International Conference on Intelligent Technologies & Science - 2021
(ICITS-2021)

Parkinson's Disease Prediction using Machine Learning Over Big Data

Vighnesh Garg[1], Anurag Jaiswal[2]
Computer Science and Engineering,
Goel Institute of Technology & Management, Lucknow, India
vighnesh017@gmail.com, anurag.jaiswal@goel.edu.in

Abstract: The Parkinson disease prediction has attracted many researchers over a decade. The Parkinson disease dataset is available on the uci machine learning repository website. Researchers have used the various techniques of data mining, machine learning and deep learning to build the prediction system for early detection of Parkinson disease at an early stage so that it can be controlled because the health care for this disease is not affordable for developing and under developed countries. In our proposed work we will pre-process the data based on the exploratory data analysis. And build a prediction system for the Parkinson disease by optimizing machine learning models like kneighbors, logistic regression, decision tree classifier, random forest classifier. All the algorithm is employed to predict the data set and a multi-dimensional result analysis and choose the best fitting algorithm. Thee dataset is generated on study of 31 person out of which 23 patients suffers with the disease.

Keywords: Parkinson's Disease, Classification, Random Forest, Support Vector Machine, Machine Learning.

1. Introduction:

Parkinson disease (PD) is a neurological disorder based on dopamine receptors. Parkinson disease mostly causes problems in moving around. It can cause a person to move very slowly. Parkinson is a progressive neurological condition, which is characterized by both motor (movement) and non-motor symptoms. Apart from many common symptoms each person will experience and demonstrate an individual presentation of the condition. A person with Parkinson disease appears stiff or rigid. At times, a person with Parkinson disease may appear to suddenly "freeze up" or be unable to move for a short period of time. Parkinson disease is a progressive neurodegenerative condition resulting from the death of the dopamine containing cells of the substantia nigra. There is no consistently reliable test that can distinguish Parkinson disease from other conditions that have similar clinical presentations. The diagnosis is primarily a clinical one based on the history and examination.

People with Parkinson disease classically present with the symptoms and signs associated with Parkinsonism, namely hypokinesia (i.e. lack of movement), bradykinesia (i.e. slowness of movement), rigidity (wrist, shoulder and neck.) and rest tremor (imbalance of neurotransmitters, dopamine and acetylcholine). Parkinsonism can also be caused by drugs and less common conditions such as: multiple cerebral infarction, and degenerative conditions such as progressive supra nuclear palsy (PSP) and multiple system atrophy (MSA).

Although Parkinson disease is predominantly a movement disorder, other impairments frequently develop, including psychiatric problems such as depression and dementia. Autonomic disturbances and pain may later ensue, and the condition progresses to cause significant disability and handicap with impaired quality of life for the affected person. Family and carers might get affected indirectly.

Neurodegenerative disorders are the results of the progressive tearing and neurons loss in different areas of the nervous system. Neurons are the functional unit of brain .They are contiguous rather than continuous. A good healthy looking neuron has extensions called dendrites or axons, a cell body and a nucleus that contains our DNA. DNA is our genome and hundred billion neurons contains our entire genome which is packaged into it .When a neuron get sick, it loses its extension and hence its ability to communicate which is not good for it and its metabolism become low so it starts to accumulate junk and it tries to contain the junk in the little packages in little pockets .When things become worse and if the neuron is a cell culture it completely loses its extension, becomes round and full of the vacuoles.

2. Related Work:

Indira R. et al. (2014) have proposed an automatically machine learning approach and detected the Parkinson disease on behalf of speech/voice of the person. The author used fuzzy C-means clustering and pattern recognition based approach for the discrimination between healthy and parkinson disease affected people. The authors of this work have achieved 68.04% accuracy, 75.34% sensitivity and 45.83% specificity.

Indira R. et al. (2014) have proposed a back propagation based approach for the discrimination between healthy and parkinson diseases affected peoples with the help of artificial neural network. Boosting was used by filtering technique, and for data reduction principle component analysis was used.

Geeta R. et al. (2012) have investigated and performed the feature relevance analysis to calculate the score to classify the Parkinson diseases Tele-monitoring dataset and dataset comparison classes Motor-UPDRS and Total-UPDRS (Unified Parkinson Disease Rating scale).

Rubén A. et al. (2013) proposed a five different classification paradigms using a wrapper feature selection scheme are capable of predicting each of the class variables with estimated accuracy in the range of 72–92%. In addition, classification into the main three severity categories (mild, moderate and severe) was split into dichotomy problems

where binary classifiers perform better and select different subsets of non-motor symptoms.

Betala E. et al. (2014) proposed a SVM and k-Nearest Neighbour (k-NN) Tele-monitoring of PD patients remotely by taking their voice recording at regular interval. The age, gender, voice recordings taken at baseline, after three months, and after six months are used as features are assessed. Support Vector Machine was more successful in detecting significant deterioration in UPDRS score of the patients.

3. Methodology

This section explains the steps taken to achieve the prediction of Parkinson's disease using various machine learning. The various steps taken are Data gathering , Data Preprocessing, Model Selection, Training, Evaluation, prediction .

3.1 Data Gathering

The first step is Data gathering .This step is very important because the quality and quantity of the data you gather will directly affects the level of your prediction model. So we have taken data of different voice recordings of the patient.

3.2 Data preparation

In this step the data is visualized well to spot the relationship between the parameters present in the data so as to take the advantage of as well as to get the data imbalances. With this ,we need to split the data into two parts .The first part for training the model like in our model we have used 70 percent of data for training and 30 percentage for testing. Which is the second part of the data

3.3 Model Selection

The next step in our workflow is model selection. There are various models that have been used till date by researchers and scientist. Some are meant for image processing ,some for sequences like text, numbers or patterns. In our case we have 26 features which defines the voice recording of various patients so we have chosen such models which will classify or differentiates the unhealthy patient with the healthy one.

3.4 Training

Training the dataset is one of the main task of machine learning .we will apply the data to progressively improve the selected model's ability to predict better ie the actual result should be approx. to predict one.

3.5 Evaluation

The metrics we have calculated are ROC, Accuracy , Specificity , Precision etc. which will highlights the best algorithm among all.

3.6 Prediction

In this phase we finally get the model ready to detect the prediction of Parkinson's disease based on the given dataset.

3.7 Dataset

The Parkinson disease data set is available on the uci machine learning repository. The data set was originally created by the University of Oxford, test had been conducted on 31 people and out of which 23 people suffered from the Parkinson disease. All the voice based test were conducted on these subjects and the data is prepared .the data set contains 198 rows and 23 columns. The columns are sex the gender of the patients, age of patients, subject# the subject ids, name of the participating patients, MDVP:Fo(Hz) - mean vocal basic frequency, MDVP:Fhi(Hz) - Max vocal basic frequency, MDVP:Flo(Hz) – Min vocal basic frequency MDVP:Jitter(%) , MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP other types measurements of variation in basic vocal frequency, MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ , Shimmer:DDA – various measures of vocal amplitude, NHR, HNR - Two measures of ratio of noise to tonal components in the voice, status contains o and 1 respectively for the patient with and without the Parkinson disease, PDE, D2 - Two nonlinear dynamical complexity measures DFA Signal fractal scaling exponent, spread1, spread2, PPE - Three nonlinear measures of fundamental frequency variation.

3.8 Proposed model

The dataset is further treated for the missing values by using the mean of values based on the status. The exploratory data analysis is performed on the dataset which includes Pearson's correlation technique, box plot of different features set, violin plot, the box plot and violin plot gives an insight of the data distribution. The data set is scaled using the standard scalers which scales the values by using the standard deviation of the values contained in the data set. Then the data is split in 80:20 ratio for the purpose of training and test set. The data set is used to train the random forest classifier. The random forest classifier is a ensemble learning based boosting technique. The boosting technique used to develop the random forest classifier is the bootstrap based boosting technique, which divides the data into multiple portion and each portion is called as the bag. Each bag is used to build a decision tree.

The test data is predicted by using the all the decision trees and the mode of the prediction is the result of the random forest classifier or we have the averaging technique. The random forest classifier is hyper tuned for the number of estimators, which is the number of decision tree to be built on the given data set. For hyper tuning of model we have the grid searchcv model which will fit the model with the given options for the number of estimators and the value with best prediction accuracy, best fitting time, best prediction time and the best training time is returned as the best fit parameter. The grid search cv internally uses the cross validation with default of 5 K-fold. The random forest model is compared to the other algorithms like Kneighbors, Logistic regression model, Decision Tree Classifier.

All the other models are also hyper tuned for different parameters like Kneighbors is tuned for the number of neighbors parameters, logistic regression model is hyper tuned for the parameters like learning rate (c), penalty for which we

have 2 options primarily l1 and l2, and the learning rate. The decision tree classifiers don't considers all the features of the features set used for training the model it considers only the important features out of the feature set. The decision tree uses a segmented tree to store the model where each node contains the best split value along with the operators for comparison like == and != if the feature is categorical and >= and < if the feature is continues. The decision tree is used for the maximum depth of tree that is the length of the tree. The data set is of type multinomial. So the result is measured by using the prediction accuracy of the models. the block diagram of the proposed model is as shown in the figure 1.

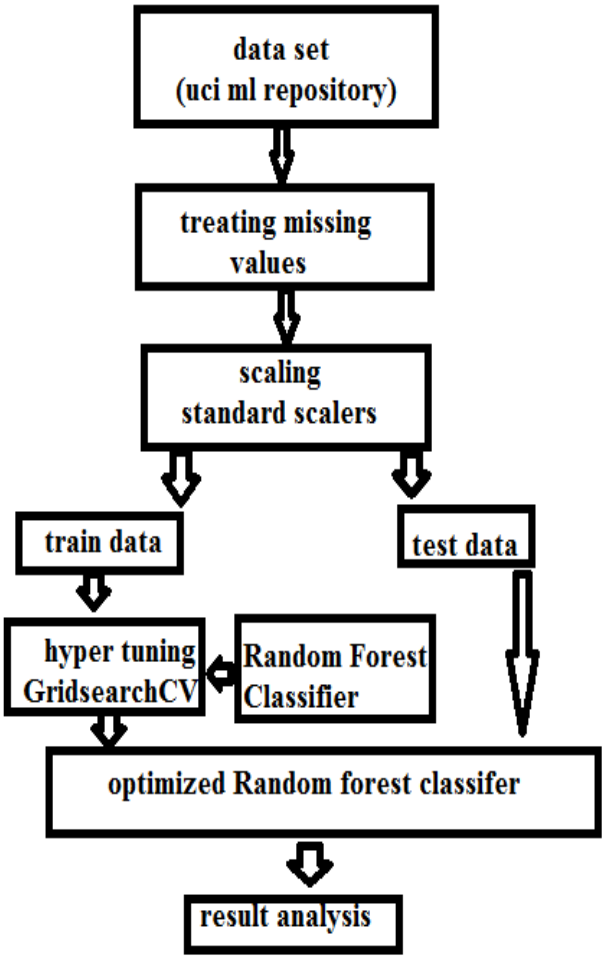

Fig. 1: The Proposed Prediction Model

4. Result and Discussion:

The Parkinson Disease data set is available on the uci machine learning repository. The data set contains 23 columns. The Parkinson disease data set is multinomial classification problem. For implementation of the model we have used the python language. It is a dynamically typed and a interpreted high level programming language. It is a popular technology now a days in field of AI, machine learning, deep learning. Firstly we have to perform the data pre-processing. For choosing the better data pre-processing models we will use the exploratory data analysis. The exploratory data analysis includes the count plot, box plot, and violin plot. With the help of box plot and violin plot we will find the correlation among the columns. For the visualization of correlation of entire data we have used the heat map. The figure 2 shows the total male and female patients records contained in dataset.

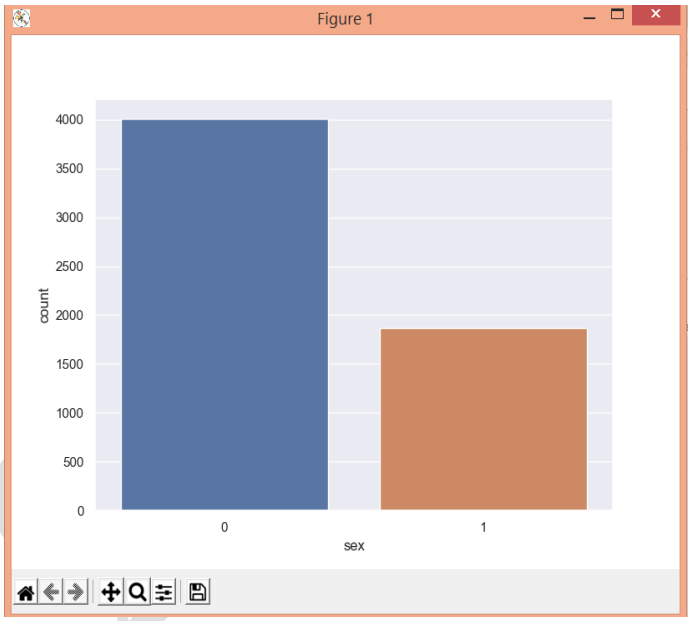

Fig 2: Total Number Of Male And Female Patients

The box plot is used to find the distribution of data upon a factor. The box plot includes min, max, 25 percentile,50 percentile (which is often termed as median), 75 percentile, the combination of 25,50,75 percentile is often termed as IQR(*interquartile* range).in the figure 3 we have used box plot to check the data distribution of sex column versus subject# and sex versus the age. Where sex is gender of patients and subject # represents either the controlled patients or the Parkinson disease patients.

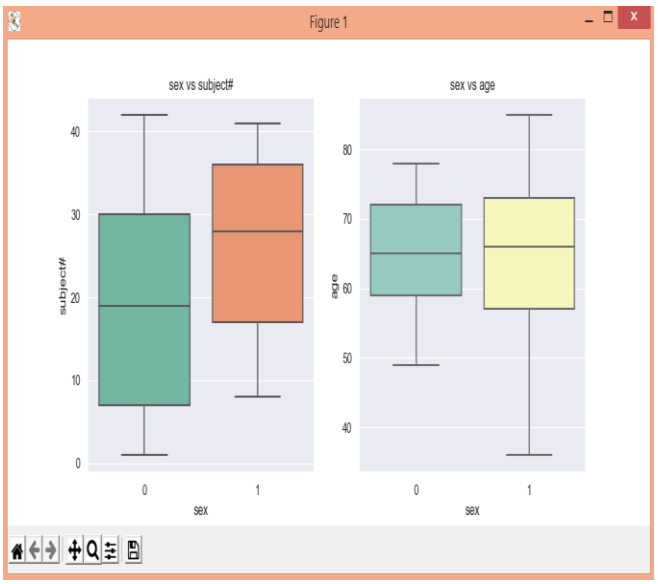

Fig 3: The Box Plot Of Sex Versus Subject# And Age

The violin plot is similar to the box plot it additionally includes the probability density of the data it is much informative when compared to the box plot. The only difference between the box plot and the violin plot is if the data contains multiple peaks (multimodal) then the box plot includes only one peak whereas the violin plot includes the multiple peaks. The figure 4 shows the violin plot of status versus the RPDE and DFA. The figure 5 shows the box plot of status versus the NHR and HNR.

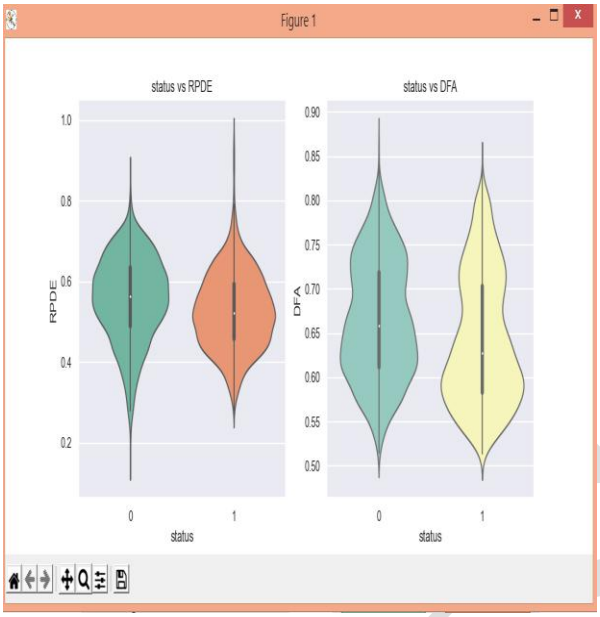

Fig 4: The Violin Plot Of Status Versus RPDE and DFA.

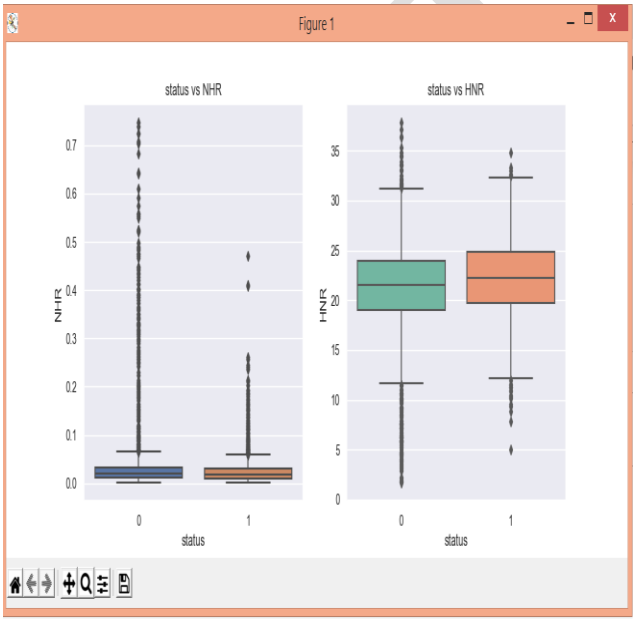

Fig 5: the box plot of status versus NHR and HNR.

The other model used for the prediction system is the ensemble learning based bagging method classifier that is Random forest classifier. The random forest classifier is hyper tuned for the parameters of the number of estimators. The number of estimators represents the number of decision tree models to be used for the purpose of building the predation model. The random forest classifier is uses all the base estimators to predict the given sample and the mode of the result is returned as an output. The random forest classifier hyper tuning resulted the best accuracy for the number of estimators to be 350. The result of prediction accuracy shows that the tuned model is better than the non- tuned model. The results are shown in figure 6. The prediction accuracy of the tuned Kneighbor model is 90.41, tuned logistic regression model is 98.98, tuned decision tree classifier is 98.75 and the tuned random forest classifier is 99.54. The results are shown in the figure 6. for the result analysis the cross validation with k-fold of 5 is used. The prediction model for the Parkinson disease is built using the tuned random forest classifer.

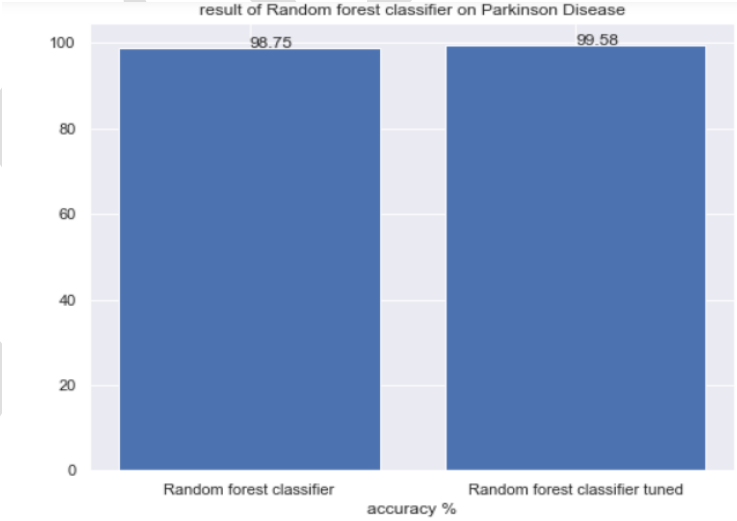

Fig 6: The Prediction Accuracy Of Random Forest Classifier.

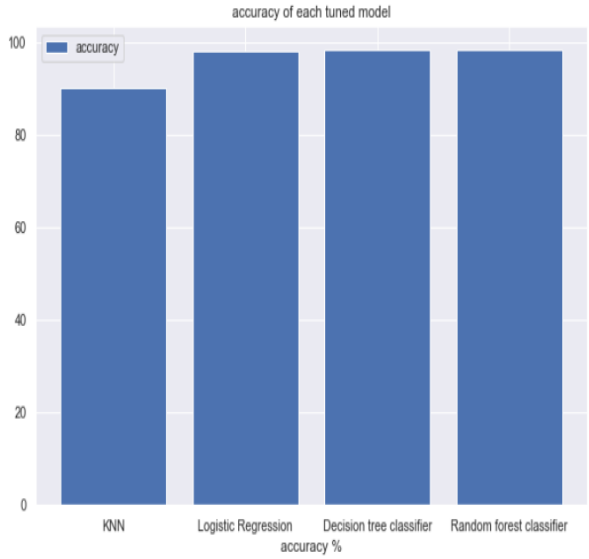

Fig. 7: The Prediction Accuracy Of All The Models.

5. Conclusion:

It is much important to detected the Parkinson disease at the initial stage itself for which we have developed the machine learning based prediction system of the Parkinson disease. The dataset is available on the uci machine learning repository which is used to build the proposed model. Various machine learning models were used on the dataset. Before using the machine learning on the dataset we have performed the exploratory data analysis and done the pre-processing which includes scaling and feature extraction. The models were hyper tuned using the gridsearchcv module. All the models were trained using the Parkinson disease and the results shows that random forest is the best model for the prediction. And the accuracy of the models is 99.54 percent

The dataset is mainly collected from people of England. The model can be improved by using various different kind of dataset from different part of the world do that the model would be versatile.

References:

[1] Rustempasic, Indira, & Can, M. (2013). Diagnosis of Parkinson's Disease using Fuzzy C-Means Clustering and Pattern Recognition. SouthEast Europe Journal of Soft Computing, 2(1).

[2] Rustempasic, I., & Can, M. (2013). Diagnosis of Parkinson's disease using principal component analysis and boosting committee machines. SouthEast Europe Journal of Soft Computing, 2(1).

[3] Ramani, D. R. G., Sivagami, G., & Jacob, S. G. (2012). Feature Relevance Analysis and Classification of Parkinson Disease Tele- Monitoring Data Through Data Mining Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 2(3).

[4] Armañanzas, Ruben, Bielza, C., Chaudhuri, K. R., Martinez-Martin, P., & Larrañaga, P. (2013). Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach. Artificial intelligence in medicine, 58(3), 195-202.

[5] Sakara, Batalu. E., & Kursunb, (2014)O. Telemonitoring of changes of unified Parkinson's disease rating scale using severity of voice symptoms.

[6] Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). Novel speech signal processing algorithms for highaccuracy classification of Parkinson's disease. Biomedical Engineering, IEEE Transactions on, 59(5), 1264-1271.

[7] Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2011). Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. Journal of the Royal Society Interface, 8(59), 842-855.

[8] Sharma, A., & Giri, R. N. (2014) Automatic Recognition of Parkinson's disease via Artificial Neural Network and Support Vector Machine.

[9] Khemphila, A., & Boonjing, V. (2012). Parkinsons disease classification using neural network and feature selection. World Academy of Science, Engineering and Technology, 64, 15-18.

[10] Revett, K., Gorunescu, F., & Salem, A. M. (2009, October). Feature selection in Parkinson's disease: A rough sets approach. In Computer Science and Information Technology, 2009. IMCSIT'09. International Multiconference on (pp. 425-428). IEEE.

[11] Shahbakhi, M., Far, D. T., & Tahami, E. (2014). Speech Analysis for Diagnosis of Parkinson's Disease Using Genetic Algorithm and Support Vector Machine. Journal of Biomedical Science and Engineering, 2014.

[12] Chen, A. H., Lin, C. H., & Cheng, C. H. New approaches to improve the performance of disease classification using nested– random forest and nested–support vector machine classifiers.

Cancer, 2(10509), 102.

[13] Bocklet, T., Noth, E., Stemmer, G., Ruzickova, H., & Rusz, J. (2011, December). Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis. In Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on (pp. 478-483). IEEE.

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

[15] Ene, M. (2008). Neural network-based approach to discriminate healthy people from those with Parkinson's disease. Annals of the University of Craiova-Mathematics and Computer Science Series, 35, 112-116.