International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

Estimation & Evaluation on Parkinson's Disease by Implementing ML Algorithms

Swetha K B, R Kruthika, Swathi R, Ranjitha V

Department of Information Science and Engineering, RR Institute of technology, Bengaluru, Affiliated to VTU, Belgaum, Karnataka, India.

ABSTRACT: Parkinson's disease (PD) is one of the prime health problems in the world. It is recognized that around one million people suffer from Parkinson's disease. Thus, it is necessary to identify Parkinson's disease in an earlystage so that early plan for the necessary treatment can be made, to save lives and advice few life style changes for a peaceful and healthier life at low costs. The proposed predictive analytics framework is a cluster of Decision Tree, Support Vector Machine, Naive Bayes Algorithm, Gradient Tress Boosting Algorithm, and Stochastic Gradient Descent which is used to acquire insights from patients. We obtain voice dataset from UCI Machine learning repository as input. The experimental outcome indicates that an early detection of PD will expedite clinical monitoring of elderly folk and increase the chances of their life span and improves lifestyle to lead peaceful life.

KEYWORDS: Parkinson's disease (PD), Decision trees, Naive Bayes Algorithm, Support vector Machine(SVM), Gradient Tree Boosting Algorithm, Stochastic Gradient Descent.

I.INTRODUCTION

PD is a progressive neurodegenerative disorder. It affects certain brain cells that support in controlling the movement and coordination. Dopamine is a hormone and neurotransmitter, a chemical that is generated by brain cell. It is used to transmit signals to other brain cells to control the muscle activity [1]. PD causes, degeneration of dopamine in the brain cell which leads to abnormal muscle activity. It is a common disorder observed in senile person which occurs in 1% of the population.[1] There are several symptoms that cause PD. Common symptoms in PD are muscular rigidity (limbs and upper half of the body is inflexible), shivering (vibration in upper and lower limbs or jaws, speech problem), expressionless face, Bradykinesia (slow movements), lethargy (unresponsiveness and inactivity), postural instability (depression and emotional changes), involuntary movements, dementia (loss of memory which is a common disorder of Alzheimer's disease), thinking inability and sleeping disorders.
Certain phases in Parkinson's disease are:

- Primary - Due to unknown reasons
- Secondary - Dopamine deficiency
- Hereditary- Genetic origin
- Multiple system atrophy - Degeneration of partsother than mid brain.

[2]For later stages, surgery is recommended for some people. It does not cure PD, but it may help to ease symptoms. Surgery, Deep Brain Stimulation (DBS) is offered to people with advanced PD.[4] Electrodes can be embedded into specific part of the brain that sends signals to your brain and may reduce the PD symptoms.DBS is a stabilized medication which reduces involuntary movements, tremor and rigidity.
Approximately 15% of people with PD have a family history of the disorder. In a few cases, the disease may be inherited through certain gene changes. PD may occur at the age of 60. [3]Due to technological development in information technology, and healthcare areas resulted in better outcomes and low-cost healthcare delivery that is possibly predicted from the PD patient's analysis.

II.LITERATURE SURVEY

The accurate diagnosis of PD has been a challenge to date, mainly due to the close relevance of PD to other neurological diseases. These close characteristics are the reasons that cause 25% inaccurate manual diagnosis of PD. In [22], they presented a Convolutional Neural Network (CNN) based automatic diagnosis system which accurately classifies PD and healthy control (HC). Parkinson's Progression Markers Initiative (PPMI) provides publicly available

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

benchmark T2- weighted Magnetic Resonance Imaging (MRI) for both PD and HC. The mid-brain slices of 500, T2-weighted MRI are selected and aligned using image registration technique.

[24] Provides approaches to quantify the motor symptoms of Parkinson's disease with wearable devices, and the quantification was developed with four paradigm maneuvers. The quantification parameters included tremor frequency, amplitude of hands movement, and speed of foot-tapping and turning duration. The approach provided quantitative measures for the tremor, bradykinesia and gait symptoms, which could be useful for optimization of drug or deep brain stimulation treatments.

[25] A study of the frontal and temporal EEG of Parkinson's disease patients using MATLAB platform. Lyapunov exponent and inverse Lyapunov exponent for both PD and healthy subjects were calculated within a given time frame. It was found that for PD subjects, the Lyapunov exponent for the temporal part of the brain is less than that of the frontal while the inverse Lyapunov exponent is reverse order of the brain. Many researchers have conducted several studies using voice recordings to produce an accurate PD diagnosis system. One unique promising way to use the speech disorder as a helping factor to predict PD is by using machine learning techniques [26] they used NNge classification algorithms to analyze voice recordings for PD classification. NNge classification is known to be an efficient algorithm for analyzing voice signals but has not been explored in details in this area. Then, an experiment using NNge classification algorithm to classify people into healthy people and PD patients was performed. The parameters of the NNge algorithm were optimized. Finally, NNge and ensemble algorithms specifically, AdaBoostM1 was implemented on the balanced data. The final implementation of NNge using AdaBoost ensemble classifier had an accuracy of 96.30%.

As the PD progresses slowly in most people. Therefore, it is difficult to be identified in the earlier stage. It resides for many years with only minor symptoms. The symptoms differ in various stages of the disease, but they mainly involve tremors, rigidity, bradykinesia, flat facial expression, and speech disorder. Since speech disorder is one of main PD symptom, recording voice signals and analyzing it automatically is the easiest and most reasonable way to identify the disease in its initial stages [28].

An attempt was been made to `distinguish PD group from the healthy control group based on voice recordings with selected features and different classification techniques such as linear classifiers, nonlinear classifiers and Probabilistic classifiers. They used recursive feature elimination algorithm (RFE) for selection of important features [29].

III.EXISTING APPROACH

In this section, we review some existing machine learning techniques for diagnoses of Parkinson Disease. Olanrewaju et al. in the article [30], proposed machine learning based technique for diagnosis of PD and developed Multilayer Feed Forward Neural Network (MLFNN). They used a data set, available in Oxford Parkinson disease datasets. The dataset consists of voice measurements of 31 people with 23 patients. They used 8 attributes which are based on frequency (tremor). They used a total of 8 input and 10 hidden nodes. For classification, they used the k-mean algorithm. The simulation result showed that they achieved a sensitivity of 83.3%, specificity of 63.6% and accuracy up to 80%. However, this method is not validated on real data.

Das et al has done a comparison based on different classification method on speech signals for effective diagnosis of PD. He used four classification methods such as neural networks, regression, for effective diagnosis of PD. He used four different classification methods such as Neural networks, Regression, DM neural, and Decision tree and found that neural network is the best among the four classifierwith accuracy of 92.9%.

Prashanth et al. in the article [31], worked on nonmotor features for PD diagnosis. The non-motor features consist of Rapid Eye Movement (REM), sleep behavior disorder, and olfactory loss. In this study, they used nonmotor features in combination with cerebrospinal fluid measurement and dopaminergic imaging markers features. They used Naïve Bayes, SVM, Boosted Tree and Random Forest for classification. The simulation result showed 96.4% accuracy rate of SVM. It was found that the combination of different non-motor features yielded better results. However, data contained imbalanced class distribution. Al-fatlawi et al. in the article [32], proposed Deep Belief Network (DBN) for diagnosis of PD. In this work, a data set PDD was obtained from the UCI data repository that contains 195 voice recordings of 31 people and 16 attributes. The proposed DBN yielded 94% accuracy. However, this work didn't calculate harming probability.

IV.PROPOSED APPROACH

In this proposed system, we follow a similar approach, however we try to use different machine learning algorithms that can help in improving the performance of model and also play a vital role in making in early prediction of PD which in turn will help us to initiate neuroprotective therapies at the right time.

The clinical Parkinson dataset consists of 195 instances and 22 attributes with one class without any missing values. The data set is collected from speech sounds produced during standard speech tests records using a microphone and

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

there recorded speech signals are analyzed using part software [10] to eliminate noise and characterize unique properties in signals. The classifier models are trained with 66 percent of the sample and tested them over the rest of it. Sufficient care has been taken such that the testing is not done over the same instances. Also the data set is standardized such that the overall standard deviation & mean is equal to 1 and 0 respectively using relation given by P(j)= (k-n)/SD where P(j) is the standardized data, k is the data to be standardized, n is the mean of the population and SD is the standard deviation. Data cleaning methods are applied for dimension reduction horizontally as well as vertically.

The flow chart of the proposed methodologies is shown in the figure 1. 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 Random Forest-Recursive Feature Elimination (RF-RFE) algorithm on the original feature sets. Feature reduction removes multicollinearity resulting in improvement of the model in use. Feature reduction is used to decrease the number of dimensions, making the data less sparse. We have used nonlinear classifier with decision tree for classification of groups are as follows Bagging classification and Regression tree, Random Forest and Boosted C5.0 and probabilistic classifier as Naïve Bayes method, and linear classifier as SVM Support vector machine is used for voiceprint analysis of Parkinson's disease patients.

For time series acquired from subjects specified, we applied the linear and nonlinear dynamics analysis, using specific parameters for the tremor symptom diagnosis. To characterize the tremor signal we used the following parameters: frequency, amplitude, type of tremor, spectral character.

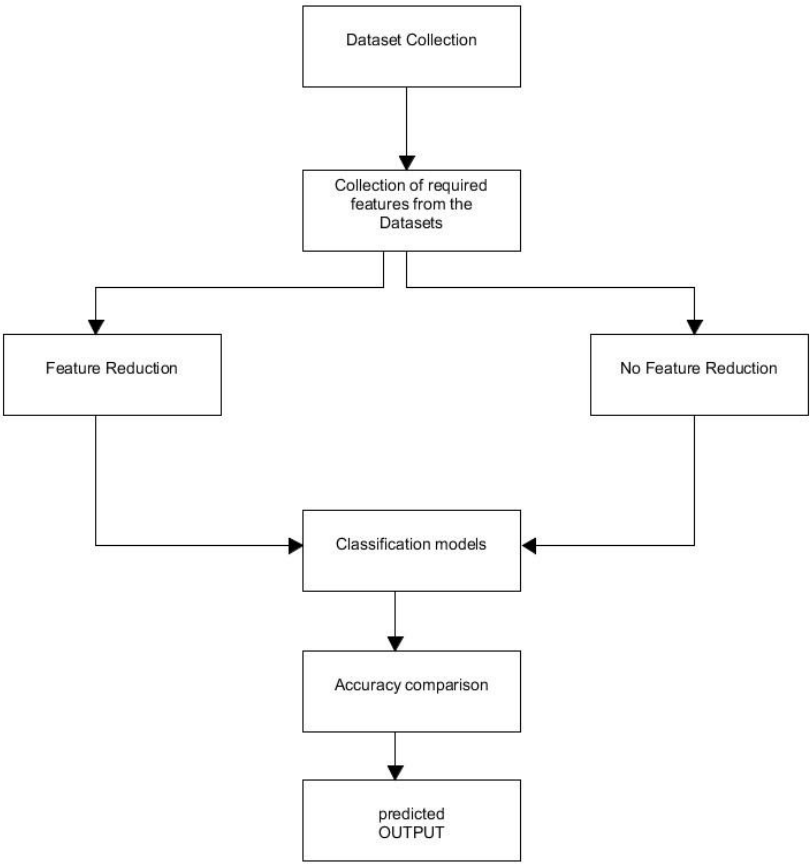

Fig 1: Flow chart of proposed method

V. EXPERIMENTAL CONFIGURATION

The architecture description is a formal description and also a representation of a system, organized in a way that supports reasoning about the structures and behavior of the system.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

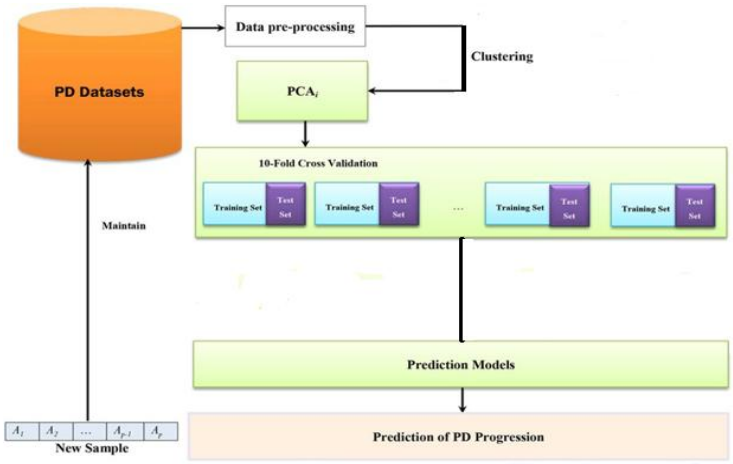

Fig 2: Architectural diagram

Dataset:

The dataset used for method evaluation was obtained from the UCI Machine Learning Repository. The dataset was designed by Max Little of the University of Oxford, in association with the National Centre for Voice and Speech, Denver, Colorado, who taped down the audio signals. This dataset contains data from voice recordings of 23 subjects with PD and 8 control subjects. There are a total of 195 recordings, from which 22 various discrete voice measure features have been obtained. Each example also includes a subject identifier and a binary classification attribute which indicates whether or not the subject has PD. The dataset is classified into two classes as per its "Status" column which is set to 0 for without PD subjects and 1 for those with PD. The features of this database are: MDVP: Fo(in Hz; Average vocal fundamental frequency), MDVP: Fhi(in Hz; Maximum vocal fundamental frequency), MDVP: Flo(in Hz; Minimum vocal fundamental frequency), MDVP: Jitter(in %), MDVP: Jitter(Abs), MDVP:RAP, MDVP: PPQ, Jitter: DDP(Numerous measures of variation in fundamental frequency), MDVP:APQ, MDVP: Shimmer, RPDE, MDVP: Shimmer(dB), Shimmer: APQ3, Shimmer:APQ5, Shimmer: DDA(Several measures of variation in amplitude), NHR, D2 (Two nonlinear dynamical complexity measures), DFA(Signal fractal scaling exponent), Spread1, Spread2, HNR(Two measures of ratio of noise to tonal components in the voice), PPE(Three nonlinear measures of fundamental frequency variation), Status(Health status of the subject (1) with PD, (0) without PD).

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a tool for data compression and information extraction. For an algorithm to be effective in this domain, it needs to be able to handle noisy data, highly correlated and redundant variables which rely on relatively few medical tests, and complement the role of physicians. They must be compressed in a manner to retain the essential information and are easy to display. Among the widely used multivariate statistical methods. Using PCA, a number of related variables are transformed to a set of uncorrelated variables. It is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. Its general objectives are data reduction and interpretation.

Cross-validation

Cross-validation is a statistical method is used for the performance evaluation of learning algorithms and performance of a predictive model on an unknown dataset in this research. The datasets used in this research are divided into several equally sized subsets. The learning model is then trained on some subsets known as training sets. After training process, the model is tested on the remaining subsets, known as test sets. According to the number of subsetspartitioned, the process is tested under *k*-fold cross-validation. In *k*-fold cross-validation, we use *k* result of *k*-fold cross-validation. K-fold cross-validation, involves partitioning the original sample into randomly partitioned *k* subsamples.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

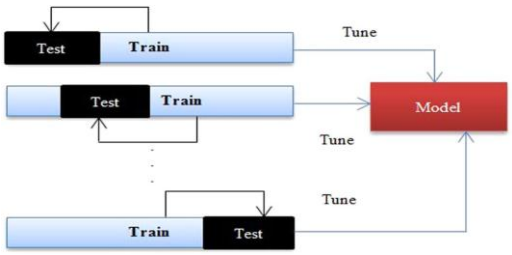

Fig 3: *k*-fold cross-validation

Decision tree

The basic algorithm used in decision trees is known as the ID3 algorithm. The ID3 algorithm creates decision trees using a top-down, greedy approach. To define information gain accurately, we specify a measure called Entropy that measures the level of impurity in a group of examples. Mathematically, it is defined as:

Entropy: $\sum i = 1 - p * \log_2(pi)$

Pi=Probability of classification.

Support Vector Machine

The support vector machine is applied as a classifier and tested with 10-fold cross-validation. This novel aspect provides a perfect PD classification with 100% accuracy, sensitivity and specificity. The SVM can be used to predict which category a new input vector belongs to. Two-dimensional and three-dimensional SVMs were developed and evaluated.

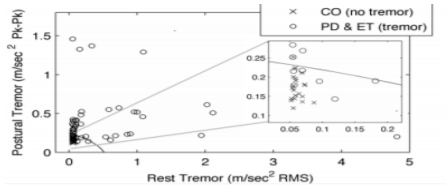

Fig 4: Classified graph based on PD *vs* non PD

The first SVM classified patients as either with PD or without PD. The second SVM classified between PD and non PD. For the two-dimensional SVMs, the 10-fold cross validation estimate of the misclassification rate was 9.5% for the PD versus non PD SVM.

Naive Bayes

Naive Bayes Classifiers depends on the Bayes Theorem i.e. it's determined from conditional probability, inother words, the tendency that an event (A) will happen given that another event (B) has already happened. Elementally, the theorem allows a hypothesis to be updated each time new information is introduced.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Stochastic Gradient Descent

It is a process that is associated with a random probability. Hence, in Stochastic Gradient Descent, a number of instances are selected randomly instead of the whole data set for every iteration.

$$for\ i\ in\ range\ (m):$$
$$\theta_j = \theta_j - \alpha\,(\hat{y}^i - y^i)\,x_j^i$$

Gradient Tree Boosting Algorithm

Gradient tree boosting is generally used with decision trees with fixed size as base learners. For a special case, Friedman proposes a modification to gradient boosting method which improves the quality of fit of each base learner.Generic gradient boosting at the *m*-th step would fit a decision tree $h_m(x)$ to pseudo-residuals. Let J_m be the number of its leaves. The tree partitions the input space intoJ_m disjoint regions $R_{1m},\ldots,R_{Jm}$m and predicts a constant value in each region. Using the indicator notation, the output of$h_m(x)$ for input x can be written as the sum:

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

$$h_m(x) = \sum_{j=1}^{J_m} b_{jm} \mathbf{1}_{R_{jm}}(x),$$

Where: b_{jm} is the value predicted in the region R_{jm}.

VI. RESULT AND ANALYSIS

We have used Python programming language to write the code.We trained each classifier based on the trained data andpredict the effectiveness of the classifier on the test data. So eachclassifier is able to show all the performance metrics based onthe test data.

All parameter changes used 10-fold cross validation for thepartitioning of training and test datasets. The highest accuracy of 87.74% wasachieved when the number of attempts of generalization wasset to 2 and 4. After that, the accuracies vary slightly andfinally, they remain the same (87.74%).

A. Comparison of Accuracy

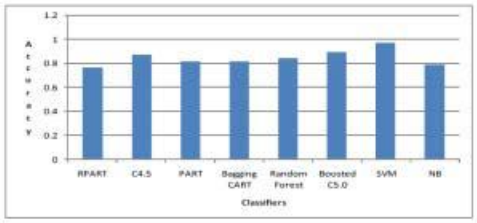

Fig 5: Shows that SVM performs better in terms of accuracy among all the classifiers. It shows the maximum accuracy of 87.74%.

B. Comparison of Sensitivity

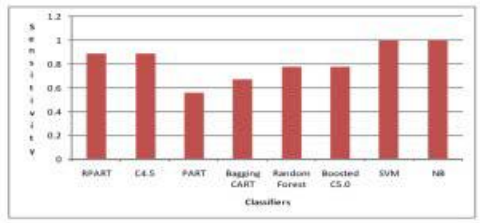

Fig 6: Shows that SVM and NB have highestsensitivity among other classifiers. Both of theclassifiers shows maximum sensitivity of 1.

C. Comparison of Specificity

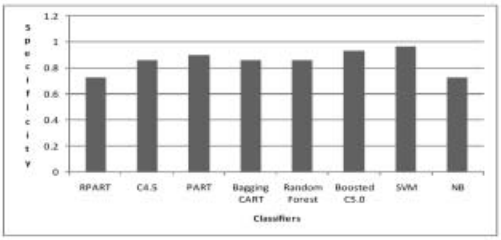

Fig 7: Shows that SVM has highest specificityamong other classifiers. It shows maximumspecificity of 0.8460.

D. Number of folders for computing the mutual information values andtheir accuracies

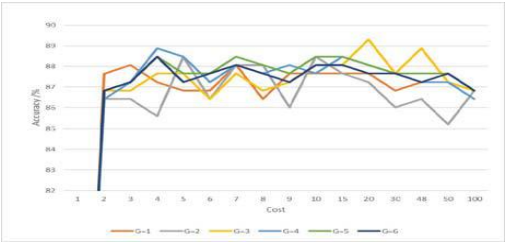

Fig 8: Shows theaccuracies of the prediction after applying the changes innumber of folders for computing the mutual informationparameter

To further consider the outcome of the implementation, theresults of these techniques used in this experiment arecompared with the results of the techniques used in thepapers presented in the literature. In a study reported in [26],an accuracy of 90% was reached using NNge. However, theydid not specify the details of their experiment. They

performed featureselection however the details of their feature selection andparameter selection method is not specified. Our studyimplemented an ensemble of voice datatsetand also specifying feature selection by achieving an accuracyof 87.74%.

VII. CONCLUSION

In this paper we have implemented different algorithms such as Decision Tree, Support Vector Machine, Naive Bayes Algorithm, Gradient Tress Boosting Algorithm, and Stochastic Gradient Descent to classify the PD. We found SVM has performed well in terms of accuracy, sensitivity, and specificity. It was presented how to use different algorithms to generate accurate assessments of Parkinson's disease at various levels. Using SVM one can attain accuracy up to 87.74% and it's been an efficient method to find the features vector for an individual patient at a given time and to predict and identify a stage in Parkinson's disease. It can be concluded that these neurophysiological measurements are most helpful for early diagnosis of PD.

REFERENCES

[1] Journal of Systems Science, 43(4), pp.597-609, 2012. 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. 310, 2003.
[2] Y.Xu, X.Wei, X.Liu, J.Liao, J.Lin, C.Zhu ...andM.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.Tjaden, "Speech and swallowing in Parkinson's disease. Topics in geriatric rehabilitation", 24(2), pp. 115126, 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] S. Shimon, L.O. Ramig, C.M. Fox, "Intensive voice treatment in Parkinson's disease: Lee Silverman voice treatment", Expert Review of Neurotherapeutics, 11(6), pp. 815- 830, 2011.
[6] A.Tsanas, M. A.Little, P. E.McSharry, and L. O.Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests", IEEE transactions on Biomedical Engineering, 57(4), pp. 884893, 2010.
[7] M.Ene, "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, pp. 112-116, 2008.
[8] 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.
[9] W.Froelich, K.Wrobel, and P. Porwik, "Diagnosis of Parkinson's disease using speech samples and thresholdbased classification", Journal of Medical Imaging and Health Informatics, 5(6), pp.13581363, 2015. [10] 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.
[11] R.Das, "A comparison of multiple classification methods for diagnosis of Parkinson disease", Expert Systems with Applications, 37(2), pp. 1568-1572, 2010.
[12] I. Bhattacharya, and M. P. S. Bhatia, "SVM classification to distinguish Parkinson disease patients", Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India. ACM, 2010.
[13] K.Polat, "Classification of Parkinson's disease using feature weighting method on the basis of fuzzy C-means clustering".
[14] Ö. Eskidere, F. Ertaş, and C. Hanilçi, "A comparison of regression methods for remote tracking of Parkinson's disease progression", Expert Systems with Applications, 39(5), pp.5523-5528, 2012.
[15] D.C.Li, C. W.Liu, and S. C.Hu, "A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets", Artificial Intelligence in Medicine, 52(1), 45-52, 2011. [16] F.S.Gharehchopogh, and P.Mohammadi, "A Case Study of Parkinson's Disease Diagnosis using Artificial Neural Networks", International Journal of Computer Applications, 73(19), pp.1-6, 2013.
[17] W.Froelich, K. Wrobel, and P. Porwik, "Diagnosis of Parkinson's disease using speech samples and thresholdbased classification", Journal of Medical Imaging and Health Informatics, 5(6), pp. 13581363, 2015. [18] M.Shahbakhi, D. T.Far, & E. Tahami, "Speech analysis for diagnosis of Parkinson's disease using genetic algorithm and support vector machine", Journal of Biomedical Science and Engineering, 7(4), pp.147156,2014. [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] P.M.Granitto, C.Furlanello, F.Biasioli, and F.Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agro industrial products", Chemometrics and Intelligent Laboratory Systems, 83(2), pp.83-90, 2006.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

[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] Detection of Parkinson Disease in Brain MRI using Convolutional Neural Network. Proceedings of the 24th International Conference on Automation &Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2018.
[23] Deep Brain Stimulation Efficiency and Parkinson's disease Stage Prediction using Markov Models. Grigore T. Popa University of Medicine and Pharmacy, Iaúi, Romania, November 19-21, 2015.
[24] Quantification of the Motor Symptoms of Parkinson's EMBS Conference on Neural Engineering Shanghai, China, May 25 - 28, 2017.
[25] Significance of Lyapunov Exponent in Parkinson`sdisease using Electroencephalography. 978-1-7281-13807/19/$31.00 ©2019 IEEE.
[26] Classification of Parkinson's Disease Using NNge Classification Algorithm 2018.
[27] R. G. Ramani, "Parkinson Disease Classification using Data Mining Algorithms," vol. 32, no. 9, pp. 17–22, 2011.
[28] Parkinson's foundation:Better lives together. [29] A Mixed Classification Approach for the Prediction Parkinson's disease using Nonlinear Feature Selection Technique based on the Voice recordings. Proceedings of International conference (ICIC) 2017.
[30] R. F. Olanrewaju, N. S. Sahari, A. A. Musa, and N. Hakiem, Application of neural networks in early detection and diagnosis of Parkinsons disease, 2014 Int. Conf. Cyber IT Serv. Manag. pp. 7882, 2014.
[31] R. Prashanth, S. Dutta Roy, P. K. Mandal, and S.Ghosh, High-Accuracy Detection of Early Parkinsons Disease through Multimodal Features and Machine Learning, Int. J. Med. Inform., vol. 90, pp. 1321, 2016.
[32] A. H. Al-fatlawi and M. H. Jabardi, Efficient Diagnosis System for Parkinson s Disease Using Deep Belief Network, pp. 13241330, 2016.