Detection of Parkinson’s Disease using Machine Learning

Kartik Nagesh Pauskar 191IT223
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: kartikpauskar.191it223@nitk.edu.in

Akshay Sreekumar Nair 191IT103
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: a.191it103@nitk.edu.in

Shashikant Kumar 191IT249
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: shashikantkumar.191it249@nitk.edu.in

Shiv Kumar Dubey 191IT250
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: shiv.191it250@nitk.edu.in

Abstract—**Machine learning is a very important aspect of today’s developing technology, with applications not just limited to Artificial Intelligence field but in almost every field, which involves modern technology. The applications of Machine learning are endless, being it in super computers, automated vehicles, and even in life saving medical fields too. This paper, as the name suggests, is based on Detection of Parkinson’s Disease using Machine Learning. Using the applications of Machine learning, detecting this neurological disorder at an earlier stage could be life saving.**

I. Introduction

Parkinson’s Disease affects the brain and deteriorates the neural functionalities of the body. It causes uneven and rapid shaking of the body and may make the body stiff. Early detection of the Parkinsons’s disease is of utmost importance as treatment is possible at only early stage of the disease.

There are many Machine learning approaches in detecting the disease using Artificial Intelligence, using various forms of datasets like, voice samples of the patients, MRI scan data, handwritten data by the patients, and many more such datasets. In this project we consider the voice samples collected from the patients with and without the disease, and train the model to detect the presence of the disease in a person, which serves as an early step in the diagnosis of the person, which intern increases the probability of saving a life on early detection leading to early treatments from medical professionals.

In this project we compare multiple classifier algorithms and analyze their accuracy, namely; SVM classifier, Random forest classifier, Decision tree, Naive bayes and Neural network.

II. Literature Survey

The paper, Early Detection of Parkinson’s Disease Using Deep Learning and Machine Learning, discusses an innovative deep-learning technique for early detection of Parkinson’s disease. The results are obtained based on whether Rapid Eye movement, Cerebro-spinal fluid data, olfactory loss, etc. The research was conducted on a relatively small number of patients, including 183 individuals without the Parkinson’s disease and 401 patients with early Parkinson’s disease. The model achieved a high accuracy, 96.45% on average. [1]
The paper, Machine Learning and Deep Learning Approaches for Brain Disease Diagnosis: Principles and Recent Advances, shows a review on the recent deep learning and machine learning techniques in detecting four major brain diseases, Parkinson’s disease, Alzhimer’s, epilepsy and brain tumour. Twenty-two datasets are discussed and 147 recent articles on the same four brain diseases were analyzed. [2]
The paper, Deep Learning-Based Parkinson’s Disease Classification Using Vocal Feature Sets, proposes two frameworks on the basis of Convolutional Neural Networks to classify Parkinson’s Disease (PD) using sets of vocal features. The paper also proposes an CNN architectures. The paper has a relatively small number of instances in the dataset, where the performance evaluation is performed by LOPO CV. [4]
The paper, Reliable Parkinson’s Disease Detection by Analyzing Handwritten Drawings: Construction of an Unbiased Cascaded Learning System Based on Feature Selection and Adaptive Boosting Model, has developed four different machine learning models. The research alleviate the problem of biasedness, also uses random undersampling method to make the training process easy. The second problem considered is low rate of classification accuracy which has limited clinical significance. [5]

III. Problem Statement

Detection of Parkinson’s Disease using Machine Learning.

A. Objectives

- Build a model to detect the presence of Parkinson’s disease in an individual.
- Early Detection of the disease to facilitate clinical monitoring of elderly people and increase their life span.

Authors	Methodology	Merits	Limitations	Additional Details
Wu Wang, Junho Lee, Fouzi Harrou, Ying Sun	Parkinson's Disease Using Deep Learning and Machine Learning	Good balance in sensitivity and specificity	Analysis on relatively small dataset	Dataset involving data from 584 individuals
Protima Khan, Fazlul Kader, S. M. Riazul Islam , Aisha B. Rahman, Shahriar Kamal, Masbah Uddin Toha, Kyung-Sup Kwak	Machine Learning and Deep Learning Approaches discussion	Analyzed many different algorithmic approaches	-	-
Amin Ul Haq, Jian Ping Li, Muhammad Hammad Memon, Jalaluddin Khan, Asad Malik, Tanvir Ahmad, Amjad Ali, Shah Nazir, Ijaz Ahad, Mohammad Shahid	Machine-learning-based prediction system	Optimal accuracy achieved the best subset of the selected features	Analysis on relatively small dataset	Dataset from the repository of the University of Oxford
Hakan Gunduz	Two frameworks based on Convolutional Neural Networks to classify Parkinson's Disease using vocal features.	Second framework seems to be very promising, since it is able to learn deep features from each feature set via parallel convolution layers.	Small number of instances in the dataset	Dataset from UCI Machine Learning repository
Liaqat Ali, Ce Zhu, Noorbaksh Amiri Golilarz, Ashir Javeed, Mingyi Zhou, Yipeng Liu	Uses random undersampling method to make the training process easy	Yields Better performance than other similar cascaded systems that used six different machine learning models	Limited clinical significance	Data includes bradykinesia, dysphonia, rigidity, tremor and poor balance

TABLE I
SUMMARY OF LITERATURE SURVEY

- Compare multiple classifier algorithms and analyze their accuracy.

IV. METHODOLOGY

Parkinson's Disease was detected using the principles of Machine Learning from datasets which had the information of sound recordings of the participants. Five algorithms, SVM classifier, Random forest classifier, Decision tree, Naive bayes, Neural network, were used as the basis to facilitate this task.

A. Data-set information

The dataset, Disease Classification (DC) Dataset, comprises of data gathered from 252 patients, 64 healthy individuals and 181 patients with the Parkinson's disease. The participants had to recite the vowel /a/ for three repitions.
The data used in this data-set were gathered from 188 patients with Parkinsons and 64 healthy individuals. Researchers recorded the participants sustaining the phonation of the vowel /a/ for three repetitions. Speech signal processing algorithms including Time Frequency Features, Wavelet Transform based Features, Vocal Fold Features, Mel Frequency Cepstral Coefficients (MFCCs), and TWQT features were also applied to the speech recordings to extract clinically useful information for PD assessment. [11]

Multiple Sound Recording (MSR) Dataset, had the training data of 20 patients with the Parkionson's disease and 20 without Parkinson's disease. A variety of sound recordings were taken from the participants, expert physicians assigned each one a Unified Parkinsonâ€™s Disease Rating Scale (UPDRS) score.
The testing data were taken from 28 patients with the Parkinson's disease. The patients were asked to recite only the sustained vowels 'a' and 'o' for a period of three times, producing 168 recordings.
For the dataset, Telemonitoring (TE) Dataset, it comprised of data from 42 participants with early-stage Parkinson's disease. Two regression measurements were used : motor UPDRS and total UPDRS. A total of 16 voice measures were used. Each row of the dataset corresponds to one voice recording. There were around 200 recordings per patient.

B. Analyzing data-sets

On analyzing the shape of all four datasets, Disease Classification Dataset, Multiple Sound Recoring Training Dataset, Multiple Sound Recoring Testing Dataset, Telemonitoring Datasets, there were similarities aming many features of different datasets while some datasets consisted of a long list of features relatively.

C. Data visualization and covariance matrices

Create visualizations for the correlation matrix of each dataset.
Telemonitoring Dataset had many dimensions which were highly correlated. This was due to having multiple features that are related inseparably, for example the jitter and shimmer

features.
This could make the Telemonitoring dataset a good for principle component analysis.

Multiple Sound Recording Datasets(testing and training). This dataset had two parts namely, testing dataset and training dataset. The data was also collected from different sets of people, the training dataset was collected from a mixture of people having Parkinson's disease and people not having the Parkinson's disease, whereas the testing dataset comprised only of patients with Parkinson's disease. Both the datasets have their differences in the covariance matrices.

D. Dimensionality Reduction

Some features are highly correlated. Performing some dimensionality reduction using PCA and LDA.

1) PCA: PCA was performed on the dataset to take a look how many components we need to recover 99% of the variance. There was a reduction in the dimensions. For each of the datasets about half of the initial dimensions are needed to recover 99% of variance.

Dataset	Number of features	Reduced number of components
Multiple Sound Recording Training	29	17
Multiple Sound Recording Testing	28	12
Telemonitoring	22	11

TABLE II
PRINCIPAL COMPONENT ANALYSIS

2) LDA: There were labels on the dataset, thus we try LDA for dimensionality reduction.

E. Training Different Classifier Algorithms

1) SVM Classifier: Support Vector Machines deal with the classification, regression and outliers detection. This method is effective in high dimensional spaces. They are also effective in cases where the number of dimensions is greater than the number of samples given in the problem.

For this classifier, we analyzed it on several kernels with increasing complexity;

1) A basic linear kernel, assuming our data being linear separable;
2) A polynomial kernel, assuming our data being nonlinear;
3) A complex Radial Basis Kernel (RBF), with a feature space of infinite dimension.

Here, when we analyze the results we see that the polynomial kernel gives the most accurate results for the MSR training data. The accuracy is pretty much similar to and can be comparable to that of the Naïve Bayes classifier. We also can conclude that the accuracy rate is higher for the combined dataset of MSR and TE.

Kernal	MSR Train	MSR Train and TE
Linear	51.7%	100%
Polynomial	79.8%	96.4%
Radial Basis Kernel	54.1%	99.4%

TABLE III
PRINCIPAL COMPONENT ANALYSIS

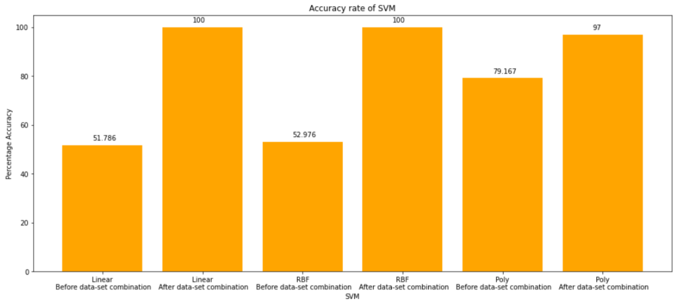

Fig. 1. SVM Accuracy

2) Random Forest Classifier: As we are dealing with the two kind of dataset in order to increase the predictive accuracy of the model the first dataset is MSR multiple sound recording and the second one is TE telemonitoring and implementing via random forest classifier algorithm. So to implement it we have considered following set of methodology:

1) We select the random sample from the mentioned dataset.
2) We use random forest algorithm to construct a decision tree for every sample. Then the algorithm itself will get the prediction result from every decision tree.
3) In this step voting will be performed for every predicted result.
4) At last it will select the most voted prediction result as the final prediction result.

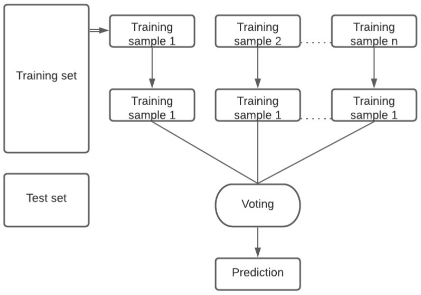

Fig. 2. Working of the Algorithm

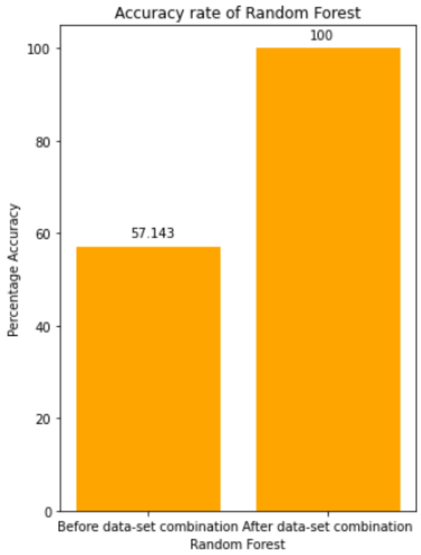

Fig. 3. Random Forest classifier Accuracy

3) Decision Tree: Decision Tree is a very powerful tool for the classification and prediction. The decision tree can be described as a flowchart like structure in which each of the internal nodes denote a test on an attribute. The branches represents outcome of the test. Each of the leaf nodes holds a class label.

The parameters default values controlling the size of the trees like min_samples, max_depth, etc. form a fully grown and unpruned trees. The numpy.argmax function is used to operate the predict method on the outputs of predict_proba.

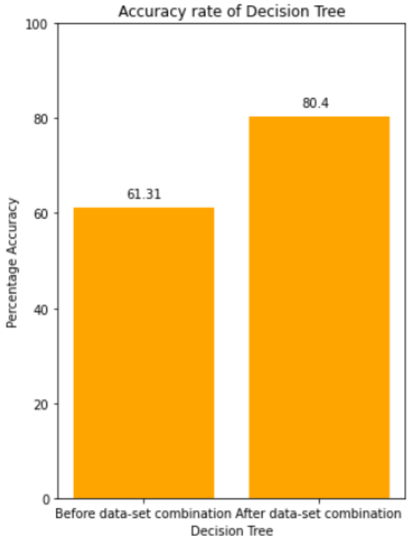

Fig. 4. Decision Tree Accuracy

4) Naive Bayes: **Gaussian Naive Bayes classifier algorithm-** It is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. As we know Naive Bayes are a group of supervised machine learning classification algorithms based on the Bayes theorem. It is a simple classification technique, but has high functionality. They find use when the dimensionality of the inputs is high. Complex classification problems can also be implemented by using Naive Bayes Classifier. Naive Bayes Classifiers are based on the Bayes Theorem. One assumption taken is the strong independence assumptions between the features. These classifiers assume that the value of a particular feature is independent of the value of any other feature. When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution.

The mathematical formula of Gaussian Naive bayes algorithm is like –

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(- \frac{(x_i-\nu_y)^2}{2\sigma_y^2}\right)$$

This approach used to create a simple model is to assume that the data is described by a Gaussian distribution with no co-variance (independent dimensions) between dimensions.

Now to implement datasets using Gaussian Naive bayes classifier algorithm, we have considered following set of methodology:

1) We are dividing dataset for training and testing purpose .
2) Now to train the dataset we select the random sample from the mentioned dataset, and Create an object classifier of name gnb.
3) Now we will call the class of naïve bayes classifier and fit the data into object classifier using gnb.fit() function, With this step our naïve bayes classifier is now trained.
4) Now we will test this model on testing datasets and print the accuracy in percentage.
5) We will call the object classifier to predict the accuracy of model using gnb.predict() function.
6) Then finally we will print the accuracy of naïve bayes classifier algorithm in percentage.

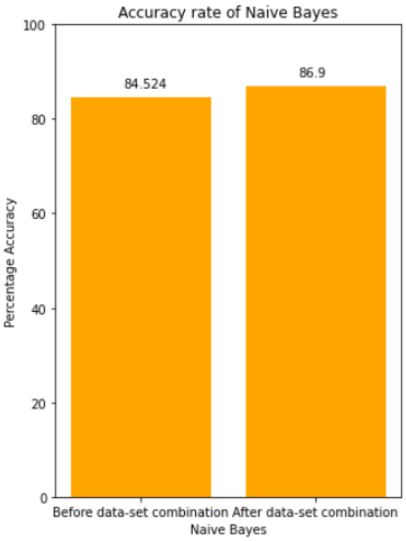

Fig. 5. Naive Bayes Accuracy

5) Neural Network: Neural networks, subset of machine learning, get their name from the way they mimick the working of biological neurons send signal to each other. When models developed with the help of neural networks and fine tuned for accuracy they help in classifying and clustering data at a high velocity.

MLPClassifier class implements a multi-layer perceptron(MLP) algorithm. It trains using Backpropagation. The Cross-Entropy loss function is used, which allows estimate probability by running the preict_proba method.

The model generated by the Neural network principle gave good results.

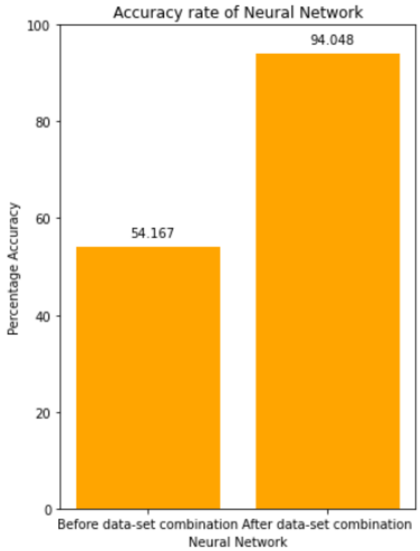

Fig. 6. Neural Network Accuracy

F. *Combining MSR and TE datasets to improve accuracy*

To improve the accuracy of the classifiers above, the Multiple Sound Recording training and Telemonitoring datasets were combined. Both these datasets had 13 overlapping features, for example Jitter and Shimmer. Next, perform predictions on the Multiple Sound Recording test dataset. After combining the datasets, an increase in the accuracy for each of the classifiers could be noticed.

V. Experimental Results and Analysis

A. *Correlation Matrices*

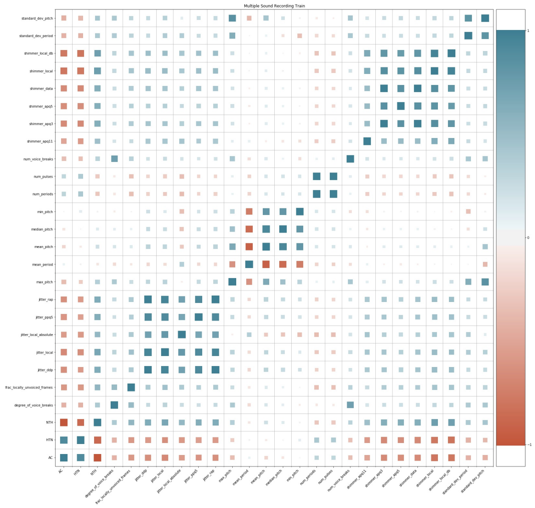

Fig. 7. Multiple Sound Recording Datasets for Training Correlation Matrix

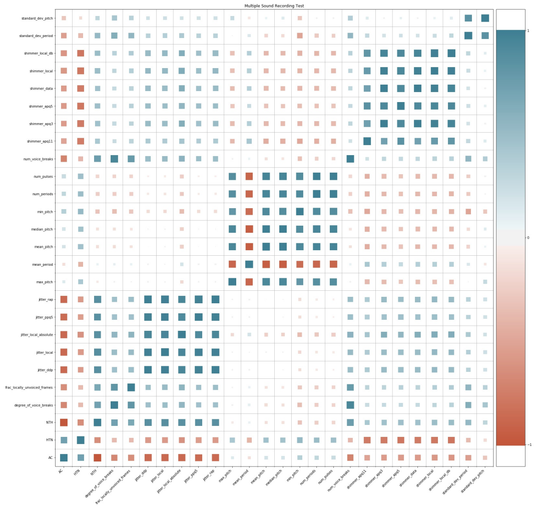

Fig. 8. Multiple Sound Recording Datasets for Testing Correlation Matrix

Multiple Sound Recording dataset was collected from different sets of patients. The training dataset was collected from a group of people with or without having Parkinson's disease. The testing dataset was collected from the data of patients who were diagnosed with Parkinson's disease. Both the datasets have their differences in the covariance matrices.

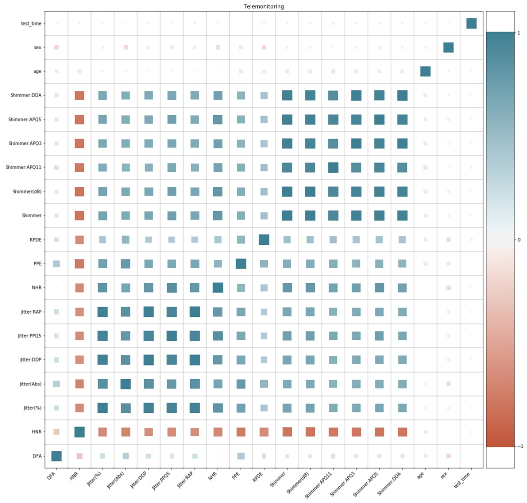

Fig. 9. Telemonitoring Dataset Correlation Matrix

Telemonitoring Dataset had many highly correlated dimensions. The reason for this was, having multiple features that are inseparably related, for ex: the jitter and shimmer features.

B. Dimensionality reduction using PCA and LDA

The PCA was carried out on the dataset to get the number of components needed to recover 99% of the variance. There was a reduction in the dimensions. When each dataset was considered, about half of the initial dimensions are needed to recover 99% of variance.
There were labels on the dataset, thus we try LDA for dimensionality reduction.

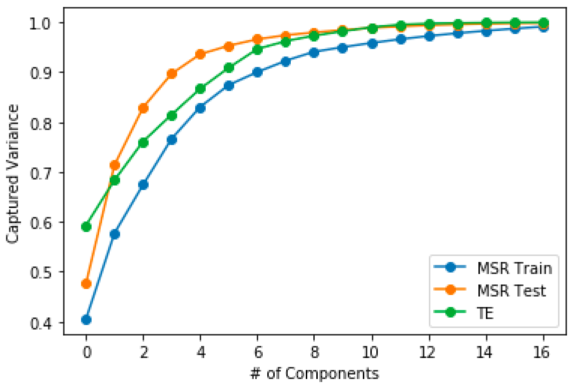

Fig. 10. PCA Variance

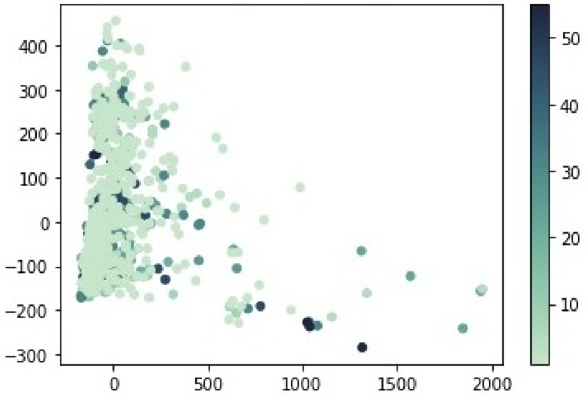

Fig. 11. Principal Component Analysis for MSR Training data-set

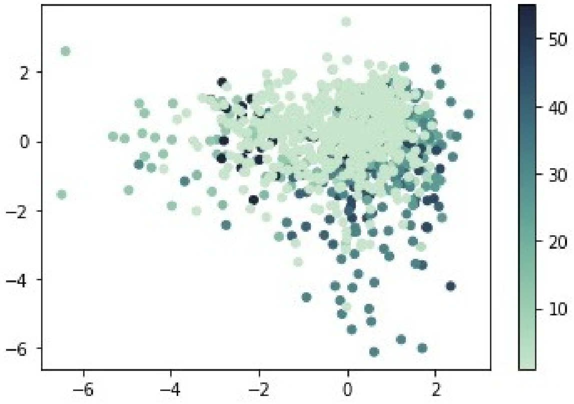

Fig. 12. Linear Discriminant Analysis for MSR Training data-set

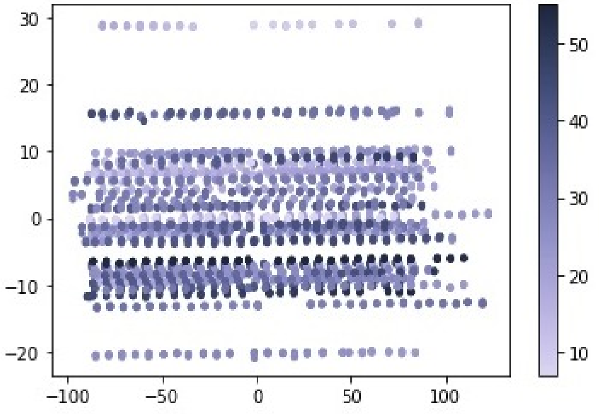

Fig. 13. Principal Component Analysis for Telemonitoring data-set

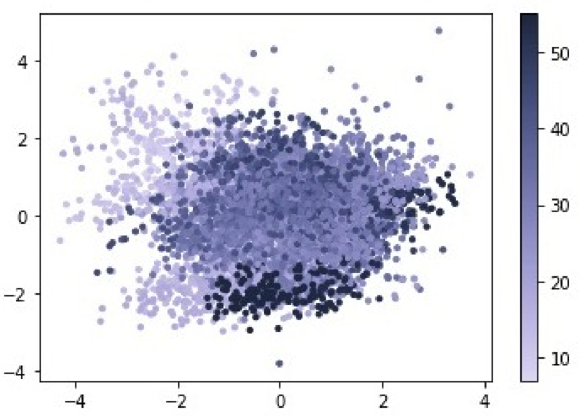

Fig. 14. Linear Discriminant Analysis for Telemonitoring data-set

C. Label Visualization

The Telemonitoring dataset and Multiple Sound Recording dataset have different UPDRS score distributions. But when their score range is analyzed, it comes out to be pretty similar to one another. Therefore, we merged these two datasets together by their common features, by which we get a new larger training dataset. Distributions of UPDRS scores:

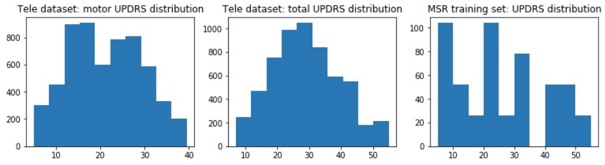

Fig. 15. Label Visualization

D. Accuracy

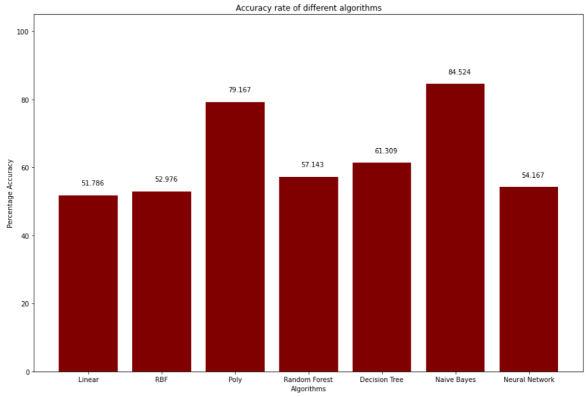

Fig. 16. Accuracy of algorithms before combination of MSR and TE data-sets

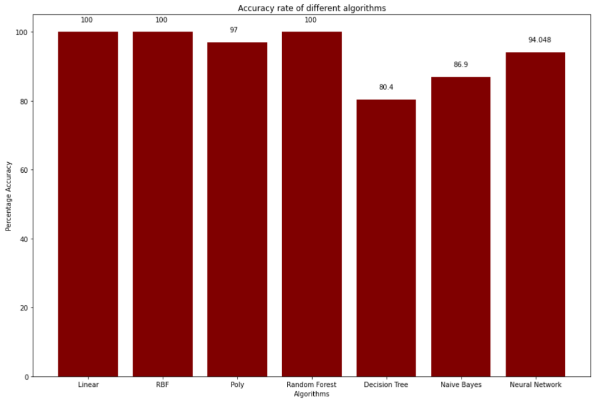

Fig. 17. Accuracy of algorithms after combination of MSR and TE data-sets

VI. Conclusion

In this project, we leverage several unsupervised techniques, like PCA, LDA, etc. and supervised techniques, like SVM, Naïve Bayes, etc. for predicting Parkinson's disease using the data of speech analysis. These methods provide highly accurate results.

But there are some limitations in the methods in its robustness. That is, all of the models are based on two dataset Telemonitoring (TE) and Multiple Sound Recording (MSR). The model is easily influenced by longer samples, pitches of numerous range and some other data variance. The Jitter, Shimmer and Harmonicity features are not sufficient enough to detect the Parkinson's disease, which was shown during the analsis of the dataset. But when considered the information of autocorrelation, range of the pitch, voice breaks, etc. makes the data more accurate.

The project being highly accurate, reliable and efficient in detecting Parkinson's disease, helps in early diagnosis of the deadly disease and serves the purpose of Artificial Intelligence and Machine Learning in the medical field.

Acknowledgment

A Special Thanks to Professor Anand Kumar M for giving valuable opinions and guiding us with this project.

VII. Individual Contribution

1) Kartik Pauskar:
   - Basic Research
   - Data-set collection
   - Data-set analysis
   - SVM classifier
   - Code-debugging
   - Report Writing
2) Shashikant Kumar:
   - Basic Research
   - Data-set collection
   - Data-set analysis
   - Random Forest classifier
   - Code-debugging
   - Report Writing

3) Akshay Sreekumar Nair:
   - Basic Research
   - Data-set collection
   - Decision Tree
   - Neural Networks
   - Code-debugging
   - Report Writing
4) Shiv Dubey:
   - Basic Research
   - Naive-Bayes
   - Data-set collection
   - Feature detection for data-set combination
   - Code-debugging

References

[1] Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning(2020)
Wu Wang; Junho Lee; Fouzi Harrou; Ying Sun
https://ieeexplore.ieee.org/document/9165732

[2] Machine Learning and Deep Learning Approaches for Brain Disease Diagnosis: Principlesand Recent Advances(2021)
PROTIMA KHAN1, MD. FAZLUL KADER 1, S. M. RIAZUL ISLAM , AISHA B. RAHMAN, MD.SHAHRIAR KAMAL, MASBAH UDDIN TOHA , AND KYUNG-SUP KWAK
https://ieeexplore.ieee.org/document/9363896

[3] Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson's Disease Using Voice Recordings(2019)
AMIN UL HAQ, JIAN PING LI, MUHAMMAD HAMMAD MEMON, JALALUDDIN KHAN, ASAD MALIK, TANVIR AHMAD, AMJAD ALI, SHAH NAZIR, IJAZ AHAD, AND MOHAMMAD SHAHID
https://ieeexplore.ieee.org/document/8672565

[4] Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets(2019)
HAKAN GUNDUZ
https://ieeexplore.ieee.org/document/8807125

[5] Reliable Parkinson's Disease Detection by Analyzing Handwritten Drawings: Construction of an Unbiased Cascaded Learning System Based on Feature Selection and Adaptive Boosting Model(2019)
Liaqat Ali; Ce Zhu; Noorbakhsh Amiri Golilarz; Ashir Javeed; Mingyi Zhou; Yipeng Liu
https://ieeexplore.ieee.org/document/8781770

[6] Deep Feature Extraction From the Vocal Vectors Using Sparse Autoencoders for Parkinson's Classification(2020)
Yanhao Xiong; Yaohua Lu https://ieeexplore.ieee.org/document/8963917

[7] Parkinson's Disease Detection Based on Running Speech Data From Phone Calls(2021)
Christos Laganas, Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, Sofia B. Dias, Sevasti Bostantzopoulou, Zoe Katsarou, Lisa Klingelhoefer, Heinz Reichmann, Dhaval Trivedi, K. Ray Chaudhuri, and Leontios J. Hadjileontiadis
https://ieeexplore.ieee.org/document/9556632

[8] Automated Detection of Parkinson's Disease Based on Multiple Types of Sustained Phonations Using Linear Discriminant Analysis and Genetically Optimized Neural Network(2019)
Liaqat Ali; Ce Zhu; Zhonghao Zhang; Yipeng Liu
https://ieeexplore.ieee.org/document/8861144

[9] Recent Advances in Computer-Aided Medical Diagnosis Using Machine Learning Algorithms With Optimization Techniques(2021)
Taki Hasan Rafi; Raed M. Shubair; Faisal Farhan; Md. Ziaul Hoque; Farhan Mohd Quayyum https://ieeexplore.ieee.org/document/9525093

[10] A Deep Learning Based Method for Parkinson's Disease Detection Using Dynamic Features of Speech(2021)
Changqin Quan; Kang Ren; Zhiwei Luo
https://ieeexplore.ieee.org/document/9321410

[11] Data-set
https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification