A Nonlinear Decision Tree based Classification Approach to Predict the Parkinson's disease using Different Feature Sets of Voice Data

Satyabrata Aich[1], Kim younga[2], Kueh Lee Hui[3], Ahmed Abdulhakim Al-Absi[4] and Mangal Sain[5]

[1]Department of Computer Engineering, Inje University, South Korea

[2]Department of Rehabilitation Science, Inje University, South Korea

[3]Department of Electrical Engineering, Dong-A University, South Korea

[4]Department of Computer Engineering, Kyungdong University- Global Campus, Gangwondo, South Korea

[5]Department of Computer Engineering, Dongseo University, South Korea

satyabrataaich@gmail.com, kya2664@hanmail.net, leehkueh@dau.ac.kr, absiahmed@kduniv.ac.kr, mangalsain1@gmail.com

Corresponding author email id: leehkueh@dau.ac.kr

Abstract— **In the past few years, lot of researchers are working to get some breakthrough for early detection of Parkinson's disease. As the old age population is increasing at a higher rate as well as it is predicted that the old age population will increase to a much higher total by 2050, it's a become a rising concern to the developed countries because the cost due to the healthcare service of these disease is really high. Parkinson's disease (PD) belongs to the group of neurological disorder, which directly affect the brain cells and the effect is shown in terms of movement, voice and other cognitive disabilities. Researchers are keep working on different fields such as gait analysis as well as on speech analysis to find the predictors of the Parkinson's disease. Recently machine learning based approach has been used by many researchers across the field because of its accuracy on the complex data. Machine learning based approach has been used in many cases of Parkinson's disease using gait data as well as voice data. However, so far no body has compared the performance metrics using different feature sets by applying non-linear based classification approach based on the voice data. So in this paper we have proposed a new approach by comparing the performance metrics with different feature sets such as original feature sets as well as Principal component Analysis based feature reduction technique for selecting the feature sets. We have used non-linear based classification approach to compare the performance metrics. We have found an accuracy of 96.83% using random forest classifiers using PCA based feature sets. This analysis will help the clinicians to differentiate the PD group from healthy group based on the voice data.**

Keywords— **Parkinson's disease, machine learning, feature selection, voice data, performance metrics**

I. INTRODUCTION

In past few years a lot of research has been going on the Parkinson's disease because the healthcare related cost due to this disease is keeping on increasing as the longevity of the population is increasing in the developed countries. Since this disease affect most of the old people, it is become necessary for the developed counties to detect the disease at the early stage. The early detection will help the developed country in economic perspective as well as social perspective because it can be assessed well. Parkinson's disease belongs to one of the category of neurodegenerative 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].The Parkinson's disease is progressive in nature. As the disease progresses more than 90% of the patients has the speech disorder [4].

The symptoms related to the vocal impairment of Parkinson's disease patients is called dysphonia.The clinicians measured some indicators related to dysphonia to assess the PD patients. The measures related to dysphonia could be treated as an important and most reliable tool to assess the voice related problem and monitor it at different stage [5, 6]. Usually the measures have lot of features which does not helpful for machine learning approaches, so feature selection method has been used for proper assessment. The feature selection method will help to evaluate the important contribution of the features in the assessment of the disease at different stage and also it helps to achieve good accuracy [7, 8]. The traditional diagnosis needs lot of observations related to the daily living activities, motor skills and other neurological parameters to assess the progression of PD, but this process is not suitable for the early detection of the PD. With respect to the past research it is found that artificial intelligence and machine learning techniques have good potential for the classification and it also found that the classification system helps to improve the accuracy and the reliability of the diagnosis and also minimize the errors as well as make the system more efficient [9]. Improvement on the prediction of accuracy on the progression of PD is getting lot of attention these days [10, 11].

In this paper an attempt has been made to check the improvement in the accuracy while classifying the PD group from the healthy control group by using nonlinear feature

selection algorithm with different feature sets such as the original feature sets (OFS) as well as the PCA based feature sets. Finally a comparison has made in terms of performance metrics using different feature sets. The structure of the paper is organized as follows: Section 2 presents the past work related to classification model used for voice datasets. Section 3 describes about the methodologies used for this research work. Section 4 describes about the result of feature selection as well as the result of classification. Section 5 describes about the conclusion and future work.

II. Related work

Shahbaba and Neal used nonlinear based approach for classification of PD. They have used Dirichlet process mixtures and compared the results with other classification model such as decision trees, support vector machine and multinomial logit model and they found Dirichlet process based method provides best classification approach of 87.7% compared to the other model [12]. Sakar and Kursun used feature selection method as well as machine learning based method for diagnosis of PD. They have used mutual information based feature selection and support vector machine as the classification approach and they found their approach gives an accuracy of 92.75% [13]. Li et al used fuzzy based method to extend the classification related information and then they have used principal component analysis based method for feature selections and the optimal features has been integrated with SVM based method provides a good accuracy of 93.47%[14].

Spadoto et al used evolutionary base techniques for feature selection and they have used Optimum-path Forest Classifier to detect the Parkinson's disease and they found this approach provides a best accuracy of 84.01% while detecting the PD [15]. Luukka have proposed a feature selection method based on the fuzzy entropy measure and used similarity classifiers to classify PD.The best classification accuracy obtained by that method was 85.03% [16]. AStröm and Koker proposed a method that is used parallel feed-forward neural network based approach to predict the PD.They have found the model is robust and the best classification accuracy obtained from that approach is 91.20% [17].Nilashi et al proposed a method for the prediction of PD progression using clustering and prediction methods. They have applied Adaptive Neuro-Fuzzy Inference system (ANFIS) and Support Vector Regression for prediction of PD progression. They found this proposed method helps to improve the accuracy of the progression of PD [18].The above past works motivated us to try a different approach. In this paper we have tried a different way of selecting feature by using Principal component Analysis (PCA) approach and compared the performance metrics with the original feature sets using nonlinear classifiers with decision tree.

III. Proposed Technique

The flow chart of the proposed methodologies is shown in the figure 1.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 [19]. 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. Principal Component Analysis (PCA) is a tool that is used for compression of data and extraction of information [20].

We have found 11 features after implementing the algorithm to the original feature sets. We have used different feature sets such as the original Feature sets (OFS) and PCA based feature sets. We have used nonlinear classifier with decision tree for classification of groups are as follows RPART, C4.5, PART, Bagging classification and Regression tree(Bagging CART), Random Forest and Boosted C5.0

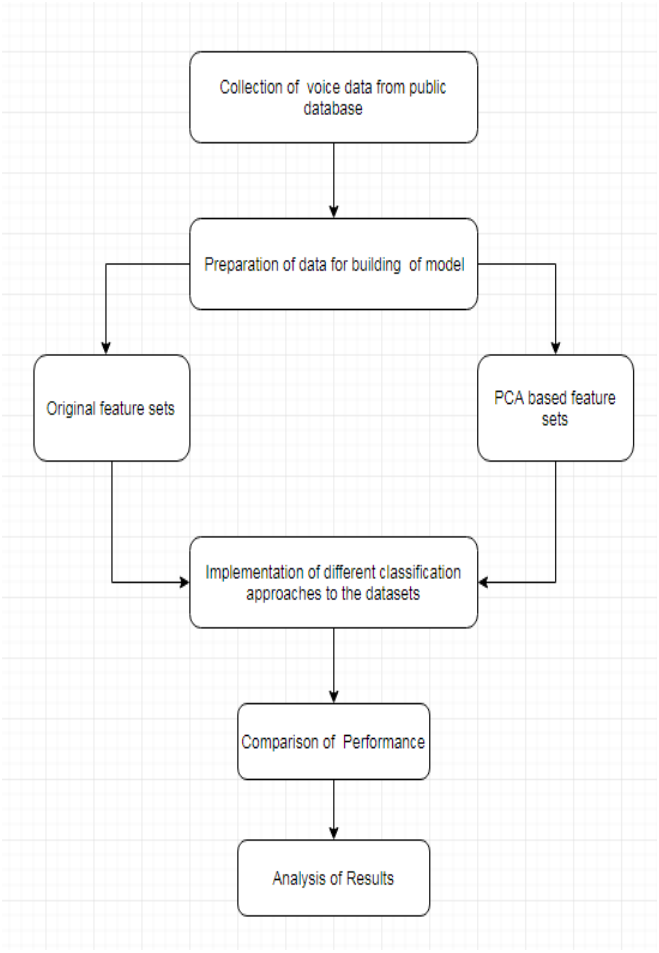

Figure 1. Flowchart of the proposed method

A. Performance Metrics

The parameters used to compare the performance and validations of classifier are as follows: 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. So the ppv is the probability that the disease is present given a positive test result and npv is the probability that the disease is absent given a negative test result [21]. Accuracy is defined as the ratio of number of correct predictions made to the total prediction made and the ratio is multiplied by 100 to make it in terms of percentage.

IV. Result and Discussions

We have used R programming language to write the code. We trained each classifier based on the trained data and predict the power of classifier on the test data. So each classifier able to show all the performance metrics based on the test data.

A. Comparison of Accuracy

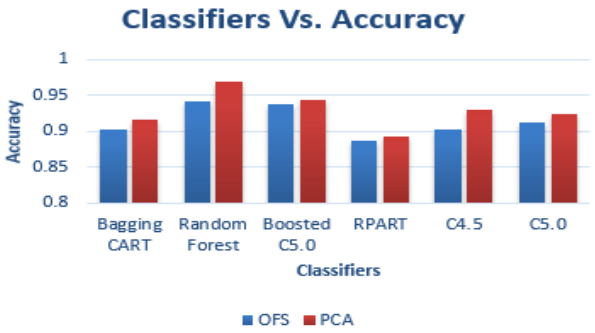

Figure 2. Accuracy of different classifiers

Fig. 2 shows that Random forest performs better with PCA based feature sets in terms of accuracy among all the classifiers. It shows the maximum accuracy of 96.83%.

B. Comparison of Sensitivity

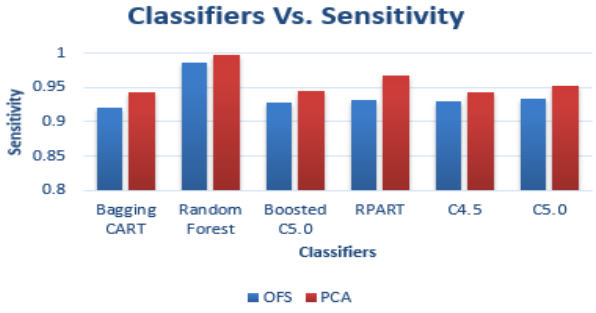

Figure 3. Sensitivity of different classifiers

C. Comparison of Specificity

Fig. 3 shows that Random forest with PCA based feature sets have highest sensitivity among other classifiers. It shows the maximum sensitivity of 0.9975

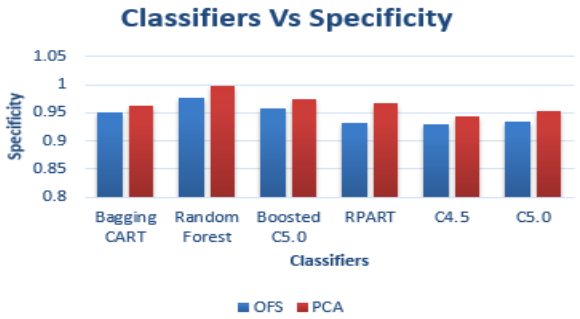

Figure 4. Specificity of different classifiers

Fig. 4 shows that Random forest with PCA based feature sets have highest specificity among other classifiers. It shows the maximum specificity of 0.9985.

D. Comparison of PPV

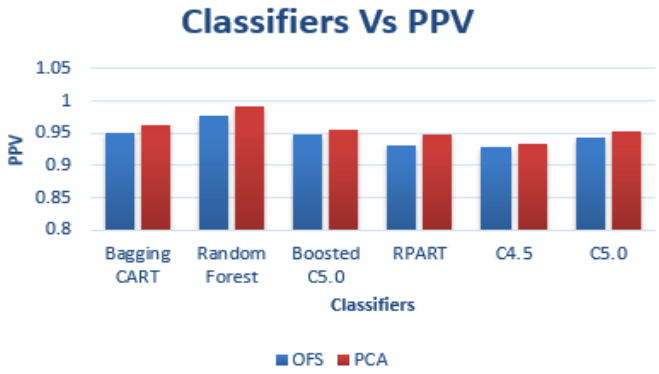

Figure 5. PPV of different classifiers

Fig. 5 shows that Random forest with PCA based feature sets have highest PPV among other classifiers. It shows the maximum PPV of 0.9912.

E. Comparison of NPV

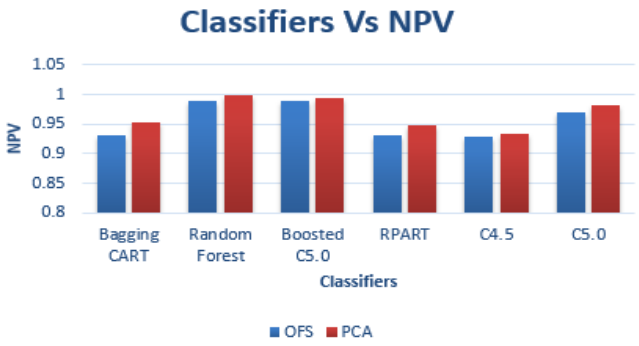

Figure 6. NPV of different classifiers

Fig. 6 shows that Random forest with PCA based feature sets have highest NPV among other classifiers. It shows the maximum NPV of 0.9994.

V. CONCLUSION AND FUTURE WORK

In this paper we have used nonlinear classifier with decision tree to classify the PD and control group and we found good result by achieving an accuracy of 96.87% using the PCA based feature sets with random forest classifier. We have compared the performance metrics of different classifiers with two different feature sets such as OFS and PCA feature sets. Overall the PCA based feature sets performed better with random forest classifier in terms of accuracy, sensitivity, specificity, PPV, NPV compared to the original feature sets. This analysis will help the clinicians to shift focus towards the important features while early diagnosis of Parkinson's disease. In the future we will try other feature reduction techniques and as well as other classification technique to compare the performance of all the parameters of the performance metrics.

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. T jaden, "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., Mc Sharry, 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., & Zhang, Y. (2007). Phonatory impairment in Parkinson's disease: evidence from nonlinear dynamic analysis and perturbation analysis. Journal of Voice, 21, 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] D. Gil, and D. J. Manuel, "Diagnosing Parkinson by using artificial neural networks and support vector machines", Global Journal of Computer Science and Technology, 9(4), pp.63-71, 2009.

[10] 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.

[11] M. Hariharan, K. Polat, and R. Sindhu, A new hybrid intelligent system for accurate detection of Parkinson's disease. Computer methods and programs in biomedicine, 113(3), pp.904-913, 2014.

[12] Shahbaba, B., & Neal, R. (2009). Nonlinear models using Dirichlet process mixtures. The Journal of Machine Learning Research, 10, 1829–1850..

[13] Sakar, C. O., & Kursun, O. (2010). Telediagnosis of Parkinson's disease using measurements of dysphonia. Journal of Medical Systems, 34, 1–9

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

[15] Spadoto, A. A., Guido, R. C., Carnevali, F. L., Pagnin, A. F., Falcao, A. X., & Papa, J. P. (2011). Improving Parkinson's disease identification through evolutionary based feature selection. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE (pp. 7857–7860).

[16] Luukka, P. (2011). Feature selection using fuzzy entropy measures with similarity classifier. Expert Systems with Applications, 38, 4600–4607

[17] AStröm, F., & Koker, R. (2011). A parallel neural network approach to prediction of Parkinson's disease. Expert Systems with Applications, 38, 12470–12474.

[18] M. Nilashi, O. Ibrahim, A. Ahani, "Accuracy Improvement for Predicting Parkinson's Disease Progression," *Scientific Reports,*vol.6,34181,2016

[19] 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.

[20] E. R. Hruschka, N.F.Ebecken, "Extracting rules from multilayer perceptrons in classification problems: A clustering-based approach, "Neurocomputing, vol.70 ,no.2,pp. 384-397,2006

[21] H. B. Wong, G. H. Lim, Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV, Proceedings of Singapore healthcare, vol.20, no.4, pp.316-318, 2011.

[22] R. J. Vidmar, On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. *21 (3).* pp. 876—880. Available: http://www.halcyon.com/pub/journals /21ps03-vidmar, Aug. 1992.

Satyabrata Aich is working as a researcher in the field of computer engineering He has over four years of teaching, research and industry experience in India and abroad. He has published many research papers in journals and conferences in the realms of Supply Chain Management and data analytics. His research interests are natural language processing, Machine learning, supply chain management, data mining.

Young a Kim is a licensed Physical therapist and has received her Masters degree in Counseling and Psychology in 2008. She is working as a lecturer for Clinical Kinesiology and Functional Anatomy at Inje University. She is also a PhD candidate in the Department of Rehabilitation Science at the same University. Her research focuses on assessment tools and rehabilitation for the elderly, especially for the detection and treatment of age related conditions. She is a co-author of the book 'Community Occupational Therapy' and a frequent contributor to the journal of 'Physical Therapy Science' and a member in the Korean Ageing Friendly Industry Association as well as the Society of Occupational Therapy for the Aged and Dementia.

Kueh Lee Hui is working as an assistant professor at the department of Electrical Engineering, Dong-A University since 2012. She completed her PhD Degrees from Department of Electrical Engineering, Dong-A University, Korea. In 2009 she completed her BS degree in Electronic and Communication, Department of Electronic Engineering, University Malaysia of Sarawak, Malaysia. She also done MS in 2007 from Malaysia. Her research interests are image processing, face recognition, digital image forensic, intelligent control and control application, power system.

Ahmed Abdulhakim Al-Absi is an assistant professor in Department of Computer Engineering (Smart Computing) at Kyungdong University in South Korea. He earned a Ph.D. in computer science from Dongseo University in 2015. He received M.Sc. degree in information technology at University Utara Malaysia in 2011, and B.Sc. degree in computer applications at Bangalore University in 2008. His research interests include Big Data processing, Hadoop, Cloud computing, IoT, Distributed systems, Parallel computing, Bioinformatics, Security, and VANETs.

Mangal Sain received the M.Sc. degree in computer application from India in 2003 and the Ph.D. degree in computer science in 2011. Since 2012, he has been an Assistant Professor with the Department of Computer Engineering, Dongseo University, South Korea. His research interest includes wireless sensor network, cloud computing, Internet of Things, embedded systems, and middleware. He has authored over 50 international publications including journals and international conferences. He is a member of TIIS and a TPC member of more than ten international conferences.