Predicting the occurrence of Parkinson's Disease using various Classification Models

ShreeragMarar[2],Debabrata Swain[1,2],Vivek Hiwarkar[2], Nikhil Motwani[2] and Akshar Awari[2]
[1]Assistant Professor in Computer Engineering,
[2]Vishwakarma Institute of Technology, Pune
Email:debabrata.swain7@yahoo.com

Abstract- Parkinson is a disease that directly degrades the functioning of central nervous system, more specifically the motor system.If diagnosed in a later stage, this disease may become incurable. Hence, it is necessary to diagnose the disease at an early stage.Voice frequency plays a vital role in the prediction of Parkinson disease. This paper presents the study for the diagnosis of Parkinson disease using various machine learning algorithmsthrough the amount of voice data attained from UCI repository [1]. The voice dataset consists of voice frequencies of 31 people with early-stage Parkinson's disease recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring. Various machine learning algorithms were applied on the dataset and among them ANN has shown highest accuracy (94.87%). Random Forest which is a Classification algorithmhasshown good accuracy (87.17%) while Naïve Bayes has shown least accuracy (71.79%). We have summarized all the results using the confusion matrix.

Index Terms- Voice dataset, SVM, ANN, Random forest.

I. INTRODUCTION

Diseases that attack nervous system are one of the major health concerns due to their increase in the mortality rate. It is said that deaths caused by disease that attack nervous system are going to surpass the death caused by cancer in the near future [2].

Parkinson is among the disease that attack central nervous system and mostly disturbs the motor system of the brain which leads to slowness of body movements, difficulty in walking, talking, thinking or carrying out other simple tasks. Parkinson is a progressive neuro-degenerative disease and as the time passes by symptoms becoming more and more severe. The cause of Parkinson disease is not identified till now. However, it is researched that combination of genetic and environmental factors play animportant part in causing Parkinson Disease. There does not exist any medical action for curing Parkinson Disease and medications are provided only to release its symptoms [7]. The symptoms of the Parkinson disease include tremor, Slow movement (bradykinesia), rigidity in muscles, impaired posture and balance, loss of automatic movements, writing and speech changes [4]. Currently the diagnosis of Parkinson Disease is done by means of neuropathological examination which consumes a lot of time and results of which are not accurate. The results given by this neuropathological examination are in the form of certain possibility level of having Parkinson disease[3], [8]. Speech disorder has been found as one of the major indicative symptoms of Parkinson Disease with 90% of people having attenuation of voice suffered from Parkinson Disease [5] – [9]. Such weakening covers the one in the speech creation of regular vocal voice (dysphonia) and the difficulties in enunciation and speaking capability (dysarthria).

The indicator of dysphonia includes tremor in speaking, panting breath, sound shrinkage and throatiness. In view of this relation between Parkinson and dysphonia, the verbal sound of patient having Parkinson Disease later develops a fear by most investigators and is hence used as the constraint for the diagnosis of Parkinson disease. In addition to this using vocal voice as a part of diagnosis, also has some benefits; some of which are correlated to itseasiness and non-intrusiveness for the medical experts. Little et al. has examined the use of vocal voice for diagnosing Parkinson disease[6], [13]. Here, the use of dimension method included the degree of sound loudness or pressure, jitter (breadth variation from one wave to other) (F0 variation from one sound wave to other, shimmer),noise-to-harmonics proportion (amplitude noise with sound signal) and the degree of tone. In this paper we have used various machine learning algorithm such as Logistic Regression, KNN, SVM, Kernel-SVM, Naïves Bayes, Decision Tree, Random Forest, ANN on voice dataset provided by the UCI Machine Learning Repository[1] and compared the results of the same with ANN showing a highest accuracy among other Machine Learning algorithms. The rest of the paper is organized as follows. Section 2 contains the Methodology. Section 3 contains the methods. Section 4 contains the detailed information about the attributes present in the dataset. Section 5 provides the results and discussions from the machine learning algorithm implemented. And finally, the conclusion of the work is provided in Section 6.

II. METHODOLOGY

The voice dataset for Parkinson disease has been obtained from UCI Machine learning repository from Centre for Machine Learning and Intelligent Systems [1]. The dataset consists of biomedical voice measurements from 31 people, where 23 people are showing Parkinson's disease. The last column indicates the "status" which is set to 0 for healthy and 1 for PD. ASCII CSV format has been used in the dataset. The data contains 23 attributesand196 number of instances. The dataset has been retrieved and executed for classification and

978-1-5386-5367-8/18/$31.00 ©2018 IEEE

supervised learning methods (Logistic regression, Decision tree, random forest, Naive Bayes,SVM,k-nearest neighbour and ANN). R language is used to perform the complete analysis and comparison of different models. R provides a good GUI and easier plotting of graphs to visualize the results. All the libraries required for each model are open source and can be downloaded and imported as required. Libraries provides the functions required for modelling, predicting and plotting (if possible). After data pre-processing, 80% instances were used for training and 20% for testing. The accuracy obtained by executing the test set is summarized with the help of confusion matrix for each model.

III. MODELS

Logistic regression is the probabilistic model and can be used when the dependent variable is binary. The logistic regression is a predictive analysis, similar to all other regression analysis. The relationship between one or more nominal/interval/ordinal/ratio-level independent variables and the binary dependent variable can be interpreted well with the help of logistic regression.

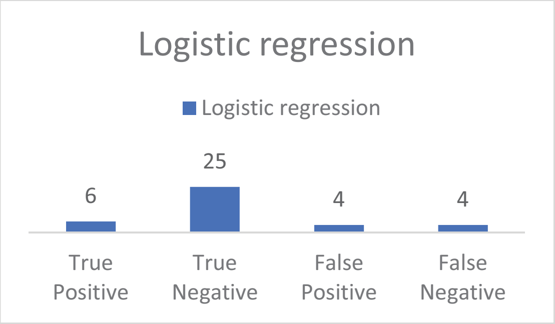

Figure 1 Confusion matrix for Logistic Regression

k-nearest neighbours algorithm (k(10)-NN) can be used as a regressor or a classifier and is a non-parametric method. It is a pattern recognition model and is also called as a slow learner because at the time of learning, it doesn't build any classification model. It just stores the training set. The output in KNN is a class membership. When a new object comes in, its nearest K neighbours are categorized into the classes they belong and the new object is categorized into that particular class that has received the maximum votes. In k-NN regression, the new object acquires the value which the average of the values of its K nearest neighbours.

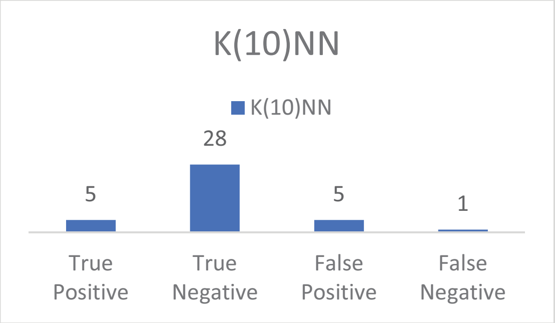

Figure 2 Classification matrix of KNN

Support vector machines (SVMs) [14] comes under supervised learning model and analyses the data used for regression/classification analyses. The SVM training algorithm uses the training samples each of which belongs to either of the 2 categories and categorizes the new entity into one class or other which makes it a non-probabilistic binary linear classifier. The SVM model considers the instances as points in space so that the instances belonging to different categories are separated by a clear wide gap. When a new object has to be classified, it is mapped to that same space. Based on the gap to which it belongs, it is predicted to belong to that category.

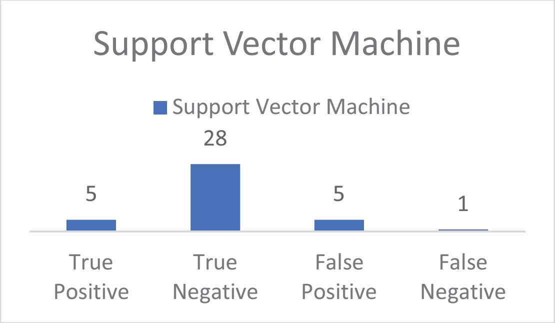

Figure 3 Confusion matrix for SVM

Kernel methods(SVM) comes under the class of pattern analysis and Support Vector Machine(SVM) is its most widely used method The main objective of pattern analyses methods is to learn and find the general types of relations (for example principal components, rankings, clusters, classifications, correlations) in datasets. The raw data in the dataset has to be obviously transformed into feature vector representations that generally is a user specified feature map. The kernel SVM usually requires a user specified kernel which can be imported from the R libraries.

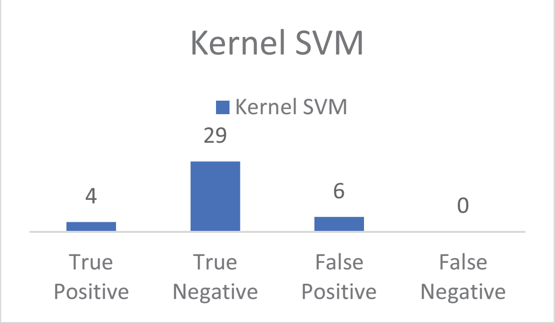

Figure 4 Confusion matrix for Kernel SVM

978-1-5386-5367-8/18/$31.00 ©2018 IEEE

Naive Bayes classifiers are application of Bayes theorem. It is based on applying Bayes theorem with strong independent assumptions between the features. It is also known as probabilistic classifier because of probabilistic relationship between the class and the features. It doesn't follow a deterministic relationship and is highly scalable. The training is done in linear time by evaluating a closed form expression unlike using iterative approximation used by many other classifiers.

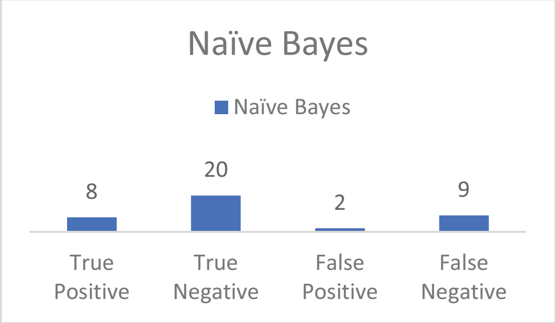

Figure 5 Confusion matrix for Naïve Bayes

Decision Tree is a tree like representation made from nodes and arcs. The decisions are made in the nodes and based on the result obtained from that node, one of the children is chosen. This continues till the leaf node is reached where the final class of the instance is decided. It is widely used in data mining and machine learning models. The process initiates from the root and terminates in one of the leaves. The conjunctions of features is represented by branches which ultimately leads to the class labels situated at the leaves. Regression trees are those where the target variables can hold continuous values.

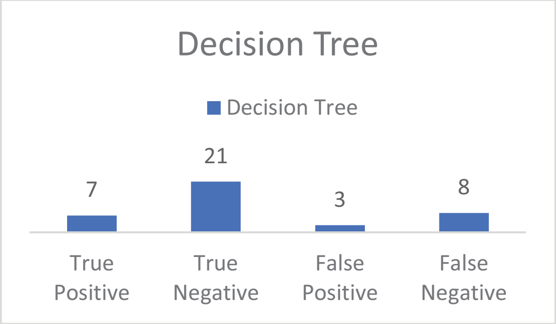

Figure 6 Confusion matrix for Decision Tree

Random forests are extension of decision trees. They construct multiple decision trees which are made from random overlapping sets of data and attributes. The combined feature of overlapping sets and random selection helps to get rid of overfitting. It uses ensemble learning technique for regression/classification and

outputs the result based on the mean prediction of the individual trees. The individual errors and limitations of decision trees can be overcome by taking combined decision of large number of trees.

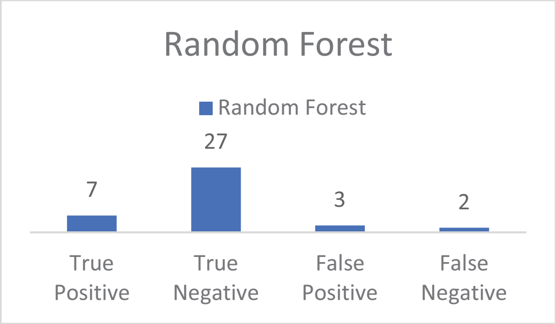

Figure 7 Confusion matrix for Random Forest

Artificial neural networks (ANNs)[14] are the computing systems elusively inspired by the biological neural networks that resemble the animal brains. They are compute intensive systems that learn progressively by improving its prediction in each iteration by considering the examples usually without task specific programming. In image recognition, (to identify dogs from dog and cat images) they might analyse images that might be labelled as 'dog' and 'cat' and use that to separate dogs from cats. It is done by the neural networks without any a priori knowledge of dogs and cats (eg-fur, tail, etc.). Instead, a set of relevant characteristics is evolved by them from the learning material they process.

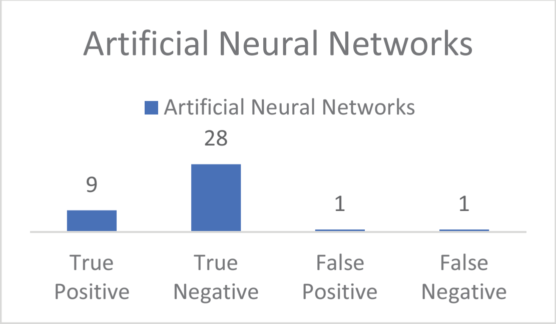

Figure 8 Confusion matrix for ANN

IV. DATASET ATTRIBUTES

Attributes	Description
Name	Subject name and recording number
MDVP: Fo(in Hz)	Average vocal essential frequency
MDVP: Fhi(in Hz)	Maximum vocal essential frequency
MDVP: Flo(in Hz)	Minimum vocal essential frequency
MDVP: jitter(in %)	Some measures of variation in essential frequency
MDVP: jitter(abs)	
MDVP: RAP	
MDVP: PPQ	
jitter: DDP	
MDVP: shimmer	Numerous measures of variation in essential amplitude
MDVP: shimmer(in dB)	
Shimmer: APQ3	
Shimmer: APQ5	
MDVP : APQ	
Shimmer : DDA	
NHR	Ratio of noise to tonal components in the voice
HNR	
Status	Health status of the subject (one) - Parkinson's, (zero) - healthy
RPDE	Two nonlinear dynamical complexity measures
D2	
DFA	Signal fractal scaling exponent
Spread1	Three nonlinear measures of essential frequency variation
Spread2	
PPE	

Efficiency of Algorithms are calculated using Sensitivity, Specificity and Accuracy given in Equitation 1, 2, 3.

$$\boxed{Sensitivity = \frac{TP}{TP + FN}} \quad \text{-(1)}$$

$$\boxed{Specificity = \frac{TN}{FP + TN}} \quad \text{-(2)}$$

$$\boxed{Accuracy = \frac{TP + TN}{P + N}} \quad \text{-(3)}$$

Where, TP is True Positive and FP is False Positive

TN is True Negative and FN is False Negative

P is Summation of all Positives

N is Summation of all Negatives

V. RESULTS

Results were calculated using a confusion matrix. The test set consisted of 20% of the original dataset and the accuracy was calculated on the basis of the confusion matrix. The best result was obtained with Artificial neural networks with an accuracy of 94.87%.

The least was obtained with decision tree which was 71.79%. All the models were implemented in R. In ANN the model was trained using H2O package and the accuracy was obtained after performing parameter tuning.

Classification Models	**Sensitivity**	**Specificity**	**Accuracy**
Logistic Regression	60	86.2	79.48
K(10)NN	83.33	84.84	84.61
SVM	83.33	84.84	84.61
Kernel SVM	100	82.85	84.61
Naïve Bayes	47.05	90.90	71.79
Decision Tree	46.66	87.50	71.79
Random Forest	77.77	90	87.17
ANN	90	96.55	94.87

Table 1 Performance evaluation results

VI. CONCLUSION

On the basis of the obtained results following conclusion were made. The Artificial Neural Networks has given the best accuracy for prediction of the status of the patient. The algorithms such as Logistic Regression and Decision trees have a prediction accuracy of around 70%. In all, 8 models were implemented to predict the occurrence of Parkinson’s Disease. Various models were implemented in order to test which one gives the best result. As ANN iscompute-intensive, one can also adopt random forest for better accuracy than other models.

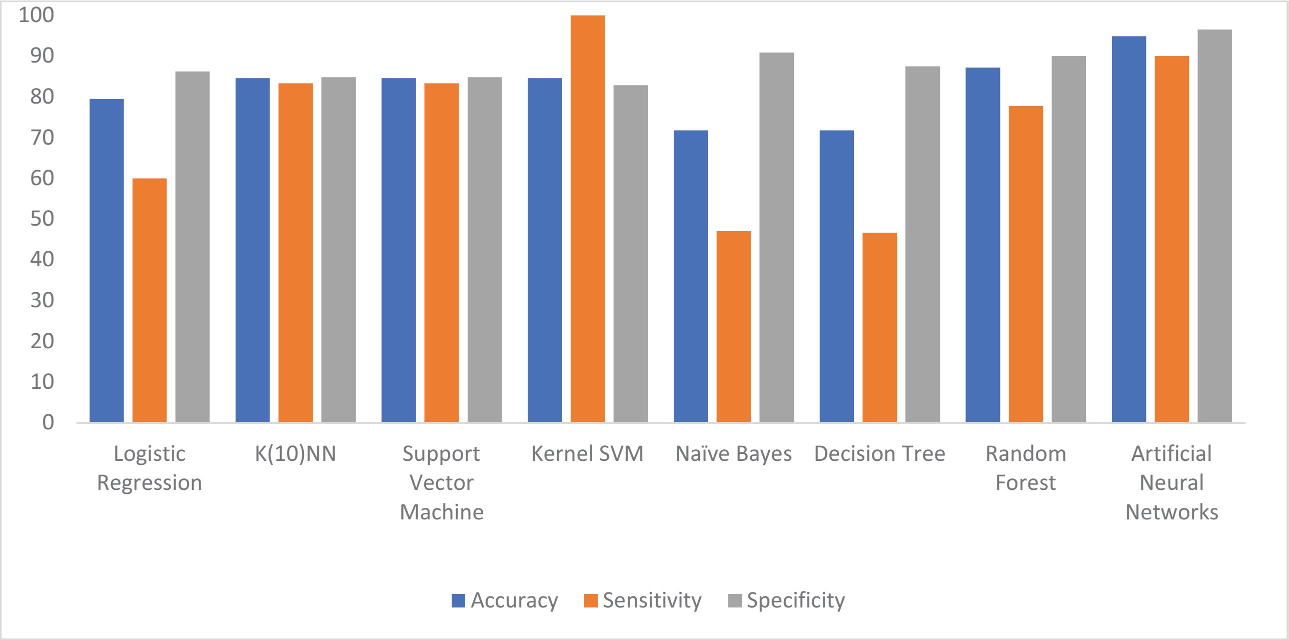

Figure 9Performance evaluation of Classification Models

REFERENCES

[1] Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM, 'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

[2] A. M. Lozano andA. E. Lang, "Parkinson's disease," New England Journal of Medicine, vol. 339, no. 15, pp. 1044–1053, 1998.

[3] I. Litvan, Z. K. Wszolek ,H. Braak, J. E. Duda, C. Duyckaerts, T. Gasser,G. M. Halliday, J. Hardy, J. B. Leverenz, K. DelTredici, andD. W. Dickson, "Neuropathological assessment of Parkinson's disease: refining the diagnostic criteria," The Lancet. Neurology, vol. 8, pp. 1150–1157, Dec. 2009.

[4] S. Fahn, "Description of Parkinson's disease as a clinical syndrome,"Annals of the New York Academy of Sciences, vol. 991, pp. 1–14, June 2003.

[5] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, "Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease," IEEE Transactions on Biomedical Engineering, vol. 56, pp. 1015–1022, Apr. 2009.

[6] S. J. Roberts, M. A. Little, P. E. McSharry D. A. Costello, and I. M. Moroz, "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection," BioMedical Engineering OnLine, vol. 6, no. 1, p. 23, 2007.

[7] Y. E. Choonara,N. Singh, andV. Pillay,, "Advances in the treatment of Parkinson's disease," Progress in Neurobiology, vol. 81, pp. 29–44, Jan. 2007.

[8] P. L. De Jager,J. M. Shulman, and M. B. Feany, "Parkinson's Disease: Genetics and Pathogenesis," Annual Review of Pathology: Mechanisms of Disease, vol. 6, pp. 193–222, Feb. 2011.

[9] D. Van Lancker Sidtis, K. Cameron, and J. J. Sidtis, "Dramatic effects of speech task on motor and linguistic planning in severely dysfluent parkinsonian speech," Clinical linguistics & phonetics, vol. 26, pp. 695– 711, Aug. 2012.

[10] J. Zhang, W. Xu, Q. Zhang, Bo Jin, X. Wei," Exploring Risk Factors and Predicting UPDRS Score Based on Parkinson's Speech Signals", 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services.

[11] Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM, 'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection,' BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

[12] M. S. Wibawa, H. A. Nugroho, N. A. Setiawan," Performance Evaluation of Combined Feature Selection and Classification Methods in Diagnosing Parkinson Disease Based on Voice Feature", ICSITech, 2015.

[13] S. Sapir, L. Ramig, and C. Fox, "Speech and swallowing disorders in Parkinson disease:," Current Opinion in Otolaryngology & Head and Neck Surgery, vol. 16, pp. 205–210, June 2008.

[14] David Gila, Magnus Johnson B," Diagnosing Parkinson by using Artificial Neural Networks and Support Vector Machines",GlobalJournel of Computer Science and Technology.

[15] T. V.S Sriram, M. V. Rao, G V Satya Narayana , DSVGK Kaladhar, T. P. R. Vital," Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms", International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 3, September 2013. pp. 1568–1572, Mar. 2010.