Multiclass based Novel Support Vector Machine for Parkinson disease detection.

Bali Devi[1], Bikkumalla Nikshep[2]
[1,2]SCIT, Manipal University Jaipur, India

Abstract: Objective: The Primary objective of this study is to address an early diagnosis of PD (Parkinson's disease) by the classification of characteristic features. **Methods:** All the features were analyzed and selected by using the feature selection algorithms to classify the subjects in 4 classes according to UPDRS (unified Parkinson's disease Rating Scale) score. We Used Various Machine Learning Classifiers such as Support Vector Machine(SVM), Logistic Regression(LR), Gaussian Naïve Bayes(GNB), K-Nearest Neighbors(KNN) and Random Forest(RF). **Results:** SVM has shown the best results for classification accuracy of 89% Using the PCA for Parkinson disease detection.

Keywords: Parkinson's Disease, Machine Learning, Classification, prediction.

1. INTRODUCTION

Parkinson's disease, a neurodegenerative disorder of the central nervous system which leads to difficulty in controlling some body functions. Because of the In-consistent flow of dopamine with the motor system, Parkinson's symptoms start gradually and get worse over the time. It is known that vocal issues of disorders can be estimated for early Parkinson's disease detection[1]. Little et al [2] have shown the measurement of the Parkinson's disease patients from healthy ones by utilizing the dysphonia indicators. Parkinson's disease condition can't be cured, however, treatment after diagnosis can allay symptoms significantly. So it is important to diagnose Parkinson's disease patients at the early stage[3]. In [4] the Parkinson's disease diagnosis is conducted by using the empirical tests and invasive techniques, but the results are less effective and also the accuracy is not satisfactory. Using machine learning algorithms diagnosis will provide better understanding for Parkinson disease[5]. SVM, LR and KNN been applied but this can be improved by applying more algorithms for accurate results of Parkinson disease detection[6]. A new method Random forest consisting of Haar wavelets used as a projection filter and integrated with logistic regression[7]. Fuzzy KNN method upon voice measurements was developed[8]. Hybrid intelligent system is proposed[9] in order to detect PD. Unsupervised methods were also used for Parkinson disease detection[10]. The accuracies are high, but the number of features were extracted made the computational time increased even though they used few features[11][12].The features obtained are voice based and these are easier than MRI based method[13][14]. The primary objective of this study is to address an early diagnosis of PD (Parkinson's disease) by the classification of characteristic features. All the features were analyzed and selected by using feature selection algorithms to classify the subjects in 4 classes according to UPDRS (unified Parkinson's disease Rating Scale) score. The rest of the paper discusses the Dataset in section 2, the Methodology in section 3, in section 4 discussed Result & Discussion, in section 5 short brief on Conclusion.

- *Bali Devi is currently working as assistant professor in Computer Software and Engineering at Manipal University Jaipur, India. E-mail: baligupta03@gmail.com*
- *Bikkumalla Nikshep is currently pursuing Bachelor's degree program in Computer Science and Engineering at Manipal University Jaipur, India. E-mail: nikshep05@gmail.com*

2. DATASET

The Dataset acquired in this study belongs to The Patient Voice Analysis[15]. The data used in this study were gathered from 620 individuals and 779 paired voice recordings with some individuals participated multiple times. All the participants were invited to pronounce the sustained vowel /a/ and hold it at their comfortable level. According to UPDRS score, we classified into 4 classes, the classes are healthy cases, early, intermediate and Advanced stages of parkinson's disease. The number of people in each class are 109 for healthy cases, 332 for early stage, 250 for intermediate stage and 49 for advanced stage.

3. METHODOLOGY

3.1. Data Pre-processing

After collecting the data, we will conduct several experiments to check any missing values. Missing values were removed and get_dummies() have been used for conversion of categorical variable into indicator variable. RDPE is used for determining the periodicity of a signal[16], PDA is a method used for power law scaling auto correlations which are known as non-stationary, Introduced a stationary signal for overcoming the scaling techniques[17]. PPE is common symptom to Parkinson disease people[18] a measure of dysphonia which is robust to many uncontrollable of stable pitch in voice frequency[2].

3.2. Principal Component analysis

All the parameters do not show the presence of Parkinson disease and cannot be distinguished them for healthy. So relevant features needs to be identified [19]. In this study we will use Principal component analysis, The purpose of PCA is to reduce the dimensionality into lower dimensionality in such a way that the variance of the data is maximized. Previous voice data assessments have shown good results for dimensionality reduce method[20]. The eigenvectors and eigenvalues of a covariance matrix are the "key" of a PCA. The eigenvectors find out the directions of the new feature space. Eigenvalues will give the variance of each data along the new axes. Covariance matrix is a matrix in which each element represents the covariance between the two features.
It is calculated using:

$$Cov(X,Y)=\sum(xi-\bar{x})(yi-\bar{y})N-1 \quad (1)$$

Table 1. Literature reviewed and key points.

Publish Date	Author	Methodology	Data set type	Dataset	Result
2013	M.Venkateswara Rao, DSVGK Kaladhar.	Early diagnosis of Parkinson's disease using various machine learning algorithms.	Voice Samples	Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado.	Random Forest has shown high accuracy with 90.26% and Naïve bayes has shown least accuracy with 69.23%
2019	F.M. Javed Mehedi Shamrat, Md. Asaduzzaman, A.K.M. Sazzadur Rahman, Raja Tariqul Hasan Tusher, Zarrin Tasnim	Implemented three supervised algorithms to improve Parkinson disease detection.	handwriting recordings (Static Spiral Test (SST), Dynamic Spiral Test (DST) and Stability Test on Certain Point (STCP)) are taken	Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University.	Accuracy of 100 % with Recall for SVM, and least accuracy of 40% with Precision for KNN.
2019	Jefferson S. Almeida	Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques	Voice Samples	Smartphone phonation and speech samples	The best feature set, using both phonation and speech achieved an accuracy of 94.55% and 92.94%, AUC of 87.84% and 92.40%, EER of 19.01% and 14.15%, respectively.
2012	Indrajit Mandal, N. Sairam	Robust methods are used for Prediction of Parkinson disease.	Voice Samples	Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado.	Accuracy of 100% percent for Linear LR with train and test samples. Using corrected t-test, accuracy of 96.75% with Additive LR
2019	Liaqat Ali	Linear Discriminant Analysis	Speech Samples	Department of Neurology in Cerrahpasa, Faculty of Medicine, Istanbul University.	Accuracy of 95% on training database and 100% on testing database using all the features of the dataset.
2009	Max little	measurement of the Parkinson's disease patients from healthy ones by utilizing the dysphonia indicators and applied Support vector machine	Voice Samples	Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado.	Accuracy of 91.4% using SVM.
2016	Gurpreet Singh	ML Based framework for Multi Class Diagnosis of Neurodegenerative Diseases (PCA, FDR, SVM)	MRI Images	ML Based framework for Multi Class Diagnosis of Neuro degenerative Diseases(PCA,FDR,SVM)	Yielded a classification accuracy of >95% to distinguish patients as binary classes. Whereas, for multi-class classification, an average classification of>85% has been achieved.

2018	Achraf Benba, Abdelilah Jilbab, Ahmed Hammouch	linear and nonlinear feature extraction techniques, principal component analysis(PCA), and nonlinear PCA.	Voice Samples	Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University	Accuracy of 87.5% using SVM.
2018	Zhennao Cai	Parkinson's Disease Diagnostic System Based on a Chaotic Bacterial Foraging Optimization Enhanced Fuzzy KNN Approach	Voice Samples	Oxford Parkinson's Disease data set	the CBFO-FKNN model performed best with an average accuracy of 96.97%, an AUC of 0.9781, a sensitivity of 96.87%, and a specificity of 98.75% when C(i) = 0.1
2016	Gunjan Pahuja, T.N.Nagabhushan	A Novel GA-ELM approach for Parkinson's disease detection using Brain Structural T1-weighted MRI Data	MRI Images	Parkinson's progression markers initiative (PPMI) database. http://www.ppmi-info.org.	With no of featured set to 40, SVM and ELM algorithms obtained testing accuracy of 81.15% and 90.97% respectively.

Table 2: Attributes description:

Field	Description	Type, format/units/values
call_timestamp	Date and time IVR voice call initiated	string, YYYY/MM/DD HH:MM:SS UTC +0000
recording_duration	Length of recording captured by IVR system	integer, seconds
callref	Unique reference number for voice recording	integer
audio_duration	Length of audio captured by IVR system during call	real, seconds
voice_indexstart	Starting audio sample index at which voicing is detected, if at all	integer, samples
voice_indexend	Ending audio sample index at which voicing is detected, it at all	integer, samples
voice_code	Quality of voice recording: 'ok' – (part of) recording usable, 'bad' – recording unusable (no features extracted)	string, 'ok'/'bad'
voice_usable_duration	Length of usable voicing detected	real, seconds
feature01	Median F0	real, Hz
feature02	Mean absolute F0 time derivative	real, Hz^2
feature03	Median absolute F0 time derivative	real, Hz^2
feature04	Mean absolute value of time derivative of RMS power	real, Hz
feature05	Median absolute value of time derivative of RMS power	real, Hz
feature06-feature19	Median cepstral coefficients 0-12 for entire voice recording	13 x real
feature20-feature32	Mean absolute time derivative of cepstral coefficients 0-12 across entire voice recording	13 x real
feature33	Recurrence period density entropy (RPDE) Hnorm	real
feature34	Detrended fluctuation analysis (DFA) scaling parameter alpha	real
feature35	Modified pitch period entropy (PPE)	real
feature36	Relative spectral power 0-500Hz	real
feature37	Relative spectral power 500-1kHz	real
feature38	Relative spectral power 1kHz-2kHz	real
feature39	Relative spectral power 2kHz-4kHz	real
user_id_hashed	Hashed PatientsLikeMe user ID	string
pvi_created_at	Date and time at which unique reference number for call was collected by PatientsLikeMe	string, YYYY-MM-DD HH:MM:SS.d
current_age	Age of participant when call was recorded	integer, years

sex	Sex of participant	string, 'M'/'F'/empty
years_since_first_symptom	Years since first Parkinson's symptoms detected by participant	integer, years
pdrs_date	Reference date at which PDRS applies	string, YYYY-MM-DD
pdrs_reported_at	Date and time at which PDRS was collected by PatientsLikeMe	string, YYYY-MM-DD HH:MM:SS.d
days_from_pvi_to_pdrs	Number of days elapsed between PDRS reference date (pdrs_date) and unique call reference number collected (pvi_created_at)	integer, days
pdrs_score	Scaled sum of all PDRS entries Q1-Q17 below	integer, 0-68
memory	PDRS Q1	integer, 0-4
hallucinations	PDRS Q2	integer, 0-4
mood	PDRS Q3	integer, 0-4
motivation	PDRS Q4	integer, 0-4
speech	PDRS Q5	integer, 0-4
saliva	PDRS Q6	integer, 0-4
swallowing	PDRS Q7	integer, 0-4
handwriting	PDRS Q8	integer, 0-4
cutting_food	PDRS Q9	integer, 0-4
dressing	PDRS Q10	integer, 0-4
hygiene	PDRS Q11	integer, 0-4
turning_in_bed	PDRS Q12	integer, 0-4
falling	PDRS Q13	integer, 0-4
freezing	PDRS Q14	integer, 0-4
walking	PDRS Q15	integer, 0-4
tremors	PDRS Q16	integer, 0-4
numbness	PDRS Q17	integer, 0-4
hoehn_yahr	Hoehn and Yahr stage at time of reporting PDRS	integer, 1-5
on_treatment_id	Whether participant was 'on' vs. 'off' medication when PDRS was reported	string, 'true'/'false'
calls_per_user	Number of calls made by this user	integer

Table 3. PLM-PVA PDRS/Hoehn and Yahr Questions:

Field	Value	Descriptive Value
memory	0, 1, 2, 3, 4	0 - 'No memory problems' 1 - 'Mild memory problems; I am often forgetful and can only partially remember some events.' 2 - 'Moderate memory problems that mean I get disoriented sometimes and can sometimes get confused easily.' 3 - 'Severe memory problems which mean I have difficulty remembering what day it is or where I am and often get confused.' 4 - 'Very severe memory problems; I sometimes forget where I am and what day it is; I have great difficulty making decisions or solving problems.'
hallucinations	0, 1, 2, 3, 4	0 - 'I never see things which are invisible to other people ('hallucinations').' 1 - 'I experience very vivid dreams.' 2 - 'Sometimes I see things which are invisible to other people but I know they are not real.' 3 - ' Sometimes I see things which are invisible to other people and I believe they are real.' 4 - 'I see things which are invisible to other people most of the time and I believe they are real.'
mood	0, 1, 2, 3, 4	0 - 'I do not get sad for long periods.' 1 - 'I sometimes get sad for longer than normal; i.e. more than a few days or a week.' 2 - 'I sometimes get sad for greater than a week at a time.' 3 - 'I get sad for extended periods during which I lose weight - cannot sleep - and have poor appetite.' 4 - 'I get sad for extended periods and have thoughts about suicide.'
motivation	0, 1, 2, 3, 4	0 - 'My levels of motivation are about normal.' 1 - 'I have become less assertive and more passive.' 2 - 'I have lost some of my initiative and am less interested in hobbies.' 3 - 'I have lost some of my initiative and am uninterested in my daily routine.' 4 - 'I am withdrawn and initiate very little activity.'

speech	0, 1, 2, 3, 4	0 - 'No; the way I speak has not changed.' 1 - 'Yes; the way I speak has changed but other people understand me without any problem.' 2 - 'Yes; the way I speak has changed somewhat and sometimes I have to repeat to make myself understood.' 3 - 'Yes; the way I speak has changed enough and frequently I have to repeat to make myself understood.' 4 - 'Yes; the way I speak has changed so much that other people has difficulty understanding me or do not understand me at all.'
saliva	0, 1, 2, 3, 4	0 - 'No; I have not noticed that I have excessive saliva and I do not drool.' 1 - 'Yes; I have noticed a slight increase in the amount of saliva and on occasion I drool at night on my pillow.' 2 - 'Yes; I have moderate excess of saliva and occasionally I drool during the day.' 3 - 'Yes; I have a marked excess of saliva and frequently drool during the day.' 4 - 'Yes; I drool so much that I have to carry a handkerchief at all times.'
swallowing	0, 1, 2, 3, 4	0 - 'No; I do not have difficulty swallowing and I do not choke.' 1 - 'Yes; I do have difficulty swallowing but I rarely choke.' 2 - 'Yes; I do have difficulty swallowing and occasionally choke.' 3 - 'Yes; I do have difficulty swallowing and I need soft food to be able to eat.' 4 - 'Yes; I am incapable of swallowing and I need nasogastric intubation or I have had a gastrostomy.'
handwriting	0, 1, 2, 3, 4	0 - 'No; I have not noticed changes in the way I write.' 1 - 'Yes; my handwriting is somewhat slower or my letter formation is smaller.' 2 - 'Yes; now my handwriting is moderately slower and my letter formation is smaller; but everything I write can be understood.' 3 - 'Yes; my handwriting is very altered and there are some words that cannot be understood.' 4 - 'Yes; my handwriting is deteriorated and most of the words cannot be understood.'
cutting_food	0, 1, 2, 3, 4	0 - 'No; I do not cut food more slowly and I have no difficulty managing my utensils.' 1 - 'Yes; I am somewhat slower and more clumsy than before but I am still capable of eating without help.' 2 - 'Yes; I am slower and clumsier than before and I need help cutting some foods.' 3 - 'Yes; someone has to cut my food but I can still eat on my own.' 4 - 'I have to be fed because I cannot do it on my own.'
dressing	0, 1, 2, 3, 4	0 - 'No; I do not find it difficult to get dressed nor am I slower than before.' 1 - 'Yes; I get dressed more slowly - but need little help.' 2 - 'Yes; I get dressed slower and sometimes need help buttoning my clothes - tying my shoes or getting my arm in the sleeves.' 3 - 'Yes; I need substantial help to get dressed - but I can still do some things on my own.' 4 - 'I have to be dressed by someone else.'
hygiene	0, 1, 2, 3, 4	0 - 'No; I have not slowed down when performing these activities.' 1 - 'Yes; I am a little slower dealing with my hygiene - but I do not need help.' 2 - 'Yes; I am slower and need help bathing and using the facilities.' 3 - 'Yes; I am slower and need help bathing - brushing my teeth - fixing my hair - and going to the bathroom.' 4 - 'I need help with everything and I wear a Foley catheter.'
turning_in_bed	0, 1, 2, 3, 4	0 - 'No I do not have difficulty turning in bed or fixing the blankets.' 1 - 'Yes; I am somewhat clumsy or slower turning in bed or fixing the blankets.' 2 - 'Yes; I am capable of turning in bed or fixing the blankets but with great difficulty.' 3 - 'Yes; I am capable of turning in bed but I need help to complete the task.' 4 - 'I am incapable of turning in bed or fixing the blankets without help.'
falling	0, 1, 2, 3, 4	0 - 'No; I have not fallen.' 1 - 'Yes; I have fallen - but rarely.' 2 - 'Yes; occasionally I have fallen but it happens less than once a day.' 3 - 'Yes; I fall on average once a day.' 4 - 'Yes; I fall every day and more than once.'
freezing	0, 1, 2, 3, 4	0 - 'No; I have not experienced freezing.' 1 - 'Yes; I have experienced freezing when I walk but on rare occasions. Or sometimes when I start walking I experience freezing.' 2 - 'Yes; occasionally I experience freezing.' 3 - 'Yes; frequently I experience freezing while walking and occasionally I have fallen because of it.' 4 - 'Yes; frequently I experience freezing while walking and frequently; I have fallen

		because of it.'
walking	0, 1, 2, 3, 4	0 - 'No; the way I walk and my arm movement has not changed.' 1 - 'Yes; the way I walk has changed but it is not a problem.' 2 - 'Yes; I have moderate difficulty while walking - but I do not need help.' 3 - 'Yes; I have great difficulty walking and need help.' 4 - 'I cannot walk alone or with help.'
tremors	0, 1, 2, 3, 4	0 - 'No; I do not have visible tremors.' 1 - 'Yes; occasionally I do have visible tremors.' 2 - 'Yes; I have moderate tremors that bother me.' 3 - 'Yes; I have intense tremors that interfere with some activities.' 4 - 'Yes; I have intense tremors that interfere with the majority of my activities.'
numbness	0, 1, 2, 3, 4	0 - 'No; I do not feel numbness tingling or discrete pain - that I can attribute to my Parkinson's disease.' 1 - 'Yes; on occasion; I do feel numbness - tingling - or discrete pain - that I can attribute to my Parkinson's disease.' 2 - 'Yes; frequently; I do feel numbness - tingling - or pain - that I can attribute to my Parkinson's disease.' 3 - 'Yes; frequently I feel painful sensations attributable to my Parkinson's disease.' 4 - 'Yes; I feel extreme pain attributable to my Parkinson's disease.'
hoehn_yahr	1, 2, 3, 4, 5	1 - Symptoms on one side 2 - Symptoms on both sides 3 - Problems with balance and walking 4 - Stand and walk with difficulty 5 - Cannot stand or walk independently

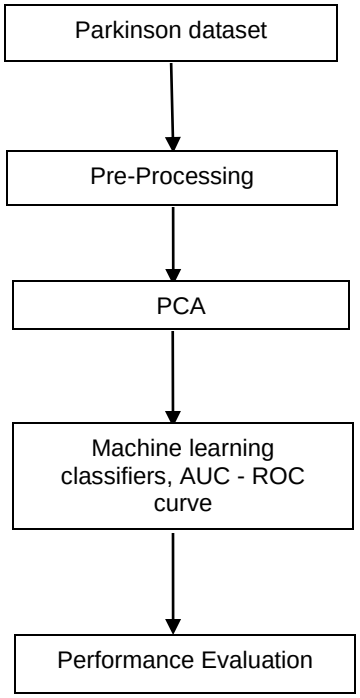

Fig 1. The Experimental Setup

3.3. Model Selection:

After PCA, Model selection is an important stage in building the Machine Learning Model for determination of the calculations. It involves choosing best Machine Learning Models. Accuracy, available resources and maintainability are many concerns while performing model selection beyond the model performance. AUC – ROC curve have also been implemented. Metrics such as Recall, Precision, f1- measure are used for evaluation.

$$\text{Accuracy} = (\mathcal{TP} + \mathcal{TN}) / (\mathcal{TP} + \mathcal{FP} + \mathcal{TN} + \mathcal{FN}) \quad (2)$$

$$\text{Recall} = \mathcal{TP} / (\mathcal{TP} + \mathcal{FN}) \quad (3)$$

$$\text{Precision} = \mathcal{TP} / (\mathcal{TP} + \mathcal{FP}) \quad (4)$$

$$\text{F1} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (5)$$

4. RESULT AND DISCUSSION

4.1. Support Vector Machine(SVM):

The table 4 represents the results obtained using SVM classification and selected PCA = 5, with 89% overall accuracy. Fig 2 represents the accuracy matrix for the classes. Fig 3 represents the ROC curve for the classes with AUC scores. In this model, Confusion matrix table:

1. Advanced class: There are 4 correctly classified, 2 are misclassified and considered as intermediate class.
2. Early class: There are 60 correctly classified, 5 are misclassified (2 as healthy class, and 3 as intermediate class).
3. Healthy class: There are 18 correctly classified, 6 are misclassified and considered as early class.
4. Intermediate class: There are 49 correctly classified, 4 are misclassified and considered as early class.

Table 4. Confusion matrix table for SVM.

	advanced	early	healthy	intermediate
advanced	4	0	0	2
early	0	60	2	3
healthy	0	6	18	0
intermediate	0	4	0	49

	precision	recall	f1-score	support
advanced	1.00	0.67	0.80	6
early	0.86	0.92	0.89	65
healthy	0.90	0.75	0.82	24
intermediate	0.91	0.92	0.92	53
accuracy			0.89	148
macro avg	0.92	0.82	0.86	148
weighted avg	0.89	0.89	0.88	148

Fig 2. Accuracy matrix for SVM.

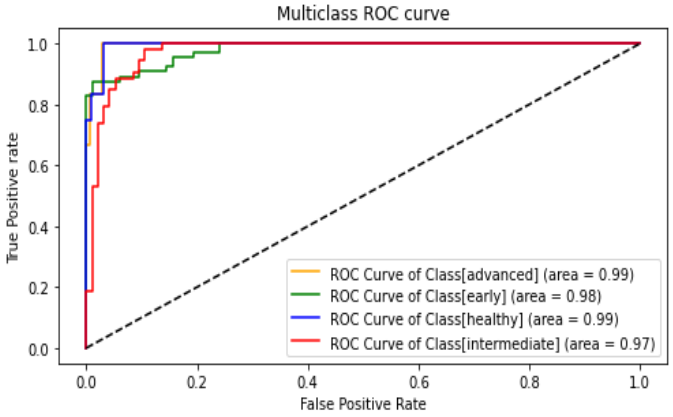

Fig 3. ROC curve for SVM.

4.2. Logistic Regression(LR):

The table 5 represents the results obtained using Logistic regression(LR) classification and selected PCA = 5, with 87% overall accuracy. Fig 4 represents the accuracy matrix for the classes. Fig 5 represents the ROC curve for the classes with AUC scores. In this model, Confusion matrix table:

1. Advanced class: There are 4 correctly classified, 2 are misclassified and considered as intermediate class.
2. Early class: There are 59 correctly classified, 6 are misclassified (4 as healthy class, and 2 as intermediate class).
3. Healthy class: There are 20 correctly classified, 4 are misclassified and considered as early class.
4. Intermediate class: There are 46 were correctly classified, 7 are misclassified (1 as advanced class, and 6 as early class).

Table 5. Confusion matrix table for LR.

	advanced	early	healthy	intermediate
advanced	4	0	0	2
early	0	59	4	2
healthy	0	4	20	0
intermediate	1	6	0	46

	precision	recall	f1-score	support
advanced	0.80	0.67	0.73	6
early	0.86	0.91	0.88	65
healthy	0.83	0.83	0.83	24
intermediate	0.92	0.87	0.89	53
accuracy			0.87	148
macro avg	0.85	0.82	0.83	148
weighted avg	0.87	0.87	0.87	148

Fig 4. Accuracy matrix for LR.

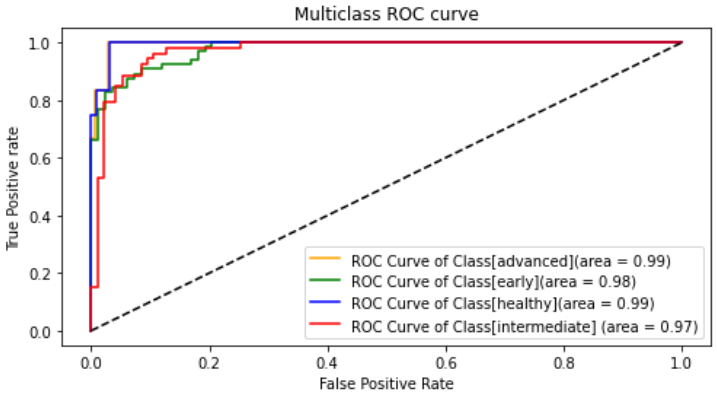

Fig 5. ROC curve for LR.

4.3. Gaussian Naïve Bayes(GNB):

The table 6 represents the results obtained using Gaussian Naïve Bayes(GNB) classification and selected PCA = 5, with 80% overall accuracy. Fig 6 represents the accuracy matrix for the classes. Fig 7 represents the ROC curve for the classes with AUC scores. In this model, Confusion matrix table:

1. Advanced class: There are 4 correctly classified, 2 are misclassified and considered as intermediate class.
2. Early class: There are 62 correctly classified, 3 are misclassified and considered as intermediate class.
3. Healthy class: There are 14 correctly classified, 10 are misclassified and considered as early class.
4. Intermediate class: There are 39 correctly classified, 14 are misclassified and considered as early class.

Table 6. Confusion matrix table for GNB.

	advanced	early	healthy	intermediate
advanced	4	0	0	2
early	0	62	0	3
healthy	0	10	14	0
intermediate	0	14	0	39

	precision	recall	f1-score	support
advanced	1.00	0.67	0.80	6
early	0.72	0.95	0.82	65
healthy	1.00	0.58	0.74	24
intermediate	0.89	0.74	0.80	53
accuracy			0.80	148
macro avg	0.90	0.73	0.79	148
weighted avg	0.84	0.80	0.80	148

Fig 6. Accuracy matrix for GNB.

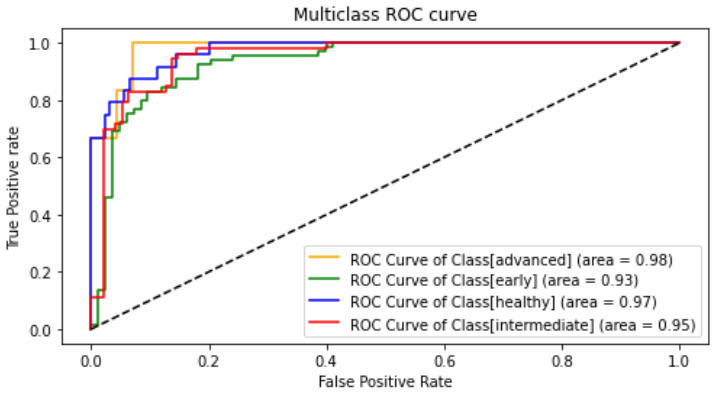

Fig 7. ROC curve for GNB.

4.4. K-Nearest Neighbors(KNN):

The table 7 represents the results obtained using KNN classification with k = 4 and selected PCA = 5, with 76% overall accuracy. Fig 8 represents the accuracy matrix for the classes. Fig 9 represents the ROC curve for the classes with AUC scores. In this model, Confusion matrix table:

1. Advanced class: There are 4 correctly classified, 2 are misclassified and considered as intermediate class.
2. Early class: There are 60 correctly classified, 5 are misclassified and considered as healthy class.
3. Healthy class: There are 10 correctly classified, 14 are misclassified and considered as early class.
4. Intermediate class: There are 38 correctly classified, 15 are misclassified (2 as advanced class, and 13 as early class).

Table 7. Confusion matrix table for KNN.

	advanced	early	healthy	intermediate
advanced	4	0	0	2
early	0	60	5	0
healthy	0	14	10	0
intermediate	2	13	0	38

	precision	recall	f1-score	support
advanced	0.67	0.67	0.67	6
early	0.69	0.92	0.79	65
healthy	0.67	0.42	0.51	24
intermediate	0.95	0.72	0.82	53
accuracy			0.76	148
macro avg	0.74	0.68	0.70	148
weighted avg	0.78	0.76	0.75	148

Fig 8. Accuracy matric for KNN.

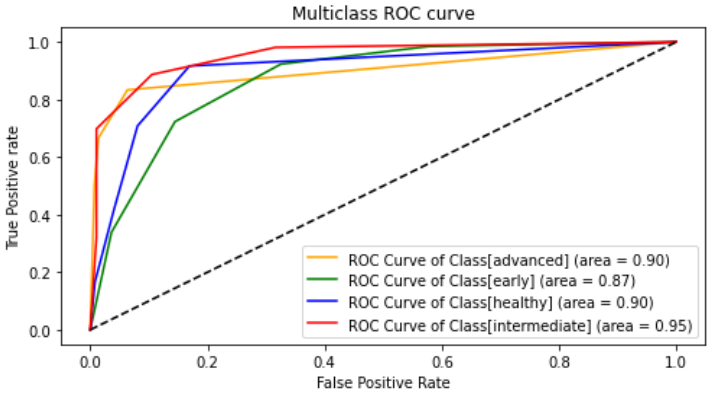

Fig 9. ROC curve for KNN.

4.5. Random Forest(RF):

The table 8 represents the results obtained using Random Forest(RF) classification and selected PCA = 5, with 82% overall accuracy. Fig 10 represents the accuracy matrix for the classes. Fig 11 represents the ROC curve for the classes with AUC scores. In this model, Confusion matrix table:

1. Advanced class: There are 4 correctly classified, 2 are misclassified and considered as intermediate class.
2. Early class: There are 60 correctly classified, 5 are misclassified (4 as healthy class, and 1 as intermediate class).
3. Healthy class: There are 16 correctly classified, 8 are misclassified and considered as early class.
4. Intermediate class: There are 42 were correctly classified,11 are misclassified (1 as advanced class, and 10 as early class).

Table 8. Confusion matrix table for RF.

	advanced	early	healthy	intermediate
advanced	4	0	0	2
early	0	60	4	1
healthy	0	8	16	0
intermediate	1	10	0	42

	precision	recall	f1-score	support
advanced	0.80	0.67	0.73	6
early	0.77	0.92	0.84	65
healthy	0.80	0.67	0.73	24
intermediate	0.93	0.79	0.86	53
accuracy			0.82	148
macro avg	0.83	0.76	0.79	148
weighted avg	0.83	0.82	0.82	148

Fig 10. Accuracy matrix for RF.

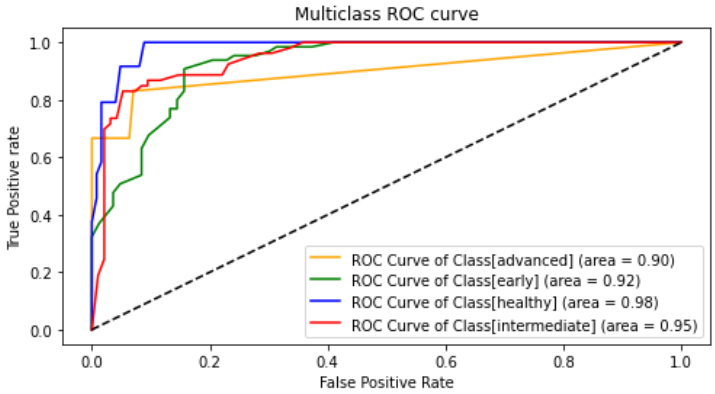

Fig 11. ROC curve for RF.

From all the performance measurements we can say, SVM has shown the best performance with 89% overall accuracy respectively.

5 CONCLUSION

We have applied various machine learning algorithms using PCA with multi classification dataset for the detection of Parkinson disease. SVM has shown the best performance with an overall accuracy of 89% and KNN has shown the worst performance with an overall accuracy of 76%. With this multi classification dataset, we can distinguish at which stage a person is, based on UPDRS score. This will be helpful for the medication purpose. In future, We intend to implement Deep Learning algorithms.

ACKNOWLEDGMENT

These Datasets were generated through collaboration between Sage Bionetworks, PatientsLikeMe and Dr. Max Little as part of the Patient Voice Analysis study (PVA). They were obtained through Synapse ID [syn2321745].

REFERENCES

[1] B. Harel, M. Cannizzaro, and P. J. Snyder, ‘‘Variability in fundamental frequency during speech in prodromal and incipient Parkinson’s disease: A longitudinal case study,’’ Brain Cogn., vol. 56, no. 1, pp. 24–29, Jun. 2004.

[2] Little, M.A., et al.: Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease. IEEE Transactions on Biomedical Engineering 56(4), 1015–1022, 2009.
[3] Singh, N., Pillay, V., Choonara, Y.E.: Advances in the treatment of Parkinson's disease. Progress in Neurobiology 81(1), 29–44, 2007.
[4] Parkinson's Disease: National Clinical Guideline for Diagnosis and Management in Primary and Secondary Care, Nat. Collaborating Centre Chronic Conditions, London, U.K., 2006.
[5] Tarigoppula V.S Sriram, M. Venkateswara Rao, G V Satya Narayana , DSVGK Kaladhar, T Pandu Ranga Vital, "Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms". International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 3, September 2013.
[6] F.M. Javed Mehedi Shamrat, Md. Asaduzzaman, A.K.M. Sazzadur Rahman, Raja Tariqul Hasan Tusher, Zarrin Tasnim , "A Comparative Analysis Of Parkinson Disease Prediction Using Machine Learning Approaches". International journal of Scientific & Technology research volume 8, issue 11, November 2019.
[7] Indrajit Mandal, N. Sairam "New machine-learning algorithms for prediction of Parkinson's disease". International Journal of Systems Science, July 2012.
[8] Cai Z, Gu J, Wen C, Zhao D, Huang C, Huang H, Tong C, Li J, Chen H. An Intelligent Parkinson's Disease Diagnostic System Based on a Chaotic Bacterial Foraging Optimization Enhanced Fuzzy KNN Approach, Jun 2018.
[9] Liaqat Ali, Ce Zhu, IEEE Fellow, Zhonghao Zhang, Yipeng Liu, IEEE Senior Member, "Automated Detection of Parkinson's Disease Based on Multiple Types of Sustained Phonations using Linear Discriminant Analysis and Genetically Optimized Neural Network", IEEE Journal of Translational Engineering in Health and Medicine, 2019.
[10] Nilashi M, Ibrahim O, Ahmadi H, Shahmoradi L, Farahmand M. A hybrid intelligent system for the prediction of Parkinson's Disease progression using machine learning techniques. Biocybern Biomed Eng 38(1):1–15, 2018.
[11] Jefferson S. Almeida , Pedro P. Rebouças Filhoa, Tiago Carneiroa, Wei Wei, Robertas Damaševicius, Rytis Maskeliunas, Victor Hugo C. de Albuquerque, "Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques", 2019.
[12] Wang Y, Wang AN, Ai Q, Sun HJ. An adaptive kernel-based weighted extreme learning machine approach for effective detection of Parkinson's disease. Biomed Signal Process Control 38:400–10, 2017.
[13] Singh G, Vadera M, Samavedham L, Lim ECH. Machine learning-based framework for multi-class diagnosis of neurodegenerative diseases: a study on Parkinson's Disease. IFAC-Papers OnLine 49(7):990–5, 2016.
[14] Gunjan Pahuja, T.N.Nagabhushan, "A Novel GA-ELM approach for Parkinson's disease detection using Brain Structural T1-weighted MRI Data" , 2016 Second International Conference on Cognitive Computing and Information Processing, 2016.
[15] Patient Voice Analysis (PVA) Synapse ID: [syn2321745] https://www.synapse.org
[16] H. Kantz and T. Schreiber, "Nonlinear time series analysis," Cambridge University Press, 2nd edition, 2004.
[17] M. A. Little, et al., "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection," Biomedical Engineering Online, vol/issue: 6(23), 2007.
[18] L.Cnockaert, et al., "Low frequency vocal modulations in vowels produced by Parkinsonian subjects," Speech Communication, vol. 50, pp. 288-300, 2008.
[19] Ferchichi S. E., et al., "Feature selection using an SVM learning machines," Proceedings of the 422 3rd International Conference on Signals, Circuits and Systems (SCS 2009), pp. 1-6, 2009.
[20] A. Benba, et al., "Voice assessments for detecting patients with Parkinson's diseases using PCA and NPCA," International Journal of Speech Technology, vol/issue: 19(4), pp. 743–75412, 2016.