Disease Prediction from Signals of Wearable devices by using Machine Learning

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 have tried to increase accuracy by using Principal Component Analysis for feature selection followed by various machine learning classifiers. We have then employed boosting approaches (Adaboost and XGBoost) and compared the results before and after the use of these methods.

II. Related Work and Literature Survey

Various classification and feature selection methods have been employed in the past to get higher accuracy in the prediction.We have classified the past research into two categories: Classification Methods and Feature Selection. This is based on the approach to PD prediction.

A. Classification Methods

Over the years, various Machine Learning and Deep Learning approaches have been used for identification of individuals as : healthy and not healthy.

- Das et al had done a comparison based on different classification methods on speech signals for effective diagnosis of PD. They used four different classification methods such as Neural Networks, Regression, DMneural, and Decision tree and found that neural network was the best among the four classifiers with accuracy of 92.9% [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].

- Polat used fuzzy c-means clustering feature weighting and k-NN classification technique for the detection of PD. They found the combined approach performs better in the classification of PD [11].

- Froelich et al presented the diagnosis of PD based on the characteristics features of a person's voice. They have used decision tree-based classification approach using the threshold value. They found classification accuracy of 90% [12].

B. Feature Selection

Various data sets have a lot of features, all of which may or may not contribute equally to PD prediction . Therefore, Feature Selection is used to reduce the dimesionality of the data set, so as to produce more accurate results.

- Li et al proposed a fuzzy based method transformation system to extract good features and then used Principal Component Analysis (PCA) to find the optimal features among them. They have used SVM for the prediction of PD [13].
- 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 [14].
- Aich et al, employed PCA and GA parallelly for feature selection and then the performance measures were compared with different machine learning classifiers [15].

III. Methodology

In this paper we have used the dataset created by Max Little University Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals [20]. The original data collected from the dataset composed of voice measurements from 31 people out of which 23 were diagnosed with PD.

We have used Principal Component Analysis (PCA) algorithm on the original feature sets. 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 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[19]. 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) and Support Vector Machine (SVM).

In parallel to this, we have also employed boosting approaches: Adaptive Boosting (AdaBoost) and Extreme Gradient Boosting (XGBoost). AdaBoost calls a given algorithm, termed 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.

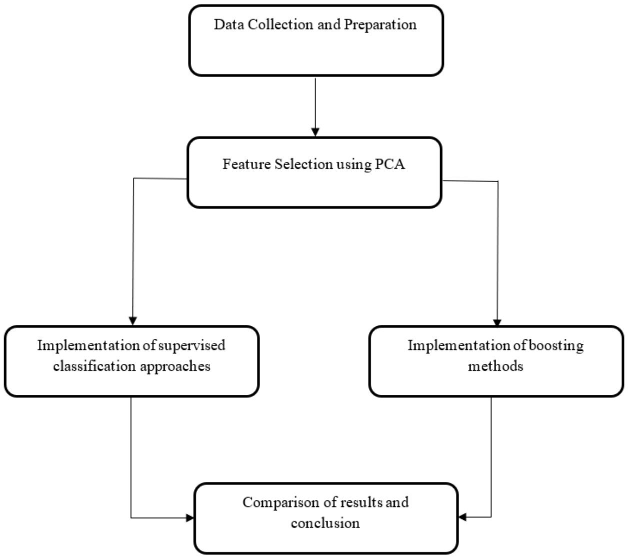
Data Collection and Preparation
Feature Selection using PCA
Implementation of supervised classification approaches
Implementation of boosting methods
Comparison of results and conclusion

Figure 1. Flowchart of the technique

IV. 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. These are summarized below:

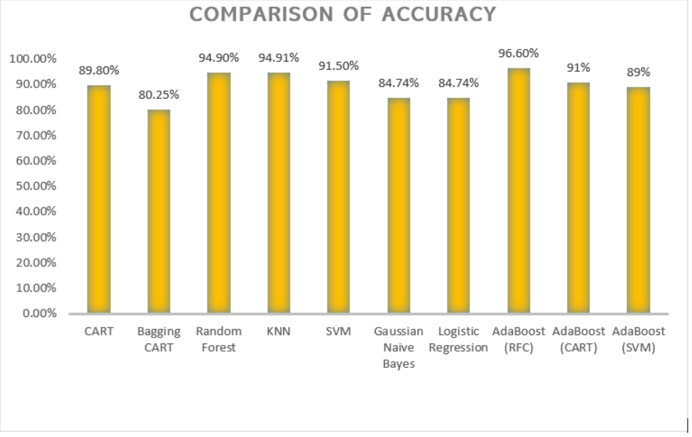

Figure 2. Comparison of accuracy of various models

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] K.Polat, “Classification of Parkinson’s disease using feature weighting method on the basis of fuzzy C-means clustering”, International Journal of Systems Science, 43(4), pp.597-609, 2012.

[12] W.Froelich, K. Wrobel, and P. Porwik, “Diagnosis of Parkinson’s disease using speech samples and threshold-based classification”, Journal of Medical Imaging and Health Informatics, 5(6), pp. 1358-1363, 2015.

[13] D.C.Li, C. W.Liu, and S. C.Hu, “A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets”, Artificial Intelligence in Medicine, 52(1), 45-52, 2011.

[14] 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.

[15] 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

[16] 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.

[17] 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.