2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019

Optimization of Features for Classification of Parkinson's Disease from Vocal Dysphonia

Walia Farzana,[1]* Dr.Quazi Delwar Hossain [2]
Department of Electrical & Electronic Engineering
Chittagong University of Engineering & Technology(CUET)
Chittagong-4349,Bangladesh
*walia.farzanalia@gmail.com

***Abstract*—Parkinsons disease is considered most prominent neurological disease after Alzheimer and Epilepsy. There is no defined test for early diagnosis of Parkinson's patient and medical decisions are provided based on the medical history of the patient and hence the possibility of misdiagnosis. Parkinsons disease influxes different prospects of a patient and in 90% cases vocal dysphonia is present an analysis of the vocal dysphonia can be considered as the early biomarker of decision making for medical practitioners and neurologists as well as biometric analysis. This study aims at vocal dysphonia analysis of Parkinson's patient from voice dataset with different machine learning algorithms with a goal to achieve better performance with less number of attributes. A comparative study is performed where k-Nearest performed approximately with 98% accuracy with 5 relevant attributes, Random Tree with 100% accuracy with 1 related attribute. In addition in the case of Multi-Layer Perceptions with different hiddenlayers, the performance is evaluated. It is observed that MDVP:Fo,MDVP:Shimmer, RPDE, Spread1 attributes contribute more to efficient classification accuracy.**

***Index Terms*—Classification, Attribute, Accuracy, Precision**

I. INTRODUCTION

Human brain which is the master controller of the whole human body has always been a paradoxical mystery as well as scientific wonder. The conventional acquainted neurological disorders range from migraines to Stroke, Dementia , Alzheimer, Epilepsy and Parkinsons Disease . More than 10 million people around the world are suffering from Parkinsons disease . In Bangladesh, every year, around 1600 patient die from this disease while many more are suffering from it. According to the latest World Health Organization data published in 2017 about Parkinsons Disease death in Bangladesh reached 539 or 0.07% of the total deaths. The death rate is 0.55 percent per 100,000 of population raked Bangladesh 152 position in the world in the area of Parkinsons disease death.

According to the report of Bangladesh Bureau of Statistics the percentage of elderly people age between 60-64, 65-69, 70-75 and over 70 were 37%, 21%, 20%,22% respectively[1]. Moreover,an increase in aging population means an the expectation of a rise in Parkinsons disease as the study suggested after the age of 60[2]. According to[3], Head of Neurology Unit at University Kebangsaan Malaysia Medical Centre (PPUKM), PD symptoms slowly develops in patients mostly over age 60, where prevails a misconception that age 60 or after 60 is the common age of prevalence of PD but it can also affect patient below 45. In addition, the disease results in increase of social isolation as well as financial burden of PD is estimated to rise in future[4].

Parkinsons Disease is a gradually progressive neurological disorder in which symptoms continue to worsen over time. The cause of Parkinsons Disease is still unknown to researchers. However, some risk factors such as exposure to pesticides and industrial fumes, drinking water from deep wells ,and head trauma is considered.There is no standardized test for detection of Parkinsons disease that means test like blood test or ECG can not ordain whether a person is suffering from Parkinsons or not. Its diagnosis is performed on the basis of patients medical history followed by some neurological examination. The misdiagnosis of Parkinsons is indicative as there is no absolute test. With a rapid shift in the arena of healthcare, decision supporting system can perform a significant preface . As mentioned before that detection, classification, and legislation of Parkinsons patient is challenging .In that case, the automatic classification of Parkinsons patient from the normal patient using machine learning can be the feasible solution over a traditional one. Nowadays, the classification method is applied in medical research where a large volume of data is needed to analyze. Classification system assists in enhance of accuracy and reliability of diagnosis and lessen diagnosis error as well as the time of diagnosis[5].Clinical decision making by doctors or physicians requires available information for the guidance. In that case, it would be helpful to provide physicians a machine learning model where models are developed using various machine learning algorithms such as SVM(Support VectorMachine), Logistic Regression, Naive Byes, Multi-Layer based on the important features or attributes. Based on the present prospects and future development of Parkinsons Patient Diagnosis developed methods are required which can fulfill decision making purpose. In this context different classification methods with low feature dimensionality can serve the defined purpose.

The paper is organized in the following manner.Section 2 contains related work.Section 3 Methodology followed by Results.And finally the conclusion of the work.

II. RELATED RESEARCH WORK

Different Researchers have applied various classification methods on datasets acquired from Parkinsons Patients to differentiate PWP (Patient With Parkinsons) from Normal one. Voice Analysis seems to be a biomarker for early diagnosis of Parkinsons Disease and analysis a large dataset using machine learning algorithms proves to be a pioneer of the beneficiary for clinical research. Zahari Abu Bakar and Nooritawati [5] have

978-1-5386-9111-3/19/$31.00 ©2019 IEEE

used namely Levenberg-Marquardt (LM) and Scaled Conjugate Gradient (SCG) of Multilayer Perceptrons (MLPs) Neural Network in diagnosing PD. The dataset they have used consists of 195 voice samples and they found out that the LM algorithm depicts 97.86% accuracy with 25 hidden layers while SCG algorithm attain an accuracy of 79.06% with 10 hidden units. Ramani[6] used a dataset created by the Max Little of the University of Oxford ,in collaboration with the National Centre for Voice and Speech ,Denver, Colorado. The aim of his work was feature relevance analysis as well as finding out the best classification rule with minimum error rate.The author of the paper[7] proposed Minimum Redundancy Maximum Relevance feature selection algorithms to sort out the most important features related to each classification methods. Multiple numbers of features such as 5 features, 8 features, and 20 features were selected as well as multiple classification algorithms such as Support Vector The machine, Random Tree, Decision Tree, Bagging ,Boosting were applied.According to the author [8], the method for Classification used is SVM with different kernel function and observation of different performance parameter with an increase in cross-validation. Tarigoppula V.S. Sriram and M. Venkateswara Rao[8] use classification methods along with Clustering Method. Some of the classification methods are Nave Bayes, Simple Logistic, Random Forest and K-star. According to a case study by Farhad and Peyman[9] Multi-layer Perception with back-propagation algorithm and Radial Basis Function (RBF) Automatic Neural Network is used for classification purpose. With MLP the network consists of three layers: the first one input layer with 22 attributes, the the second one is the hidden layer and one output node. Resul Das[10] proposed four classification methods (Neural Network, DMneural,Regression,Decision Tree) and the comparative study of the classification methods on the Oxfords Parkinsons Disease Database(OPDD). He has used 65% data as training set and the remaining as a test set with SAS base software 9.1.3 which includes SAS Enterprise Guide Program 4.3 for data pre-processing and SAS Enterprise Miner for comparison purpose among classifier methods. For performance evaluation Receiver Operating Curve (ROC) and SAS software based Cumulative curve is used. A.Bourouhou and co-authors[11] have applied three distinct classifiers: Support Vector Machine, k-Nearest Neighbors, Nave Bayes, on the same database to find out the efficient classifiers. and support vector machine is an efficient one. Vasily Sachnev and Hyoung Joong Kim[12] conducted a research on ParkDB database consists of 22283 gene expressions information from 22 normal patients and 50 Parkinsons Patent. They have proposed a Binary Coded Genetic Algorithm for feature selection and Extreme Machine Learning for Classification purpose. The author [13] proposed a new algorithm method for classification of Parkinsons Disease where he had used a genetic algorithm for multiple numbers of feature selection and Support Vector Machine for classification using MATLAB software with 75% as training and 25% as a testing dataset. The classification accuracy is 94.50 percent with 4 features, 93.66 percent with 7 features and 94.22 percent with 9 features. Pei-Fang Guo[14] applied a combination of genetic programming and expectation maximization algorithm to discriminate healthy Patient from Parkinsons patient. The proposed method acquire an accuracy of 93.12% and they have mentioned the advantage of the algorithm as its ability to reduce feature dimensionality.

III. METHODOLOGY

A. Dataset

The database used for this study composed of 195 sustained vowel phonations among them 23 patients were diagnosed with Parkinsons disease. Around six phonations from per patient were recorded and the time span of recording ranges from 1 to 36 second. The age range of the subjects varies from 46 to 85 years (mean 65.8 ,standard deviation of 9.8)[15]. The dataset was created by Max Little of the University of Oxford in collaboration with National Centre for Voice and Speech , Denver, Colorado. Each column in the dataset refers to a particular attribute or feature and each row corresponds to one the 195 voice recordings or instances. The last column refers to status where 0(zero) stands for healthy patient and 1(one) stands for Parkinsons patient and the main aim of the data is to classify the healthy and Parkinsons patient. The characteristics attributes are given in tabular form [16] where MDVP stands for Multi-Dimensional Voice Program.

TABLE I
CHARACTERISTICS ATTRIBUTES OF PARKINSON'S DATASET

Attribute Number	Attribute Name	Attribute Description
1	Name	ASCII subject name and recording number
2	MDVP:Fo(Hz)	Average vocal fundamental frequency
3	MDVP:Fhi(Hz)	Maximum Vocal fundamental frequency
4	MDVP:Fho(Hz)	Minimum Vocal fundamental frequency
5	MDVP:Jitter(%)	Five measures of variation in fundamental frequency (5-9)
6	MDVP:Jitter(Abs)(%)	
7	MDVP:RAP	
8	MDVP:PPQ	
9	Jitter:DDP	
10	MDVP:Shimmer	Six measures of variation in amplitude (10-15)
11	MDVP:Shimmer(dB)	
12	Shimmer:APQ3	
13	Shimmer:APQ5	
14	MDVP:APQ	
15	Shimmer:DDA	
16	NHR	Two measures of ratio of noise to tonal components in the voice (16-17)
17	HNR	
18	RPDE	Two nonlinear dynamical complexity measures(18-19)
19	DFA	
20	D2	Correlation Dimension
21	spread1	Three nonlinear measures of fundamental frequency variation (21-23)
22	spread2	
23	Status	Health of the subject (1) - Parkinson's, (0) -healthy

B. Status vs Attribute Analysis

Different attributes have different impact on the status of patient. Here in the figures shows range of values of attributes for denoting the status of a patient.

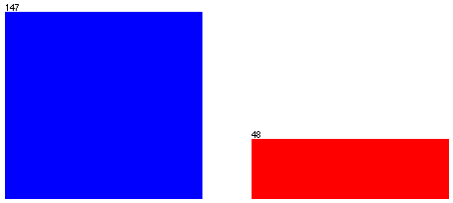

Fig. 1. Pictorial Representation(Blue='1',Parkinson's Patient;Red='0',Normal Patient) of Patient Status

In the above figure the blue bar represents 147 instances are from Parkinsons Patient and red bar represents 48 instances from Normal Patient.

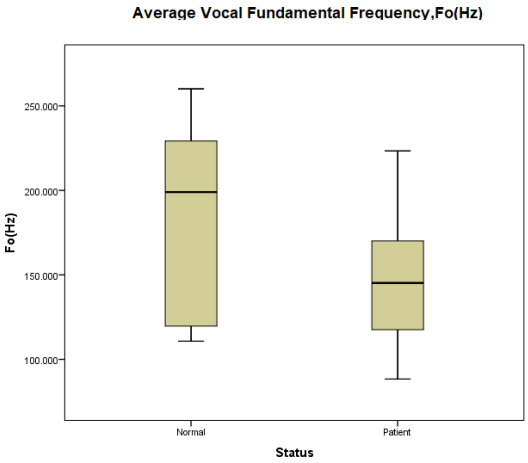

Fig. 2. Status vs Average Fundamental Frequency

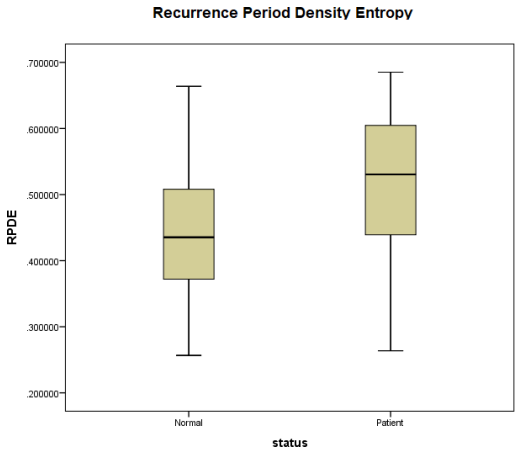

Fig. 3. Status vs Recurrence Period Density Entropy

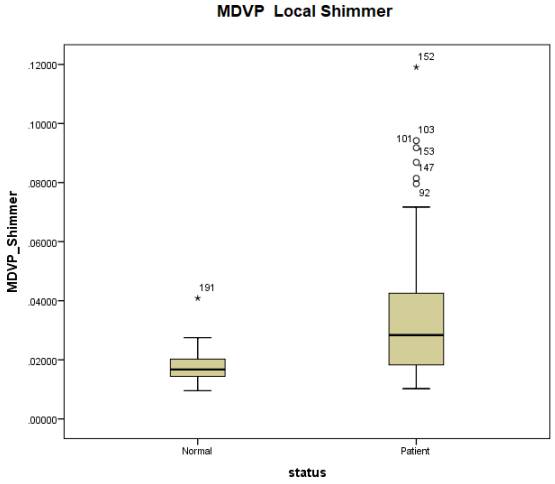

Fig. 4. Status vs MDVP Local Shimmer

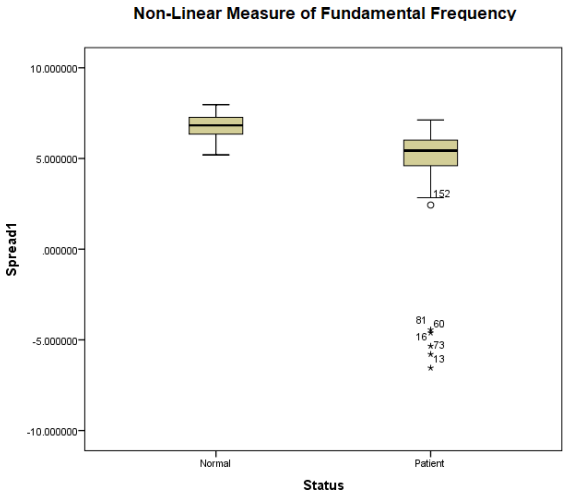

Fig. 5. Status vs Spread1

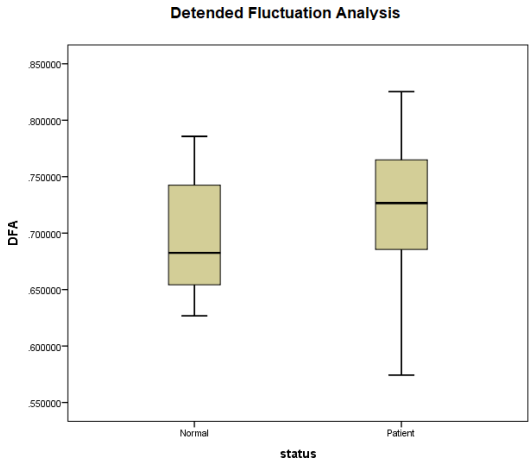

Fig. 6. Status vs Detended Fluctuation Analysis

C. Observations

It is observed from the above figures that in the case of Parkinson's Patient the fundamental frequency ranges from 110-260 Hz while in case of Normal Patient the frequency range is 88-223 Hz.Moreover, in the case of shimmer, the range for normal patient .00954-.04087 whilst for Parkinson's Patient the range is .01022-.11908 which is higher than the normal patient.For Recurrence Period Density the range for Parkinson's Patient is slightly higher than the Normal patient. On the other hand in the case of Spread1, the variation range for Parkinson's Patient is lower than normal Patient.

D. Classification Algorithms

The classification algorithms have used in our study are :
1.Support Vector Machine (SVM),
2.K-Nearest Neighbors,
3.Random Tree,
4.Naive Bayes,
5.Logistic Regression.
We have also used Multi-Layer Perceptions with different number of hidden layers and relevant attributes.

E. Relevant Attribute Selection

Before Classification, it is necessary to sort out the most relevant attributes for a particular classification method . Because the irrelevant attributes have a negative impact on the performance of classification and may also result in a poor outcome.

TABLE II
RELEVANT FEATURES FOR CLASSIFICATION ALGORITHMS

Classification Algorithms	Number of Attributes	Name of the Attributes
Support Vector Machine	5	MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Shimmer,HNR, Spread1
K-Nearest Neighbors	5	MDVP:Fo(Hz), MDVP:Jitter(%), MDVP:Shimmer(dB), RPDE,DFA
Random Tree	1	MDVP:Fo(Hz)
Naive Bayes	3	MDVP:Fo(Hz), MDVP:Fhi(Hz),PPE
Logistic Regression	3	RPDE, DFA,Spread1

IV. RESULTS

A. Comparison

Different Classification Models such as Support Vector Machine(SVM),k-Nearest Neighbors,Naive Bayes, RandomTree, and Logistic Regression are used for classification and different performance parameters such as Accuracy,Recall,Precision, and F-score are evaluated. For analysis, at first the whole Training Dataset is used and after that, the dataset is divided into train and test data where only one parameter, accuracy, is evaluated. Following tables corresponds to the result. It is visible that k-Nearest Neighbors perform better than other classification methods for each test dataset. If consideration of each classification algorithms individually . For Support Vector Machine performs best with 20% test data.Following that for Nave Bayes the performance is better for 20% test data while for Random Tree the accuracy is greater for 25% data split. Finally, for Logistic Regression better accuracy is observed with 20% test data.

TABLE III
CLASSIFICATION EVALUATION ON TRAINING DATASET USING WEKA

Algorithms Name	Accuracy (%)	Recall (Weighted Average)	Precision (Weighted Average)	F-score (Weighted Average)
Support Vector Machine	88.7179	0.887	0.887	0.880
K-Nearest Neighbors	99.4872	0.995	0.995	0.995
Random Tree	100	1.000	1.000	1.000
Naive Bayes	86.1538	0.862	0.858	0.857
Logistic Regression	89.7436	0.897	0.895	0.895

TABLE IV
CLASSIFICATION MODEL ACCURACY ON TRAIN-TEST DATASET USING WEKA

Train Set (%)	Test Set (%)	Support Vector Machine	K-Nearest Neighbor	Naive Bayes	Random Tree	Logistic Regression
85	15	89.6552%	96.5517%	86.2069%	79.3103%	96.5517%
80	20	92.3077%	94.8718%	89.7436%	79.4872%	97.4359%
75	25	89.7959%	91.8367%	85.7143%	81.6327%	91.8367%
70	30	86.2069%	93.1034%	82.7586%	74.1379%	87.931%
65	35	85.2941%	92.6471%	80.8824%	67.6471%	88.2353%
50	50	84.5361%	89.6907%	84.5361%	75.2577%	87.6289%

B. Multi-Layer Perception

Multi-layer perception is a feed-forward neural network working with the backpropagation algorithm. The number of inputs equal to a number of columns in dataset or number of total attributes. There is a hidden layer between the input and output layer. Here we started with one hidden layer and increase the number of hidden layer up to three layers.

TABLE V
RELEVANT ATTRIBUTES FOR DIFFERENT NUMBER OF LAYERS

Number of Layers	Name of the Attributes
One	MDVP:Fo,MDVP:Flo, MDVP:RAP,Shimmer:APQ3,Spread1
Two	MDVP:Fo ,MDVP:RAP,RPDE,DFA,PPE
Three	MDVP:Fo, MDVP:Fhi, MDVP:RAP, DFA,Spread1,Spread2,PPE

The relevant attributes are selected for different number of layers. It is observed that the dependency on different attributes due to difference in number of layers because back propagation algorithm results in selection of more related attributes in order to acheive disired result.. It is observed from the above table that the relevancy of attributes with respect to Number of hidden layers varies and Fundamental Frequency(MDVP:Fo)is the prominent attribute for each number of layers.Moreover,the learning rate, learning momentum ,batch size and Training time is considered same for all the layers.At first, the algorithm is applied on only training dataset and after that it is applied on train-test dataset for further evaluation.

TABLE VI
SUMMARIZED PERFORMANCE EVALUATION OF TRAINING SET

Number of hidden Layers	Accuracy(%)	Batch Size	Learning Rate	Learning Momen-tum	Training Time
One	98.4872	100	0.1	0.3	1000
Two	99.4872	100	0.1	0.3	1000
Three	98.9744	100	0.1	0.3	1000

TABLE VII
SUMMARIZED PERFORMANCE ACCURACY ON TRAIN-TEST WITH DIFFERENT HIDDEN LAYERS

Train Set (%)	Test Set (%)	One Hidden Layer	Two Hidden Layer	Three Hidden Layer
90	10	100%	94.7368%	89.4737%
80	20	87.1795%	97.4359%	92.3077%
75	25	87.7551%	89.7959%	91.8367%
70	30	87.931%	91.3793%	84.4828%
65	40	89.7436%	93.5897%	78.2051%
50	50	89.6907%	93.8144%	86.5979%

It is observed the performance accuracy increases with the number of hidden layers because with increasing number of hidden layers between the input layer and output layer enable the whole model to acknowledge more related features from the input layers. In the case, just one hidden layer with 13 neurons and relevant attributes the performance accuracy is 98.4872% which is greater than the accuracy 92.31% found by David Gil A,Magnus and JohnsonB[18]. Moreover, it is observed with appropriate selection of learning rate, learning momentum and with the increase of hidden layers the performance accuracy increases.

V. CONCLUSION

It might get difficult for bioinformatics practitioners or medical practitioners to analyze a large volume of medical history and analyze the relevancy of various features concerned with a disease and shed light on a particular medicine or recommended exercise . With a view to easing the analysis through Classification Algorithms with selection of relevant attributes using Wrapper method, can pave the way of complex problem-solving. In this study,performance comparison between Classification algorithms is presented and in most of the cases the performance accuracy, precision found better than previous work. Such comparative analysis on Parkinsons Voice Database may further encourage and provide insight into improvement in classifying Patient with Parkinsons.

REFERENCES

[1] Antoni Barikdaar, Tahera Ahmed, and Shamima Parvin Lasker, "The situation of Elderly in Bangladesh,"*Bangladesh Journal of Bioethics.*, vol. 7,no.1, pp. 27 36, 2016.

[2] S. K. V. D. Eeden, C. M. Tanner, A. L. Bernstein, R. D. Fross, A. Leimpeter,D. A.Bloch, and L. M. Nelson, "Incidence of Parkinsons disease: Variation by age,gender, and race/ethnicity," *Amer. J.Epidemiol.*, vol. 157, pp. 1015 1022, 2003.

[3] Gil, D., Manuel, "Diagnosing Parkinson by using Artificial Neural Networks and Support Vector Machines," , 2009.

[4] D. M. Huse, K. Schulman, L. Orsini, J. Castelli-Haley, S. Kennedy, and G.Lenhart, "Burden of illness in Parkinsons disease," *Movement Disord.*, vol. 20, pp. 1449 1454, 2005.

[5] Zahari Abu Bakar, Nooritawati Md Tahir, Ihsan M. Yassin, "Classification of Parkinsons Disease Based on Multilayer Perceptrons Neural Network," in*6th International Colloquium on Signal Processing & Its Applications (CSPA)*, 2010

[6] Dr. R. Geetha Ramani, G. Sivagami, "Parkinson Disease Classification using Data Mining Algorithms," in*International Journal of Computer Applications.*, vol.32,no.9,October 2011

[7] Arvind Kumar Tiwari, "Machine Learning Based Approaches for Prediction of Parkinsons Disease," *Machine Learning and Applications: An International Journal(MLAIJ).*, vol. 3,no.2 , June 2016.

[8] Ipsita Bhattacharya, M.P.S Bhatia,"SVM Classification to Distinguish Parkinson Disease Patients,"in*Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India.*, September 16-17,Tamilnadu, India,2010

[9] Farhad Soleimanian Gharehchopogh, Peyman Mohammadi, "A Case Study of Parkinsons disease Diagnosis using Artificial Neural Networks," *International Journal of Computer Applications (0975 8887).*, vol 73, no.19,July 2013.

[10] R.Das,"A comparison of multiple classification methods for diagno- sis of Parkinson disease,"*Expert Systems with Applications.*, vol. 37,pp. 1568 1572, 2010.

[11] A.Bourouhou,A.Jilbab,C.Nacir, A.Hammouch, ", Comparison of Classification methods to detect Parkinsons Disease," in*2nd International Conference on Electrical and Information Technologies (ICEIT)*, 2016

[12] V. Sachnev and H.J. Kim,, Ihsan M. Yassin, "Parkinson Disease Classification based on binary coded genetic algorithm and Extreme learning machine," in*Ninth IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, 2014

[13] M. Shahbakhi, D.T. Far and E. Tahami,"Speech Analysis for Diagnosis of Parkinsons Disease Using Genetic Algorithm and Support Vector Machine,"*Journal of Biomedical Science and Engineering*, vol. 7,pp. 147 156, 2014.

[14] P. F. Guo, P. Bhattacharya,N. Kharma,"Advances in Detecting Parkin- sons Disease,"*Medical Biometrics. Lecture Notes in Computer Science*, vol.6165,pp. 306 314, 2010.

[15] Marius Ene, "Neural network-based approach to discriminate healthy from those with Parkinson's disease,",Annals of the University of Craiova, Math. Comp. Sci. Ser.Volume 35, 2008.

[16] Fie Ye, "Evaluating the SVM model based on a hybrid method using swarm optimization techniques in combination with a genetic algorithm for medical diagnosis" ,Multimedia Tools Application,2016.

[17] Tarigoppula V.S. Sriram, M. Venkateswara Rao, G.V. Satya Narayana and D.S.V.G.K.Kaladhar,"Diagnosis of Parkinson Disease Using Machine Learning and Data Mining Systems from Voice Dataset,"*Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA).*, vol 1, pp. 151 156, 2014.

[18] David Gila, Magnus Johnson B, "Diagnosing Parkinson by using Artificial Neural Networks and Support Vector Machines," *Global Journal of Computer Science and Technology.*, vol 9, Issue 4,pp. 63−71, 2009.