Parkinson's disease prediction from speech signals by using Machine Learning

***Abstract*—In this paper we have tried to predict the presence of Parkinson's Disease through speech signals and machine learning methods.The input to our model is a set of features based Jitter, HNR, amplitude etc. We have then used Principal Component Analysis for feature selection followed by various linear and non linear machine learning algorithms. After that two popular boosting algorithms - Adaboost and XGBoost were used on the feature set. The results after both the processes were compared and analysed.**

***Index Terms*—feature selection, Principal Component Analysis, Boosting**

I. Introduction

WEARABLE medical sensing and actuating devices with wireless capabilities have become the cornerstone of many revolutionary digital health applications. Devices like these are equipped with motion and audio sensors. These sensors therefore can acquire signal data like heart rate, frequency of speech, breath rate etc. Technical advancements in this field has led to more and more data in the above form. Since more data are available through this process,the preliminary diagnosis of the disease can be possible; hence diseases can be preventable.

In this paper, we have analyzed speech signals for disease prediction. The analysis is done for Parkinson's Disease (PD) and uses various machine learning based classification methods. PD belongs to one of the categories of neuro-degenerative disease which directly as well as indirectly affects the brain cells that will affect the movement, speech and other cognitive parts [1, 2, and 3]. As the disease progresses more than 90% of the patients have speech disorders [4]. The symptoms related to the vocal impairment of Parkinson's disease patients is called dysphonia. As a result, medical professionals rely on indicators related to dysphonia to assess the PD patients. These measures/indicators related to dysphonia are important and reliable methods to assess the voice related problem and monitor it at different stage [5, 6].

With respect to the past research it is found that artificial intelligence and machine learning techniques have potential for the classification and it was also found that the classification system helps to improve the accuracy and the reliability of the diagnosis.

Usually the measurements have lot of features which is not helpful for machine learning approaches, so feature selection method is used for proper assessment. The feature selection method helps us to evaluate the contribution of the various features in the assessment of the disease at different stages and it helps us achieve good accuracy [7,8].

In this paper we followed the work of Aich et al [12]. We have then employed boosting approaches (Adaboost and XGBoost) and compared the results before and after the use of these methods. In the paper , we start with Section-II which is a review of related work and literature. This is followed by Section-III in which we have discussed briefly about the data set that we have used for our study. Section - IV deals with our proposed technique and Section - V summarises our results.

II. Related Work and Literature Survey

In literature, different studies can be found which focus on speech measurement for general voice disorders and for PD in particular. Some of the studies use a regression approach to detect the level of PD utilizing the UPDRS (Unified Parkinson's Disease Rating Scale) measurements while other studies approach to the problem as a classification problem to detect whether the patient has PD or not. We have discussed the latter type i.e. studies which approach this as a classification problem and have used the same dataset as us to test their models.

- Das et al had done a comparison based on different classification methods on speech signals for effective diagnosis of PD [9]. Four independent classification schema were applied, and a comparative study was carried out. These were Regression Neural Networks, DMneural, and Decision Tree respectively. Various evaluation methods were employed for calculating the performance score of the classifiers.
  The selected classifiers, were implemented with the SAS base software 9.1.3 (Licence number: 291468).

The software includes two different programs and nodes are available for classification and model comparison. These programs are called SAS Enterprise Guide 4.3 and SAS Enterprise Miner 5.2.While SAS Enterprise Guide program 4.3 was used for data pre-processing, SAS Enterprise Miner 5.2 program was used to analyze and recognize the PD by combining several classification methods with model comparison node.
65% of the input dataset was used for training and the rest of the dataset was used for testing. The adjustable parameters of each classifier were tuned. For neural networks classifier, the following adjustments were carried out: The backpropagation learning algorithm has been used in the feed-forward, single hidden layer neural network. The algorithm used in the study is the Levenberg–Marquardt (LM) algorithm. A tangent sigmoid transfer function was used for both the hidden layer and the output layer. 10 neurons were used in the hidden layer. The initial weights were chosen randomly. In regression node, logistic regression was used. Moreover, default values were chosen for both DMneural and Decision tree nodes. The accuracy after testing was a 92.9% for Neural Network, 84.3% for DMNeural, 88.6% for Regression and 84.3% for Decision Tree. Therefore, neural network was the best among the four classifiers with accuracy of 92.9%

- Bhattacharya and Bhatia used SVM based method with different kernels to distinguish the Parkinson group from the healthy group by using Weka data mining tool. They have analyzed the accuracy based on the variation of Receiver Operating Characteristics (ROC) [10].
  The dataset used was the same as that in our study. It was pre-processed using Weka, a data mining tool and it was found that jitter and shimmer measure values are all very close to zero, with some rare examples of exceptionally high values. Also it was found that there exists high correlation between various attribute values. Hence with the help of Weka‘s filtering tool these attributes were removed.
  Classification was done using LIBSVM . It is an integrated software for support vector classification, regression and distribution estimation. It supports multi class classification as well and was used to find the best possible accuracy. The pre-processed dataset obtained was split randomly and was transformed from the .csv format to the libsvm format using a Perl script. For each kernel value : Linear, Polynomial, Radial Basis Function and Sigmoid, the value of the complexity parameter(C) was changed. For each of the kernel values, C value was increased , starting from 100 to 1000.
  The result obtained was plotted using MATLAB , i.e the Receiver Operating Characteristic (ROC) curve. This was represented by plotting the fraction of true positives vs. the fraction of false positives.
  It was inferred that RBF Kernel was not suitable as the test set accuracy kept on decreasing with the increasing value of C and training set accuracy kept on increasing, which indicated overfitting. In case of Sigmoid Kernel for C value ranging from 100 to 1000 the training set accuracy is 83.89% and the test set accuracy is 47.8261% , which was unaffected by changing value of C. For Polykernel as well it was seen that the test set accuracy decreases. Hence the best possible result is obtained for linear kernel which is 65.22%

- Shahbakhi et al proposed a method for diagnosis of PD based on speech analysis by using genetic algorithm (GA) and support vector machine. They have found accuracy of 94.50%, 93.66% and 94.22% on the basis of 4, 7 and 9 optimized features [11].
  This study proposed a new algorithm for diagnosing of Parkinson's disease based on voice analysis. In the first step, genetic algorithm (GA) is undertaken for selecting optimized features from all extracted features. Genetic Algorithm (GA) is an adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection. It is one of the most influential methods in the process of data classification, which is effectively used to select optimized features. In genetic algorithm, the solution is called chromosome or string. This method requires a population of chromosomes (strings) representing a combination of features from the solution set, and requires a cost function (called an evaluation or fitness function). This function calculates the fitness of each chromosome.
  Afterwards a network based on support vector machine (SVM) was used for classification between healthy and people with PD. The dataset of this research is composed of a range of biomedical voice signals from 31 people, 23 with Parkinson's disease and 8 healthy people. The subjects were asked to pronounce letter "A" for 3 seconds. 22 linear and non-linear features were extracted from the signals that 14 features were based on F0 (fundamental frequency or pitch), jitter, shimmer and noise to harmonics ratio, which are main factors in voice signal. A change in these factors is noticeable for the people with PD, optimized features were selected among them. Of the various numbers of optimized features, the data classification was investigated.
  Results showed that the best classification accuracy of 94.50% was achieved for 4 optimized features. It can be observed this was achieved using Fhi (Hz), Fho (Hz), jitter (RAP) and shimmer (APQ5).

- Aich et al, employed PCA and GA parallelly for feature selection and then the performance measures were compared with different machine learning classifiers [12].They found an maximum accuracy of 97.57% using SVM with RBF by using genetic algorithm-based feature sets.
  All of the code was written in R Programming lan-

guage.The ratio of train data to test data was 70:30. They found 11 features after implementing PCA to the original dataset and 10 features using GA based feature sets. After feature selection, they have used different classification approaches such as RPART, C4.5, PART, Bagging classification and Regression tree (Bagging CART), Random Forest, Boosted C5.0 and SVM.Results obtained from each classifier were then compared. The parameters used to compare the performance and validations of classifier were: accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV). The sensitivity is defined as the ratio of true positives to the sum of true positives and false negatives. The specificity is defined as the ratio of true negatives to the sum of false positives and true negatives. In our research we have used the Positive predictive value and negative predictive value to check the present and absent of disease.

In case of SVM different Kernel function such as linear, polynomial, and radial basis function were used. SVM with RBF had the highest accuracy of 97.57% with GA based feature sets followed by random forest and RPART classifiers with the same feature sets. Random forest had the highest sensitivity of 0.9985 with GA based feature sets followed by SVM with RBF classifiers with the same feature sets. The SVM classifier has a sensitivity of 0.9756. SVM with GA based feature set had highest specificity of 0.9987. It was also seen that GA based feature sets perform better compared to the PCA based feature set with random forest classifier. GA based feature sets showed maximum NPV of 0.9995.

III. Dataset

The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders [13]. It is composed of 23 columns and 197 rows and consists of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The time since diagnoses ranged from 0 to 28 years, and the ages of the subjects ranged from 46 to 85 years (mean 65.8, standard deviation 9.8). Details of the data set are summarised in Table 1.

Table 1 : Matrix column entries in the Data Set

Column Entries	Description
Name	ASCII subject name and recording number
MDVP: Fo (Hz)	Average vocal fundamental frequency
MDVP: Fhi(Hz)	Maximum vocal fundamental frequency
MDVP: Flo(Hz)	Minimum vocal fundamental frequency
MDVP: Jitter (%)	Several measures of variation in fundamental frequency
MDVP: Jitter(Abs)	
MDVP:RAP	
MDVP:PPQ	
Jitter: DDP	
MDVP: Shimmer	Several measures of variation in amplitude
Shimmer: APQ	
Shimmer:APQ5	
MDVP:APQ	
Shimmer: DDA	
NHR	Two measures of ratio of noise to tonal components in the voice
HNR	
status	Health status of the subject (one) - Parkinson's, (zero) - healthy
RPDE, D2	Two nonlinear dynamical complexity measures
DFA	Signal fractal scaling exponent
spread1	Three nonlinear measures of fundamental frequency variation
spread2	
PPE	

The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD. The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient; the name of the patient is identified in the first column.

IV. Methodology

The proposed technique is as per Fig-1 and we have used the programming language Python3 for implementing our model. In PCA , components or the latent variables are obtained from

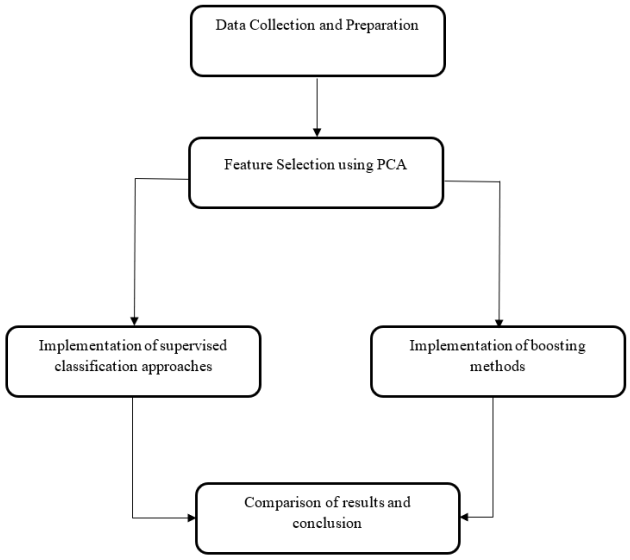

Figure 1. Flowchart of the technique

the variance of the data by maximizing it. The number of principal components is lesser than the regular variables. PCA reduces the dimensionality of the space so that the data can be visualized in the low dimensional space. The feature selection process is done by removing the redundant variables[14]. There were 11 features obtained, after implementing the algorithm to the original data set.

We have then used different classification approaches such as:Classification and Regression Trees (CART), Random Forest , Bagging Classification and Regression Tree (Bagging CART), Support Vector Machine (SVM), KNN, Logistic Regression and Gaussian Naive Bayes. In this paper we have used radial basis function as kernel function for SVM.

The CART or Classification Regression Trees methodology was introduced in 1984 by Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone as an umbrella term to refer to the following types of decision trees:

- Classification Trees: where the target variable is categorical and the tree is used to identify the "class" within which a target variable would likely fall into.
- Regression Trees: where the target variable is continuous and tree is used to predict it's value.

It is structured as a sequence of questions, the answers to which determine what the next question, if any should be. The result of these questions is a tree like structure where the ends are terminal nodes at which point there are no more questions.

Bagging CART is an extension over CART , as it constructs n classification trees using bootstrap sampling of the training data and then combines their predictions to produce a final meta-prediction. It is therefore an ensemble learning mechanism.

Random Forest is also a decision tree-based algorithm. Random Forest as a Classifier is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object.

In parallel to this, we have also employed boosting approaches : Adaptive Boosting (AdaBoost) and Extreme Gradient Boosting (XGBoost). AdaBoost calls a given weak or base learning algorithm repeatedly in a series of rounds One of the main ideas of this algorithm is to maintain a distribution or set of weights over the training set. Initially, all weights are set equally, but on each round, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training. We have used AdaBoost with various base estimators such as : Random Forest Classifier, SVM and CART.

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

V. Results

The data set was split in a ratio of 70:30 for training and testing. We achieved varying accuracy measurements for various models. Maximum accuracy was achieved when AdaBoost was used with Random Forest as base estimator. Fig-2 shows the accuracy as obtained through various methods. Table -2 compares the results when SVM, CART and Random Forest (RF) Classifier are used after feature selection and when each of these is used as a base estimator for Adaboost.

Table 2 : Comparison of accuracy when SVM, CART and RF are used as base estimators

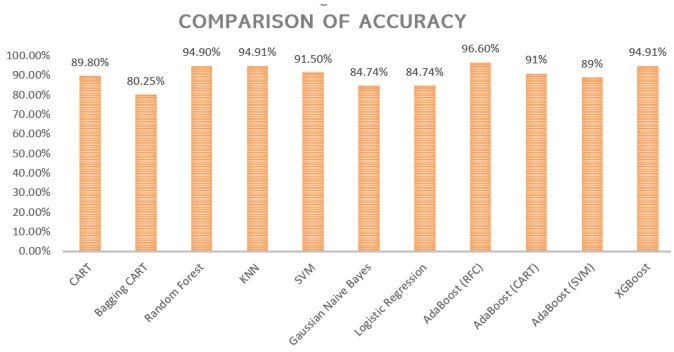

Figure 2. Comparison of accuracy of various models

Sl. No	Classification Method	Accuracy (%)
1.	Random Forest	94.90
	Adaboost with Random Forest Classifier as base estimator	96.60
2.	CART	89.80
	Adaboost with CART as base estimator	91
3.	SVM	91.5
	Adaboost with SVM as base estimator	89

VI. Conclusion

In this study we found that boosting algorithms yield better results when base classifiers are decision tree based. The highest accuracy in our study was achieved when Adaptive Boosting was used with Random Forest as base estimator.XGBoost also gave good results. Boosting is on of the ensemble learning methods, therefore in other ensemble learning approaches may also have potential to give better results.

References

[1] S.Przedborski, M.Vila, and V.Jackson-Lewis, "Series Introduction: Neurodegeneration: What is it and where are we?", Journal of Clinical Investigation, 111(1), pp. 3-10, 2003.

[2] Y.Xu, X.Wei, X.Liu, J.Liao, J.Lin, C.Zhu and M.Cheng, "Low cerebral glucose metabolism: a potential predictor for the severity of vascular Parkinsonism and Parkinson's disease", Aging and disease, 6(6), pp. 426-436, 2015.

[3] K.Tjaden, "Speech and swallowing in Parkinson's disease", Topics in geriatric rehabilitation, 24(2), pp. 115-126, 2008.

[4] A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, and S. Gates, "Speech impairment in a large sample of patients with Parkinson's disease", Behavioural Neurology, 11(3), pp.131–137,1998.

[5] Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., Ramig, L.O. (2009). "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease", IEEE Transactions on Biomedical Engineering, 56(4), 1015–1022.

[6] Rahn, D. A., Chou, M., Jiang, J. J., and Zhang, Y. (2007), "Phonatory impairment in Parkinson's disease: evidence from nonlinear dynamic analysis and perturbation analysis", Journal of Voice, 21, pp. 64–71.

[7] T. Hastie, R. Tibshirani, and J. H. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-Color Illustrations", New York: Springer-Verlag, 2001.

[8] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J.Mach. Learn. Res., vol. 3, pp. 1157–1182, 2003.

[9] R.Das, "A comparison of multiple classification methods for diagnosis of Parkinson disease", Expert Systems with Applications, 37(2), pp. 1568-1572, 2010.

[10] I. Bhattacharya, and M. P. S. Bhatia, "SVM classification to distinguish Parkinson disease patients", Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India. ACM, 2010.

[11] M.Shahbakhi, D. T.Far, and E. Tahami, “Speech analysis for diagnosis of Parkinson’s disease using genetic algorithm and support vector machine”, Journal of Biomedical Science and Engineering, 7(4), pp.147-156, 2014.

[12] S.Aich , H.C. Kim, K. Young A, K.L. Hui, A.A Al-Absi and M. Sain, “A Supervised Machine Learning Approach using Different Feature Selection Techniques on Voice Data sets for Prediction of Parkinson’s Disease”, ICACT Transactions on Advanced Communications Technology (TACT) Vol. 7, Issue 3, pp 1116-1121, 2018

[13] M. A.Little, P.E. McSharry, E. J.Hunter, J.Spielman, and L. O.Ramig, “Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease” , IEEE transactions on biomedical engineering,2009, 56(4), pp.1015-1022.

[14] Guo, Q., Wu, W., Massart, D. L., Boucon, C., and De Jong, S. (2002). “Feature selection in principal component analysis of analytical Data”, Chemometrics and Intelligent Laboratory Systems, 61(1-2), 123-132.