine learning to data sets of numerous patients has led to high diagnostic accuracies. Thereupon the adaptation of machine learning models and bio markers can help in more accurate decision making.

A computer application called the Multi-Dimensional Speech Program (MDVP) can extract 33 features/parameters from a voice sample. In both scientific and clinical applications, the MDVP has the ability to provide rapid quantitative voice assessments. Using the obtained data, a model can be built to detect early Parkinson's. It should be acclaimed that theearly detection of Parkinson's disease helps in quick treatment and alleviates symptoms significantly. For the above topic, webuilt a better machine learning model using Standard Vector

Classifiers (Standard Vector Machines). The main aim is tomaximize accuracy and F1 - Score using speech data set.

II. METHODOLOGY

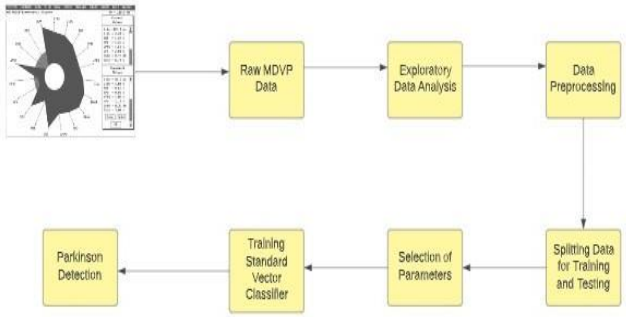

Fig. 1. General Overview of Methodology

A. EXPLORATORY DATA ANALYSIS

We analysed the raw data using histograms and correlation matrices to gain insights into the features. We understood the range by producing histograms for each and every one of the features.

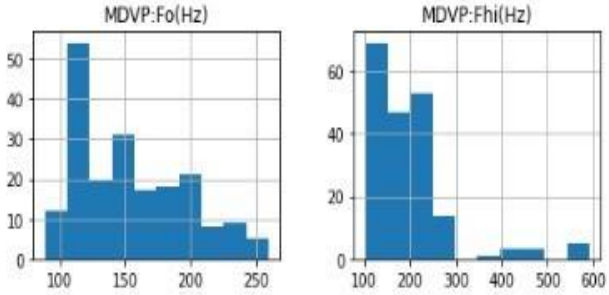

Fig. 2. Sample Dataset Histograms

These are the histograms for two of the parameters. A correlation matrix helps determine the correlation among the features. There are three types of correlations : positive, neg- ative and zero correlations.A positive correlation means that the variables change their values in tandem, while a negative correlation suggests that the variables alter their values inopposite directions. i.e, when one increases the other decrease and vice - versa. The variables are not linked at all and are completely independent in a zero correlation.

Observing only the status column of the data, we can note that HNR and fundamental frequencies have high negative correlations with the output category. A point to be made now is that these features influence the status column i.e, detection of Parkinson's Disease.

B. DATA PREPROCESSING

Firstly, the data was imported as a .csv file into the Jupyter Notebook. We used Pandas, NumPy, Mat- PlotLib, Seaborn and SKLearn libraries. The data consisted of 195 patient records, out of which 147 were diag-nosed with Parkinson's and 48 patients were healthy.

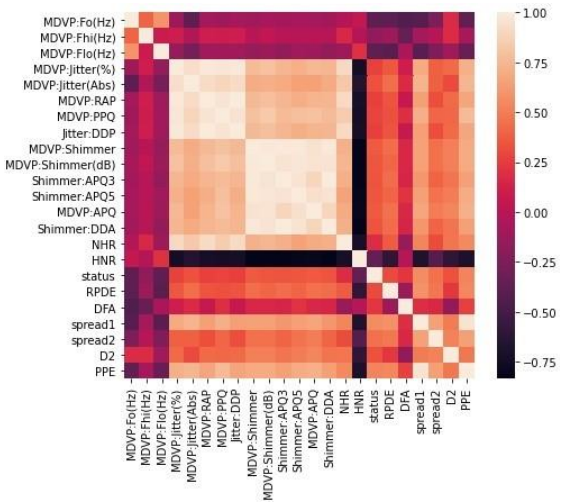

Fig. 3. Correlation Matrix

We built the model based on 23 parameters of Multidimen- sional Voice Program (MDVP) analysis data.

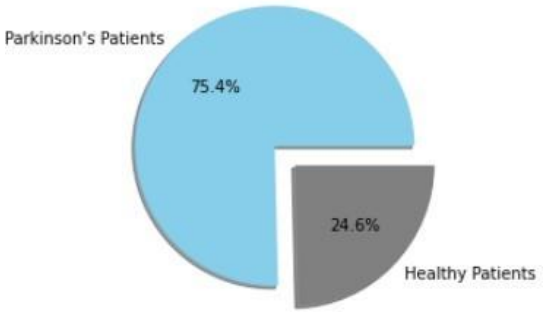

Fig. 4. Distribution of Healthy and Unhealthy Patients

The parametersused for the model are:

1) Vocal fundamental frequencies [4] [5]
   - Multidimensional Voice Program : Fo (Hz)
   - Multidimensional Voice Program : Fhi (Hz)
   - Multidimensional Voice Program : Flo (Hz)
2) Variations in fundamental frequency [4] [5]
   - MDVP : Jitter (%), MDVP : Jitter (Abs) - Feelings of nervousness ,absolute value in milliseconds. Abnormalities in vibration of vocal chords.
   - Perturbation - Perturbation is the deviation of a process from its current state or path, mainly dueto outside influence.
   - MDVP : RAP (%) - To compute RAP (Relative Average Perturbation), the difference is considered between a period and its average and the two neighbors of it, then divide it by the overall averageperiod.

- MDVP : PPQ (%) - The value of Pitch Period Perturbation Quotient is mainly influenced by breathy or hoarse voices.
- Jitter : DDP (%) - To get the value of Jitter : DDP, multiply the value of RAP by 3.

3) Variations in amplitude [4] [5]
   - MDVP: Shimmer - The value of shimmer is calculated by computing the average total difference between the amplitudes of successive periods and dividing it by the average amplitude.
   - MDVP : Shimmer (dB) - The log average of base
     - 10 of the differences between the consecutive periods of the amplitudes are multiplied with 20 to get the value of shimmer (in decibels). [6]
   - MDVP : APQ - Amplitude Perturbation Quotient is a measurement of variation in fundamental frequencies.
   - Shimmer : DDA - The average of the changes between amplitudes of successive periods is usedto calculate the value of shimmer (DDA). [7].
4) Ratio of noise to tonal components in voice [4] [5]
   - NHR: NHR - Noise to Harmonic's Ratio. [8].
   - HNR - Harmonic to Noise Ratio.
   - Status - Healthy patients are denoted by 0 and PDpatients by 1.
5) Nonlinear dynamical complexity measures [4] [5]
   - RPDE - Recurrence Period Density Entropy deter-mines the periodicity of a signal.
   - D2 - Also known as correlation dimension.

Using StandardScaler the values of the columns were scaled down to make plotting easier. Data standardization is a common process for many machine learning models as they might disrupt the learning process if the individual features do not resemble standard normally distributed data. Using StandardScaler the values are scaled down such that the mean of the total column is 0 and standard deviation is 1. Thisstep improves the learning process of the machine due tohigh resemblance in data values. This method scales downthe values, feature-wise.

C. MODEL BUILDING :

The dataset contains 195 rows and 24 columns. The machinelearning model classifies the patients based on 'Status' and theother features on one side. The column of 'Status' is set tozero for healthy patients and one for patients with Parkinson's.Test data was taken to be 20 percent of the total dataset using test train split method from sklearn library. We used standard vector classifiers (standard vector machines) as an estimator for this project, we used the following parameters for theclassifier:

1) C :- C is known as the regularization parameter. It basically means assigning a penalty whenever the valuescross the hyperplane. Regularizing ability is inversely proportional to the magnitude of C [9]. A low regularization value generalises the model without depending too much on the training data. The value of C must always be positive.
2) Kernel :- It specifies the type of kernel to be used. A kernel transforms the input given to the model into the required form [9]. Various types of kernels are used in standard vector such as linear, nonlinear, polynomial, gaussian rbf, sigmoid etc. are used in standard vector classifiers
3) Gamma :- The gamma parameters are the reciprocal of the radius of influence of support vectors. Support vectors are particular values selected from the sample.It shows how far one training sample influences themodel. A very high gamma will result in the modelbeing overfitted to the training dataset. Low value of gamma implies a general fit of the dataset.
4) Degree :- The degree parameter influences the pliability of the decision boundary [9]. Lower values of degree result in more rigid kernel boundaries.

The best parameters were selected by utilizing Grid-SearchCV. It searches for the pre - eminent combination over the specified features to yield the highest accuracy. We fed different Kernels and ranges of Gamma and C values were fed.High accuracy was easily achievable, so the main motive was to achieve substantial F1 - Score. Using the above mentioned method, we arrived to a conclusion that **Kernel=rbf, C=10, Gamma=0.1, Degree=1** were the best suited parameters. These features were used for cross - validation of the model. For a cross - validation count of 45, the average accuracy scorewas **95.2222%**.

D. MODEL EVALUATION:

To understand better how our model works, we used a confusion matrix [9] to demonstrate our results. Since our aim is to achieve a binary classifier, we used a 2x2 confusion matrix.

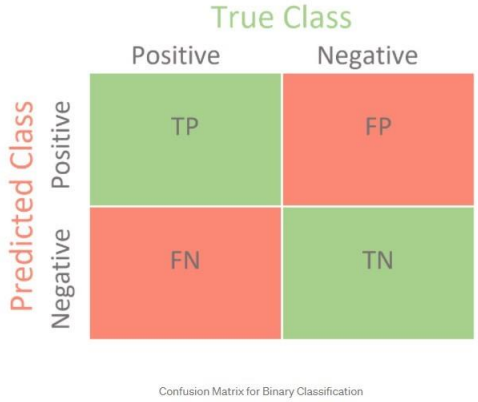

Fig. 5. General 2x2 Confusion Matrix

A 2x2 confusion matrix will have four blocks, namely

1) True Positive:- Number of instances where the positive value was predicted

2) False Positive:- Number of instances where the negative value was predicted as positive[Incorrect prediction]as positive[Correct prediction].
3) False Negative:- Number of instances where the positivevalue was predicted as negative[Incorrect prediction]
4) True Negative:- Number of instances where the negative value was predicted as negative[Correct prediction]

According to the definitions,a model must have denser TP and TN blocks whilst minimizing FP and FN.

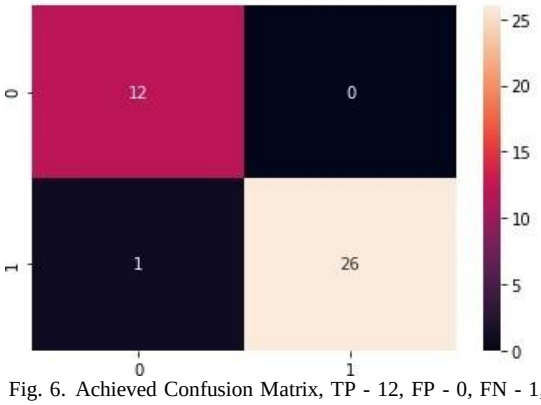

Fig. 6. Achieved Confusion Matrix, TP - 12, FP - 0, FN - 1, FN - 26

1) Precision:- Actual positive values divided by the classof predicted positive values

$$P\frac{T}{TP + FP}$$

2) Recall:- Actual positive values divided by all the positivevalues of the data set.

$$P\frac{T}{TP + FN}$$

3) F1-Score:- Produces a single result combining both precision and recall.

$$\frac{2(TP)}{2(TP) + FP + FN}$$

4) Support:- It is the quantity of samples for the particular performance measure.

	precision	recall	f1-score	support
0	0.92	1.00	0.96	12
1	1.00	0.96	0.98	27
accuracy			0.97	39
macro avg	0.96	0.98	0.97	39
weighted avg	0.98	0.97	0.97	39

Fig. 7. Model Classification Report

```
In [122]: metrics.accuracy_score(Y_predicted,Y_test)
Out[122]: 0.9743589743589743

In [123]: metrics.f1_score(Y_predicted,Y_test)
Out[123]: 0.9811320754716981
```

Fig. 8. Accuracy and F1 - Score

III. RESULTS

Hence,using the aforementioned parameters and classifier, we were able to achieve a highest accuracy of **97.4358 %**, Cross - Validation accuracy of **95.2222 %**, F1 score of **0.9811**and an AUC of **0.9615**. An area under the curve graph demonstrates how good a model is suited to be implemented. Itdescribes how best the model can distinguish between negative and positive values. In comparison to previous models, we have established a higher accuracy using standard vector classifiers.

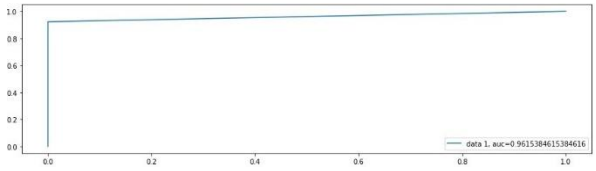

Fig. 9. AUC - ROC Graph

IV. CONCLUSION

A model was created to detect Parkinson's with high efficiency. In this comparative study, the usefulness of voice data set of Parkinson's Disease patients and Standard Vector Classifiers (Standard Vector Machines) was shown. The used techniques were compared with previous machine learningmodels. This model's future enhancements include using an additional Parkinson's Disease patients spiral drawing analysisdata set. This can improve the scalability of the model. A hybrid deep learning and machine learning architecture can be implemented to produce an enhanced model. Nonetheless, a better feature selection for our model can be made in the futureto make it viable to the public. This model can be used by hospitals or clinics by feeding the test results and diagnosing whether the patient could get Parkinson's Disease.

TABLE I
COMPARISON WITH PREVIOUS MODELS

Reference	Number of Parameters	Source of Parameters	Model Used	Accuracy Achieved (in %)
[10]	18	Voice	KNN	94.55
[3]	23	Voice, Image	Decision Trees	88
[11]	23	Voice	Stacking	92.2
[12]	23	Voice	NN	92.9
[13]	66	Voice	SVM	91.25
[14]	13	Voice	SVM	93.84
PD Detection with SVM	**23**	**Voice**	**SVM**	**97.43**

REFERENCES

[1] Jankovic, Joseph, and Eduardo Tolosa, eds. Parkinson's disease and movement disorders. Lippincott Williams & Wilkins, 2007.
[2] DeMaagd, George, and Ashok Philip. "Parkinson's disease and its management: part 1: disease entity, risk factors, pathophysiology, clinicalpresentation, and diagnosis." Pharmacy and therapeutics 40.8 (2015): 504.
[3] Jayaprakash, S., Nagarajan, M.D., Prado, R.P.D., Subramanian, S. and Divakarachari, P.B., 2021. A systematic review of energy management strategies for resource allocation in the cloud: Clustering, optimization and machine learning. *Energies*, *14*(17), p.5322.
[4] Little, M., Mcsharry, P., Roberts, S., Costello, D., Moroz, I. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Nature Precedings, 1-1.
[5] Hema, N., Sangeetha Mahesh, and M. Pushpavathi. "Normative data for Multi-Dimensional Voice Program (MDVP) for adults-A computerized voice analysis system." Journal of All India Institute of Speech and Hearing 28.1 (2009): 1-7.
[6] Benba, Achraf, Abdelilah Jilbab, and Ahmed Hammouch. "Hybridiza- tion of best acoustic cues for detecting persons with Parkinson's disease." 2014 Second World Conference on Complex Systems (WCCS). IEEE, 2014.
[7] Lahmiri, Salim. "Parkinson's disease detection based on dysphonia measurements." Physica A: Statistical Mechanics and its Applications 471 (2017): 98-105.
[8] Hüseyin Gürüler. "A novel diagnosis system for Parkinson's disease using complex-valued artificial neural network with k-means clustering feature weighting method".
[9] Ben-Hur, Asa, and Jason Weston. "A user's guide to support vector machines." Data mining techniques for the life sciences. Humana Press, 2010. 223-239.
[10] Almeida, Jefferson S., et al. "Detecting Parkinson's disease with sus- tained phonation and speech signals using machine learning techniques."Pattern Recognition Letters 125 (2019): 55-62.
[11] Younis Thanoun, Mohammed, and M. O. H. A. M. M. A. D. T. YASEEN. "A comparative study of Parkinson disease diagnosis in machine learning." 2020 The 4th International Conference on Advances in Artificial Intelligence. 2020.
[12] Das, Resul. "A comparison of multiple classification methods for diagnosis of Parkinson disease." Expert Systems with Applications 37.2 (2010): 1568-1572.
[13] Yaman, Orhan, Fatih Ertam, and Turker Tuncer. "Automated Parkinson's disease recognition based on statistical pooling method using acoustic features." Medical Hypotheses 135 (2020): 109483.
[14] Senturk, Zehra Karapinar. "Early diagnosis of Parkinson's disease using machine learning algorithms." Medical hypotheses 138 (2020): 109603.