International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

The Big Data Prediction on Parkinson's Disease using Machine Learning Approach

Neelesh Rawat
Electronics & Telecommunication
MITS, Gwalior, India
Rawatneelesh0@gmail.com

Dr. Anshu Srivastava
Dept of CSE
QRAT, Lucknow, India
Anshuqrat14@gmail.com

Abstract: The Parkinson disorder prediction has attracted many researchers over a decade. The Parkinson disease dataset is available on the uci machine studying repository website. Researchers have used the diverse techniques of statistics mining, system getting to know and deep getting to know to build the prediction gadget for early detection of Parkinson disorder at an early stage so that it could be managed due to the fact the fitness take care of this disease isn't less costly for developing and underneath developed countries. In our proposed paintings we will pre-technique the facts primarily based at the exploratory data analysis. And build a prediction machine for the Parkinson sickness by optimizing machine gaining knowledge of fashions like kneighbors, logistic regression, choice tree classifier, random wooded area classifier. All the algorithm is employed to predict the information set and a multi-dimensional end result evaluation and pick out the pleasant becoming algorithm. Thee dataset is generated on take a look at of 31 individual out of which 23 patients suffers with the disorder.

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

1. Introduction:

Parkinson disease (PD) is a neurological disorder based totally on dopamine receptors. Parkinson disease commonly reasons problems in moving around. It can reason a person to move very slowly. Parkinson is a innovative neurological situation, which is characterised via each motor (motion) and non-motor signs. Apart from many not unusual signs every body will experience and display an individual presentation of the condition. A person with Parkinson disease appears stiff or inflexible. At times, someone with Parkinson disorder may seem to all at once "freeze up" or be unable to transport for a short time frame. Parkinson ailment is a revolutionary neurodegenerative circumstance on account of the demise of the dopamine containing cells of the substantia nigra. There is no consistently dependable check that may distinguish Parkinson ailment from different situations that have comparable medical shows. The prognosis is more often than not a medical one based totally on the records and exam.

People with Parkinson disease classically present with the signs and signs and symptoms related to Parkinsonism, specifically hypokinesia (i.E. Lack of movement), bradykinesia (i.E. Slowness of motion), rigidity (wrist, shoulder and neck.) and relaxation tremor (imbalance of neurotransmitters, dopamine and acetylcholine). Parkinsonism also can be due to drugs and less commonplace situations together with: multiple cerebral infarction, and degenerative situations consisting of innovative supra nuclear palsy (PSP) and a couple of device atrophy (MSA).

Although Parkinson disease is predominantly a motion sickness, different impairments frequently increase, including psychiatric issues inclusive of melancholy and dementia. Autonomic disturbances and pain might also later happen, and the circumstance progresses to motive sizable disability and handicap with impaired pleasant of lifestyles for the affected individual. Family and carers would possibly get affected indirectly.

Neurodegenerative problems are the outcomes of the modern tearing and neurons loss in exclusive areas of the fearful gadget. Neurons are the practical unit of mind .They are contiguous in preference to continuous. A exact wholesome looking neuron has extensions known as dendrites or axons, a cellular body and a nucleus that carries our DNA. DNA is our genome and hundred billion neurons incorporates our whole genome that's packaged into it .When a neuron get sick, it loses its extension and consequently its ability to talk which is not properly for it and its metabolism become low so it starts to build up junk and it attempts to contain the junk within the little programs in little wallet .When matters end up worse and if the neuron is a cell subculture it absolutely loses its extension, will become spherical and full of the vacuoles.

2. Related Work:

Indira R. et al. (2014) Have proposed an automatically machine gaining knowledge of method and detected the Parkinson sickness on behalf of speech/voice of the person. The author used fuzzy C-approach clustering and pattern popularity based totally approach for the discrimination between wholesome and parkinson ailment affected people. The authors of this paintings have completed 68.04% accuracy, seventy five.34% sensitivity and forty five.Eighty three% specificity.

Indira R. Et al. (2014) have proposed a returned propagation based approach for the discrimination among healthful and parkinson illnesses affected peoples with the help of artificial neural community. Boosting was utilized by filtering method, and for records discount precept issue evaluation become used.

Geeta R. Et al. (2012) have investigated and performed the feature relevance evaluation to calculate the rating to categorise the Parkinson diseases Tele-tracking dataset and

dataset contrast classes Motor-UPDRS and Total-UPDRS (Unified Parkinson Disease Rating scale).
Rubén A. Et al. (2013) proposed a 5 one of a kind category paradigms using a wrapper feature choice scheme are capable of predicting each of the elegance variables with predicted accuracy inside the variety of 72–92%. In addition, classification into the primary 3 severity classes (mild, moderate and severe) become break up into dichotomy issues wherein binary classifiers perform higher and pick out distinct subsets of non-motor signs and symptoms.
Betala E. Et al. (2014) proposed a SVM and k-Nearest Neighbour (okay-NN) Tele-tracking of PD patients remotely by using taking their voice recording at regular c program languageperiod. The age, gender, voice recordings taken at baseline, after 3 months, and after six months are used as features are assessed. Support Vector Machine changed into greater a success in detecting large deterioration in UPDRS rating of the patients.

3. Methodology

This segment explains the stairs taken to acquire the prediction of Parkinson's disorder using numerous gadget studying. The diverse steps taken are Data accumulating , Data Preprocessing, Model Selection, Training, Evaluation, prediction .

3.1 Data Gathering
The first step is Data gathering .This step could be very crucial due to the fact the great and amount of the data you acquire will directly influences the extent of your prediction version. So we have taken facts of different voice recordings of the affected person.

3.2 Data education

In this step the statistics is visualized well to spot the relationship between the parameters present inside the data so one can take the gain of in addition to to get the data imbalances. With this ,we need to split the records into two parts .The first element for education the model like in our model we've got used 70 percentage of information for education and 30 percentage for trying out. Which is the second one part of the information

3.3 Model Selection

The subsequent step in our workflow is model choice. There are diverse fashions which have been used until date via researchers and scientist. Some are supposed for picture processing ,a few for sequences like text, numbers or patterns. In our case we've got 26 features which defines the voice recording of numerous patients so we've selected such models with a purpose to classify or differentiates the bad affected person with the wholesome one.

3.4 Training

Training the dataset is one of the predominant project of gadget mastering .We are able to practice the statistics to regularly enhance the selected model's potential to expect better ie the actual end result need to be approx. To expect one.

3.5 Evaluation

The metrics we have calculated are ROC, Accuracy , Specificity , Precision and many others. So one can highlights the nice set of rules amongst all.

3.6 Prediction

In this segment we finally get the model ready to discover the prediction of Parkinson's sickness based on the given dataset.

3.7 Dataset

The Parkinson sickness facts set is to be had at the uci system getting to know repository. The information set changed into initially created with the aid of the University of Oxford, check have been conducted on 31 human beings and out of which 23 people suffered from the Parkinson ailment. All the voice based totally take a look at had been conducted on these topics and the information is ready .The statistics set consists of 198 rows and 23 columns. The columns are sex the gender of the patients, age of sufferers, situation# the situation ids, name of the collaborating patients, MDVP:Fo(Hz) - mean vocal simple 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 different kinds measurements of variant in fundamental 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 additives within the voice, reputation includes o and 1 respectively for the patient with and without the Parkinson ailment, PDE, D2 - Two nonlinear dynamical complexity measures DFA Signal fractal scaling exponent, spread1, spread2, PPE - Three nonlinear measures of essential frequency variant.

3.8 Proposed model

The dataset is in addition dealt with for the missing values by way of the usage of the imply of values primarily based at the repute. The exploratory facts evaluation is performed at the dataset which incorporates Pearson's correlation approach, container plot of various features set, violin plot, the field plot and violin plot offers an insight of the facts distribution. The facts set is scaled the use of the same old scalers which scales the values by means of using the standard deviation of the values contained within the statistics set. Then the information is cut up in eighty:20 ratio for the cause of education and test set. The statistics set is used to teach the random woodland classifier. The random woodland classifier is a ensemble mastering primarily based boosting approach. The boosting approach used to develop the random woodland classifier is the bootstrap based boosting method, which divides the data into more than one component and every element is referred to as as the bag. Each bag is used to build a decision tree.

The check facts is anticipated by way of the usage of the all the decision trees and the mode of the prediction is the result of the random forest classifier or we have the averaging method. The random wooded area classifier is hyper tuned for the range of estimators, which is the quantity of selection tree to be constructed at the given information set. For hyper tuning of version we've the grid searchcv model to be able to match the model with the given options for the quantity of estimators and the value with excellent prediction accuracy, excellent fitting time, satisfactory prediction time and the quality schooling time is lower back because the nice fit parameter. The grid seek cv internally makes use of the move validation with default of five K-fold. The random woodland version is in comparison to the opposite algorithms like Kneighbors, Logistic regression version, Decision Tree Classifier.

All the alternative models are also hyper tuned for unique parameters like Kneighbors is tuned for the range of buddies parameters, logistic regression version is hyper tuned for the parameters like mastering rate (c), penalty for which we have 2 options on the whole l1 and l2, and the studying rate. The selection tree classifiers don't considers all the capabilities of the features set used for education the model it considers most effective the crucial features out of the characteristic set. The choice tree makes use of a segmented tree to save the model wherein each node incorporates the satisfactory cut up cost in conjunction with the operators for comparison like == and != if the characteristic is express and >= and < if the characteristic is continues. The decision tree is used for the most depth of tree that is the period of the tree. The information set is of kind multinomial. So the result is measured through the usage of the prediction accuracy of the fashions. The block diagram of the proposed version is as shown within the figure 1.

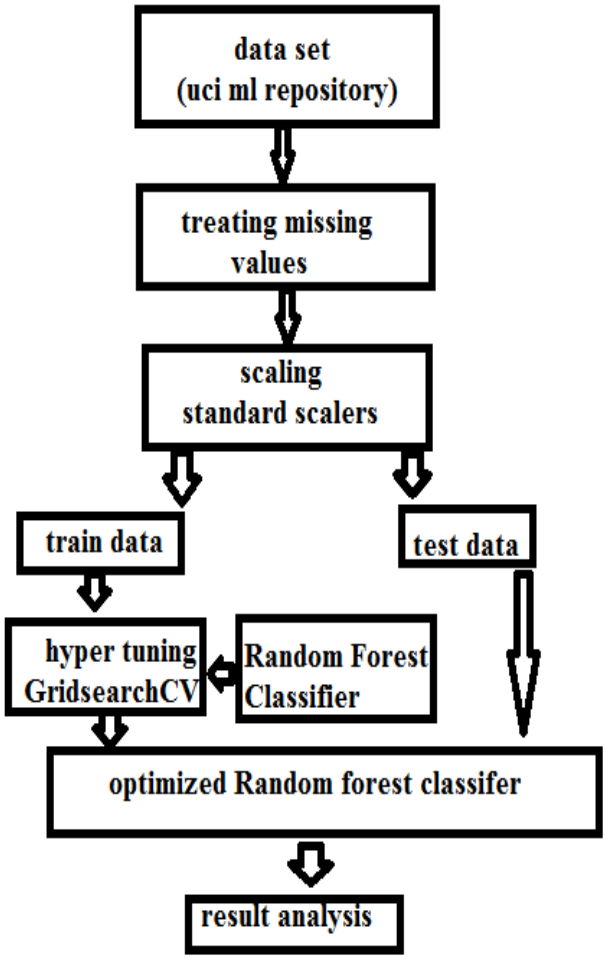

Fig. 1: The Proposed Prediction Model

4. Result and Discussion:

The Parkinson Disease information set is available at the uci device studying repository. The facts set consists of 23 columns. The Parkinson disease data set is multinomial classification hassle. For implementation of the model we've used the python language. It is a dynamically typed and a interpreted excessive stage programming language. It is a popular era now a days in subject of AI, gadget mastering, deep gaining knowledge of. Firstly we must perform the information pre-processing. For selecting the higher records pre-processing fashions we can use the exploratory information evaluation. The exploratory statistics analysis consists of the depend plot, field plot, and violin plot. With the help of box plot and violin plot we will discover the correlation most of the columns. For the visualization of correlation of complete records we've used the warmth map. The parent 2 suggests the full male and female patients facts contained in dataset.

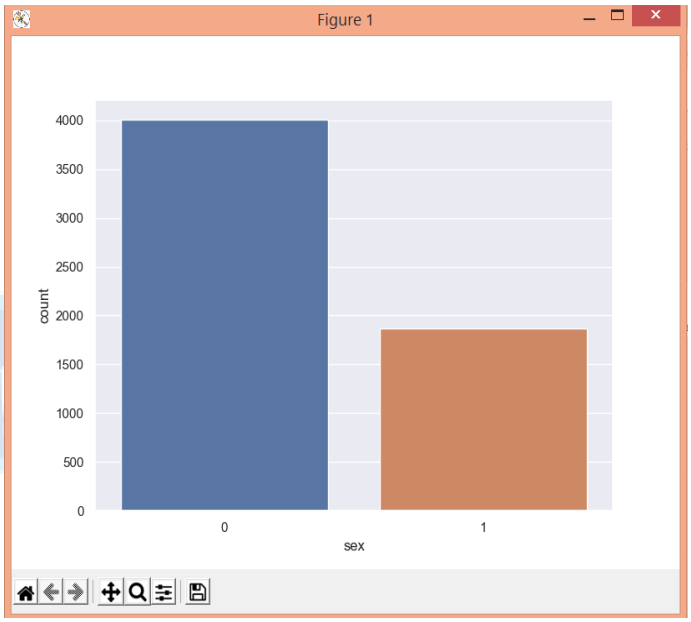

Fig 2: Total Number Of Male And Female Patients

The container plot is used to find the distribution of information upon a element. The field plot consists of min, max, 25 percentile,50 percentile (that's often termed as median), seventy five percentile, the aggregate of 25,50,75 percentile is often termed as IQR(interquartile range).Within the discern 3 we have used container plot to test the information distribution of intercourse column versus subject# and sex as opposed to the age. Where intercourse is gender of patients and difficulty # represents both the managed patients or the Parkinson disorder patients.

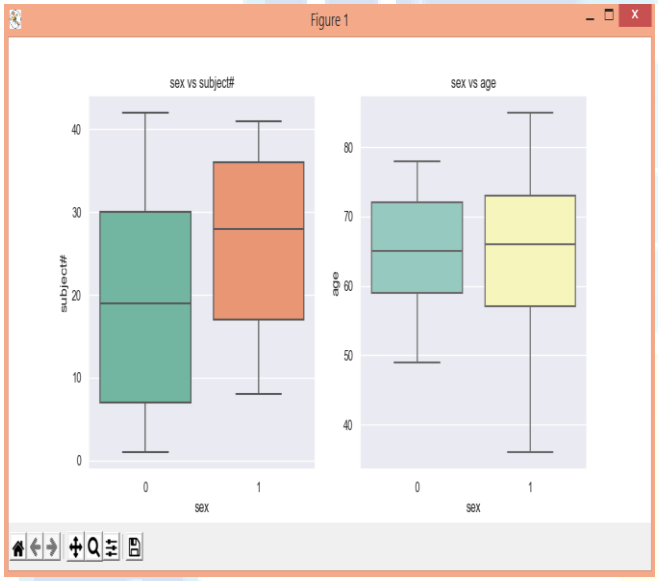

Fig 3: The Box Plot Of Sex Versus Subject# And Age

The violin plot is much like the field plot it moreover includes the possibility density of the statistics it's far an awful lot informative while compared to the container plot. The most effective difference among the box plot and the violin plot is if the statistics contains multiple peaks (multimodal) then the box plot includes simplest one top whereas the violin plot includes the more than one peaks. The figure four suggests the violin plot of reputation versus the RPDE and DFA. The discern five suggests the container plot of repute as opposed to the NHR and HNR.

The different version used for the prediction device is the ensemble getting to know primarily based bagging technique classifier this is Random woodland classifier. The random woodland classifier is hyper tuned for the parameters of the quantity of estimators. The wide variety of estimators represents the range of choice tree fashions for use for the purpose of building the predation model. The random forest classifier is uses all the base estimators to are expecting the given pattern and the mode of the result is lower back as an output. The random woodland classifier hyper tuning resulted the nice accuracy for the wide variety of estimators to be 350. The result of prediction accuracy shows that the tuned model is higher than the non- tuned model.

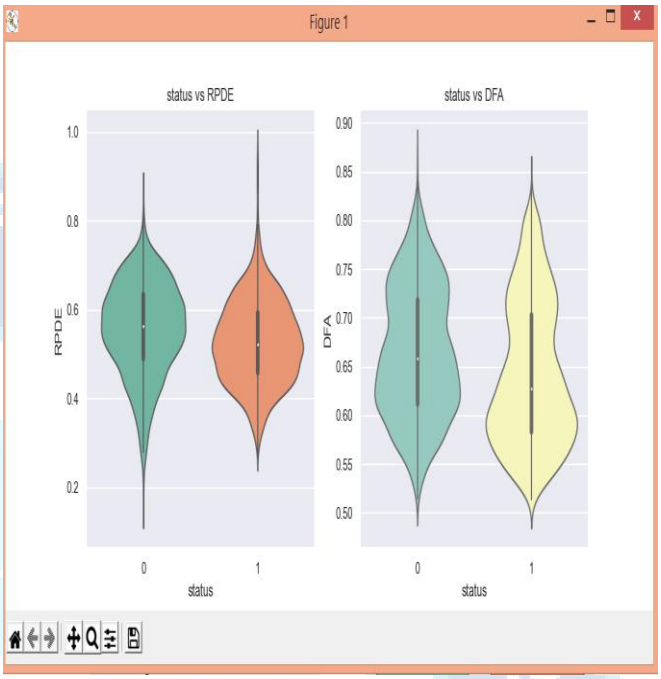

Fig 4: The Violin Plot Of Status Versus RPDE and DFA.

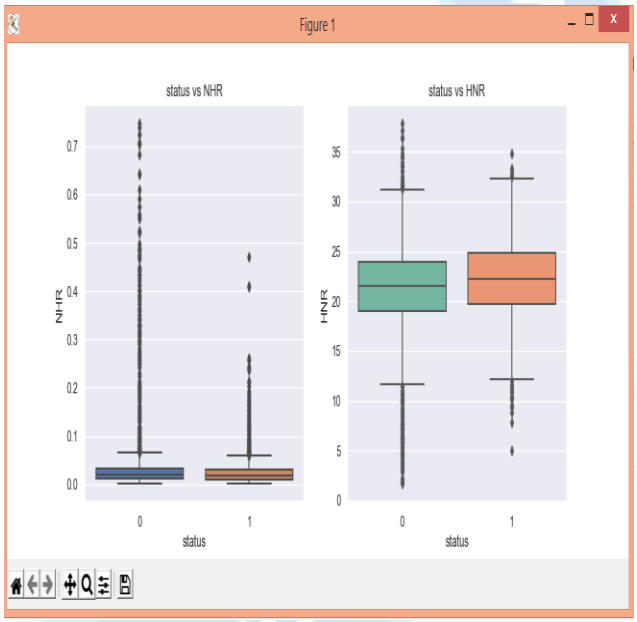

Fig 5: the box plot of status versus NHR and HNR.

The consequences are proven in parent 6. The prediction accuracy of the tuned Kneighbor model is ninety.Forty one, tuned logistic regression model is 98.98, tuned choice tree classifier is 98t.75 and the tuned random forest classifier is 99.54. The results are proven inside the figure 6. For the result evaluation the move validation with k-fold of 5 is used. The prediction model for the Parkinson sickness is built using the tuned random woodland classifer.

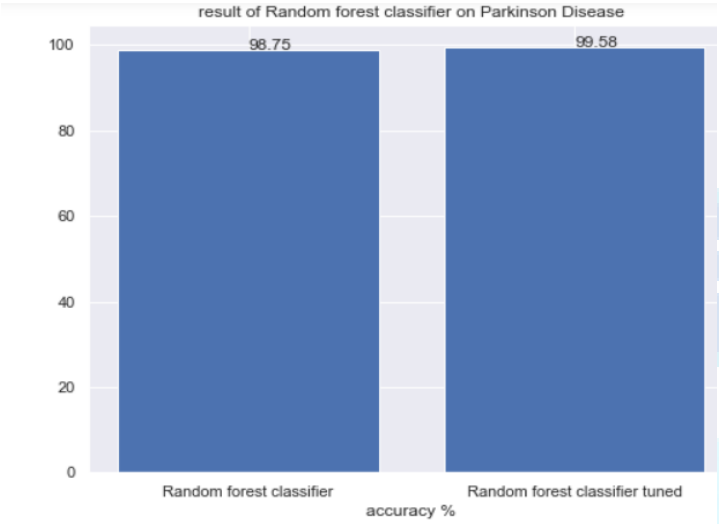

Fig 6: The Prediction Accuracy Of Random Forest Classifier.

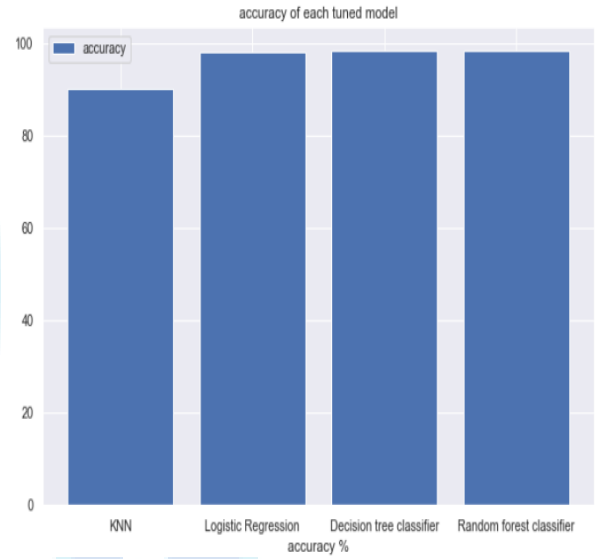

Fig. 7: The Prediction Accuracy Of All The Models.

5. Conclusion:

It is a great deal vital to detected the Parkinson ailment at the initial degree itself for which we have advanced the gadget getting to know based prediction gadget of the Parkinson disorder. The dataset is to be had on the uci device learning repository that's used to construct the proposed model. Various device mastering fashions were used at the dataset. Before using the system getting to know at the dataset we've got accomplished the exploratory statistics analysis and finished the pre-processing which includes scaling and characteristic extraction. The models have been hyper tuned using the gridsearchcv module. All the fashions have been skilled the use of the Parkinson disease and the consequences suggests that random forest is the satisfactory version for the prediction. And the accuracy of the models is 99.54 percent

The dataset is specially collected from human beings of England. The model may be advanced through the use of diverse distinctive form of dataset from distinct part of the sector try this the version might be flexible.

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.