ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 3, September 2013

Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms

Tarigoppula V.S Sriram[1], M. Venkateswara Rao[2], G V Satya Narayana[3] , DSVGK Kaladhar[4], T Pandu Ranga Vital[5]
[1]MCA, Raghu Engineering College, Visakhapatnam, India
[2]IT, GITAM University, Rushikonda, Visakhapatnam, India
[3]IT, Raghu Institute of Technology, Visakhapatnam, India
[4]Dept. Of Bioinformatics, GITAM University, Visakhapatnam, India
[5]CSE, Raghu Engineering College, Visakhapatnam, India

Abstract*— Diagnosis of the Parkinson disease through machine learning approache provides better understanding from PD dataset in the present decade. Orange v2.0b and weka v3.4.10 has been used in the present experimentation for the statistical analysis, classification, Evaluation and unsupervised learning methods. Voice dataset for Parkinson disease has been retrieved from UCI Machine learning repository from Center for Machine Learning and Intelligent Systems. The dataset contains name, MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), MDVP:Jitter(%), MDVP: Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP, MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA, NHR, HNR, status, RPDE, DFA, spread1, spread2, D2, PPE attributes. The parallel coordinates shows higher variation in Parkinson disease dataset. SVM has shown good accuracy (88.9%) compared to Majority and k-NN algorithms. Classification algorithm like Random Forest has shown good accuracy (90.26) and Naïve Bayes has shown least accuracy (69.23. Higher number of clusters in healthy dataset in Fo and less number in diseased data has been predicted by Hierarchal clustering and SOM.

Index Terms- **Data mining, Voice dataset, Parkinson disease.**

I. INTRODUCTION

The clinical diagnosis of Parkinson disease (PD) can be confirmed basing on neuro-pathologic and histo-pathologic criteria [1]. Clinical diagnostic classification of PD can be done on comprehensive review of the literature data and selection basing on the sensitivity and specificity of the characteristic clinical features. Prospective with clinic-pathologic studies in representative population of patients showing PD are needed to investigate the clinical, pathologic, and nosologic studies based on frequency of occurrence, characteristics, and risk factors in patients [2].

Neural Networks, DMneural, Regression and Decision Trees are previously employed for calculating the performance score of the classifiers reliable diagnosis of PD [3, 4]. PD causes vocal impairment that effects speech, motor skills, and other functions like behavior, mood, sensation and thinking. Tele_monitoring of the disease using voice measurement has a vital role in its early diagnosis of PD. The conventional bootstrapping or leave-one-out validation methods with Support Vector Machine (SVM) for building a classification to build a predictive model for assessing the relevance and the statistical significance of the PD relations to attributes [5].

Diagnostic and predictive value of various clinical features provide diagnosis of multiple systems Parkinson's disease [6]. Classification accuracy (ACC), Kappa Error (KE) and Area under the Receiver Operating Characteristic (ROC) Curve (AUC) with two base classifiers, i.e. KStar and IBk provide diagnosis model in PD diagnosis accuracy [7].

Medical biometrics plays an important role in Diagnosing disorders like PD. Medical decision boundaries for detecting Parkinson's disease (PD) demonstrates the effectiveness and computational efficiency of the mechanism of class boundaries to separate healthy subjects from those with PD [8]. The accuracy obtained by training the probabilistic neural network using Parkinson disease dataset has been previously predicted using WEKA 3 and MatLab v7 [9]. The neural network based diagnosis of medical diseases in the present decade shows a great deal of attention in the prediction of PD [10].

II. METHODOLOGY

The voice dataset for Parkinson disease has been retrieved from UCI Machine learning repository from Centre for Machine Learning and Intelligent Systems [11]. The dataset range of biomedical voice measurements from 31 people, where 23 people are showing Parkinson's disease. The class column represents "status" which is set to 0 for healthy and 1 for PD. The data is in ASCII CSV format. The following are the Attribute Information:

name,MDVP:Fo(Hz),MDVP:Fhi(Hz),MDVP:Flo(Hz),MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP,MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA,NHR,HNR,status,RPDE,DFA,spread1,spread2,D2,PPE

Matrix column entries (attributes) show the following 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(%),MDVP:Jitter(Abs),MDVP:RAP,MDVP:PPQ,Jitter:DDP - Several measures of variation in fundamental frequency

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 3, September 2013

MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3, Shimmer:APQ5,MDVP:APQ,Shimmer:DDA - Several measures of variation in amplitude

NHR, HNR - Two measures of ratio of noise to tonal components in the voice

status - Health status of the subject (one) - Parkinson's, (zero) - healthy

RPDE, D2 - Two nonlinear dynamical complexity measures

DFA - Signal fractal scaling exponent

spread1, spread2, PPE - Three nonlinear measures of fundamental frequency variation

The data contains 5875 number of instances and 26 attributes. The dataset has been retrieved and executed in Orange software v2.0b for data visualization (parallel coordinates, Sieve graphs and SOM), classification (Majority, k-nearest neighbor and SVM), Evaluation and unsupervised learning methods (Hierarchal clustering). The dataset is also used for correctly classified instances (classification) using weka v3.4.10 for Bayes Net, Naïve Bayes, Logistic, Simple Logistic, KStar, ADTree, J48, LMT and Random Forest.

III. RESULTS

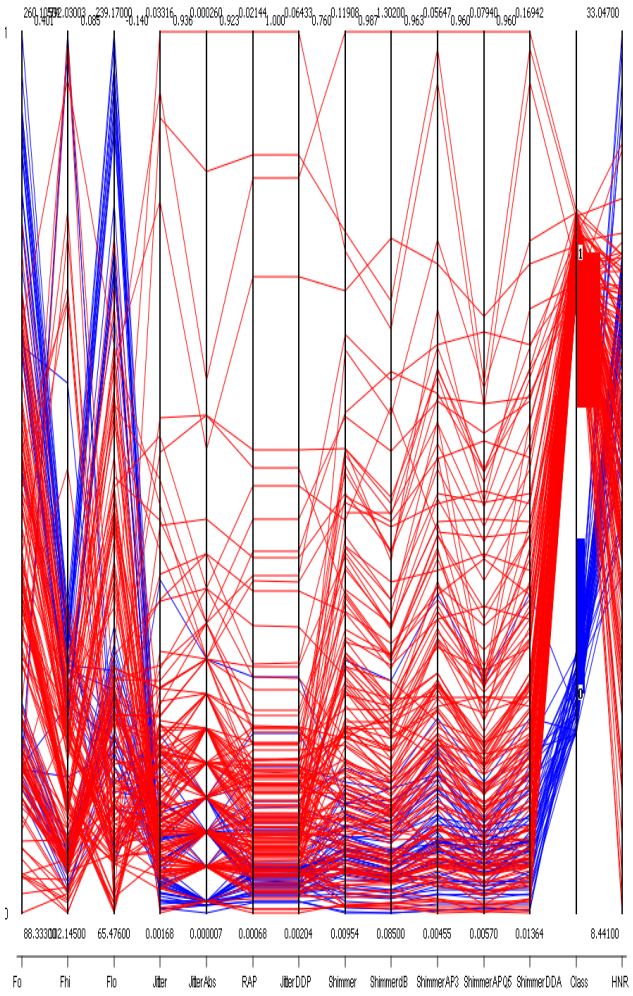

Fig. 1. Parallel coordinates.

The dataset that has been retrieved has shown the statistical analysis, classification, Evaluation and providing unsupervised learning methods. These algorithms provide the diagnosis methods for finding the attributes that are probable towards presence of PD. Fig 1 and Fig. 2 shows the data visualization for the submitted voice dataset. The parallel coordinate's shows higher variation in Parkinson disease Red colored. All the attributes has network of links using sieve graphs showing inter and intra connections with the healthy and diseased data like Fo, Flo Jitter(%), Jitter(Abs), Jitter, Shimmer,Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,APQ,Shimmer:DDA Fig. 3 has shown the ROC for classification algorithms (Majority, k-nearest neighbor and SVM). ROC plot for Majority has shown specificity at 0.25. k-NN has shown 82.5% of accuracy. SVM has shown 88.9% accuracy based on ROC results. Hence SVM has predicted good results in comparison with the other two algorithms.

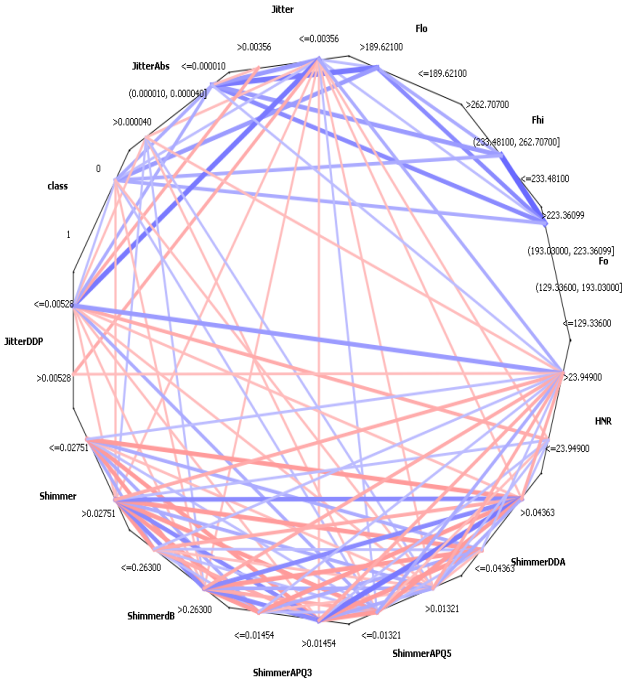

Fig. 2. Sieve graph

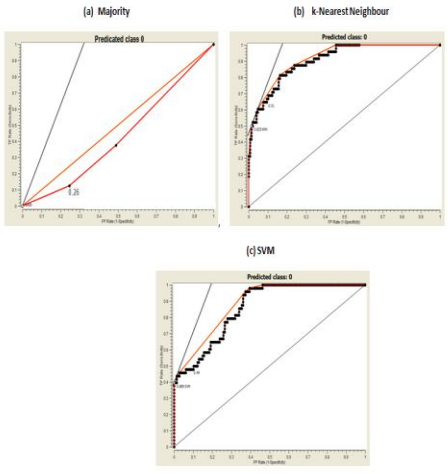

Fig. 3. ROC for Classification Algorithms

Table 1 has shown the accuracy (Correctly classified instances) based on the algorithms using weka software. Based on the results, Random Forest has shown good

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 3, September 2013

accuracy (90.26) followed by KStar (89.74). Naïve Bayes has shown least accuracy (69.23) based on PD dataset.

Table 1. Correctly classified instances based on Algorithms using Weka v3.4.10

Algorithm	Correctly Classified Instances
Bayes Net	80.00
Naïve Bayes	69.23
Logistic	83.66
Simple Logistic	84.61
KStar	89.74
ADTree	86.15
J48	80.51
LMT	86.15
Random Forest	90.26

Fig. 4 has shown the higher number of clusters in healthy dataset in Fo and less number in diseased data. Healthy dataset range from 95.7 to 252.45 and diseased dataset has range from 144.18 to 202.63 for Fo attribute.

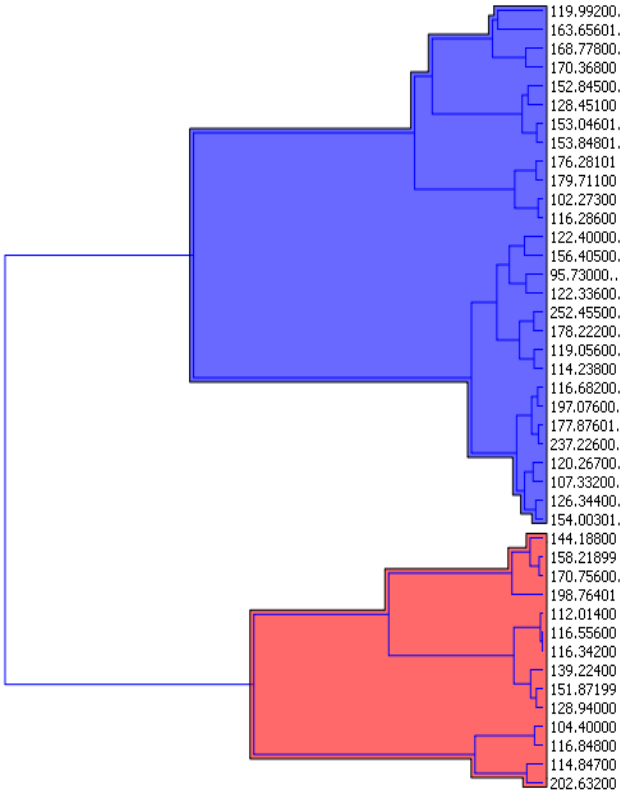

Fig. 4. Hierarchal clustering for Fo attribute.

Fig. 5 has shown the Self-Organizing Map (SOM), a most powerful algorithm in data visualization and exploration. Visualization and clustering of the data is relevant to Fo provides qualitative information towards healthy and PD datasets. Most of the data occupies diseased class.

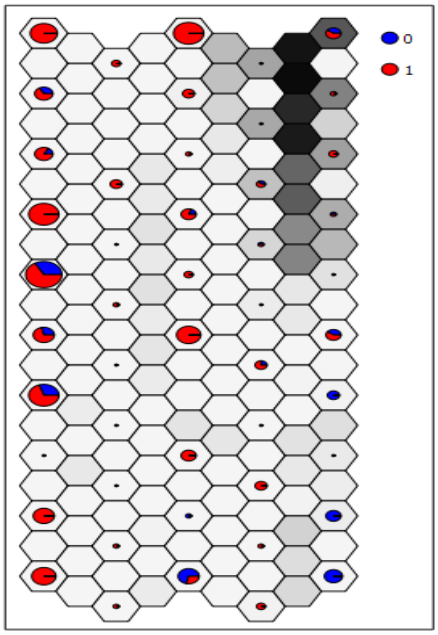

Fig. 5. SOM for Fo attribute

IV. CONCLUSION

Analysis of voice data is important in the present decade to understand and diagnostic methods for human diseases. The present method provides the diagnosis of PD using voice dataset through machine learning algorithms.

ACKNOWLEDGEMENT

Author would like to thank management and staff of Raghu Engineering College and GITAM University Visakhapatnam, India for their kind support in bringing out the above literature and providing lab facilities.

REFERENCES

[1] D.J. Gelb, E. Oliver, and S. Gilman, "Diagnostic criteria for Parkinson disease." Archives of neurology, vol. 56, no.1, pp. 33 1999.

[2] D. Aarsland, K. Andersen, J.P. Larsen, and A. Lolk, "Prevalence and characteristics of dementia in Parkinson disease: an 8-year prospective study." Archives of Neurology. Vol. 60, no. 3, pp. 387, 2003.

[3] R. Das, "A Comparison of multiple classification methods for diagnosis of Parkinson disease." Expert Systems with Applications, vol. 37, no. 2, pp. 1568-1572, 2010.

[4] R. Polikar, A. Topalis, D. Green, J. Kounios, and C. M. Clark, "Comparative multiresolution wavelet analysis of ERP spectral bands using an ensemble of classifiers approach for early diagnosis of Alzheimer's disease." Computers in biology and medicine, vol. 37, no. 4, pp. 542-558, 2007.

[5] C. O. Sakar, and O. Kursun, "Telediagnosis of Parkinson's disease using measurements of dysphonia." Journal of medical systems, vol. 34, no. 4, pp. 591-599, 2010.

[6] G. K. Wenning, Y. Ben-Shlomo, A. Hughes, S. E. Daniel, A. Lees, and N. P. Quinn, "What clinical features are most useful to distinguish definite multiple system atrophy from

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 3, September 2013

Parkinson's disease?." Journal of Neurology, Neurosurgery & Psychiatry, vol. 68, no. 4, pp. 434-440, 2000.

[7] A. Ozcift, "SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease." Journal of medical systems, vol. 36, no. 4, pp. 2141-2147, 2012.

[8] P. F. Guo, P. Bhattacharya, and N. Kharma, "Advances in detecting Parkinson's disease." In Medical Biometrics. Springer Berlin Heidelberg. pp. 306—314, 2010.

[9] D. S. V. G. K. Kaladhar, P. V. Nageswara Rao, and N. R.B. L. V. Ramesh, "Confusion matrix analysis for evaluation of speech on Parkinson disease using Weka and MatLab." International Journal of Engineering Science and Technology, vol. 2, no. 7, pp. 2734-2737, 2010.

[10] F. Åström, and R. Koker, "A parallel neural network approach to prediction of Parkinson's Disease." Expert systems with applications, vol. 38, no. 10, pp. 12470-12474, 2011.

[11] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, and I. M. Moroz, "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection'." Biomedical Engineering Online, vol. 6, pp. 23, 2007.

[12] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests." Biomedical Engineering, IEEE Transactions, vol. 57, no. 4, pp. 884-893, 2010.

[13] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression." In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference, pp. 594-597, 2010.

[14] K. Revett, F. Gorunescu, and A. Salem, "Feature selection in Parkinson's disease: A rough sets approach." In Computer Science and Information Technology, IMCSIT'09. International Multi conference, pp. 425-428, 2009.

AUTHORS BIOGRAPHY

Mr. Tarigoppula V.S Sriram: Mr. Sriram is a faculty member presently working in Raghu Engineering college, Visakhapatnam. His areas of specilizations are Data mining , mechine learning and Network security.

Dr. M. Venkateswara Rao: Dr. Rao is presently working as Associate Professor, GITAM University, Visakhapatnam. His area of specialization are Embeded systems, Robotics and Machine learing.

Dr. G V Satya Narayana: Dr. Satyanarayana is Professor in Dept. of IT, Raghu Institute of technology, visakhapatnem. His areas of specializations are Embeded systems and Robotics.

Dr. DSVGK Kaladhar: Dr. Kaladhar is an Assistant professor in Bioinformatics, GITAM University, Visakahapatnam. His areas of specializations are Competitional biology and Data mining

Mr. T Pandu Ranga Vital: Mr. Vital is a faculty member from Raghu college, Visakhapatnam. His areas of specialization are Data mining, Artificial intelligence and machine learning.