International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

Application of Data mining in Analysis and detection of Parkinson's Disease

[1]Omini Rathore, Mrs., [2]P. Akilandeswari, [3]Namrata Yadav

ABSTRACT.--*Parkinson's disease (PD) is a neurodegenerative disorder which often affects patients' movements. Some of the most common symptoms of Parkinson's disease are tremors, rigidity, akinesia, walking disability, and postural instability. The primary motor symptoms are collectively called "parkinsonism". This paper provides a brief description of the existing techniques used in detecting Parkinson's Disease with the help of various data mining algorithms such as Multiple Instance Learning (MIL), K-means clustering, Decision Tree Classification, Moving Average Algorithm etc., their accuracies and drawbacks and also gives an outline of the proposed system. Since all of the existing models consider a single symptom for detecting Parkinson's, the proposed approach aims at building an analytical model with two different symptoms i.e. speech and finger tapping keystroke, so as to increase the accuracy and find the co-relation between these symptoms.*

Keywords*--Parkinson's disease, data mining, SVM, Logistic regression, keystroke.*

I. INTRODUCTION

Presently, Specialists diagnose Parkinson's Disease via different neurological examinations. Since there are not many standard tests for detecting Parkinson's disease, therefore, a statistical approach has been proposed. The datasets are based on particular symptoms, some of which are described below-

i. Tremors: The most common symptom which can be seen in PD patients is a course slow tremor of the hands when they are at rest. However, it gradually disappears while doing some kind of voluntary movement of the affected arm and in sleep at the deeper stages. It mostly appears in only one hand, slowly affecting both as the disease progresses. Tremor results in micrographia which is a disorder that features abnormally small, cramped handwriting or progressively smaller handwriting.

iiBradykinesia: PD patients experience slowness and hindrance in their movements which is because of motor planning disturbances in the movement initiation. It has some problems associated along the whole course of the movement, from starting until the completion of the movement.

iii. Rigidity: It is resistance or firmness in the movements of the limb caused due to increment in muscle tone, and incessant and extra contraction of muscles.

[1] *Bachelors in technology, Computer Science and Engineering, SRM IST, Chennai, India,omini1997@gmail.com*

[2] *Bachelors in technology, Computer Science and Engineering, SRM IST, Chennai, India,yadavny28@gmail.com*

[3]*Assistant Professor, Computer Science and Engineering, SRM IST, Chennai, India, akilandeswari.p@ktr.srmuniv.ac.in*

Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020

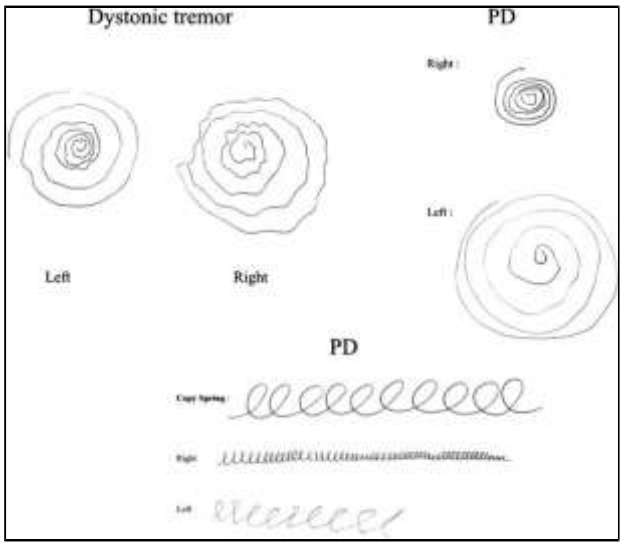


Fig 1.1. Handwriting Samples of patients diagnosed with Parkinson's.

ivPostural instability: It is common in the later phase of this disease, resulting in disability to balance and falling frequently which may cause fractures, underconfidence, and even decreased mobility.

The data mining algorithms used for classification includes-

i. Multiple Instance Learning: It is a type of supervised learning. The learner receives a set of labelled bags each of which contains many instances, and not just receiving a separately labelled group of instances. Consider the basic case of multiple-instance learning classification, if the bag contains only negative instances, then the bag may be labelled negative, while if the bag contains at least one positive instance, it is labelled as positive. The machine tries to generate a concept for properly labelling individual instances from a group of labelled bags or it will try to learn a way to label bags without generating a concept.

ii. K-Means: A collection of objects is given, with each of them having n measurable attributes, k-means algorithm builds an analytical model. K groups or clusters are identified (for any randomly chosen value of K) on the basis of the nearness of objects. K-means is an iterative algorithm, local maxima is generated after each iteration, it is an analytical technique, it identifies k clusters based on the objects' proximity for any chosen value of k to the center of the k groups. The objects are re-grouped after each iteration into the cluster with most proximity. The center of the cluster is calculated by estimating the mean of each cluster's n- dimensional vectors of attributes.

iii. Decision tree classification: Decision tree is a graphical representation of all possible outcomes of a decision which is based n some conditions. Decision trees are easy to understand and interpret. Decision tree uses CART- Classification and Regression tree algorithm. According to this algorithm, the root receives the entire training data as input and the internal nodes receive a list of rows. Atrue/false question is asked at each node about one of the attributes in the dataset, in response to which the dataset is partitioned into two subsets. These subsets then become input to the two child nodes and so on. The goal is to unmix the labels as the tree branches into the leaf nodes or termination nodes. The features or attributes must be categorical, if they are continuous, they must be converted into categorical values before proceeding with

the algorithm. In order to decide which attribute must be questioned at each node, two selection measures are used: a) Information gain b) Gini index.

Information gain, measures decrease in entropy and decides which attribute should be selected as a decision node at each level. Entropy is the measure of randomness in the dataset. If the dataset is completely pure or completely impure the value of entropy will be 0 but if the dataset is equally pure and impure i.e. the number of yes is equal to the number of no, then the value of entropy is 0.5 which is the highest. Gini index is used to calculate the amount of uncertainty. Entropy is calculated using the formula:

$$E = -\sum i * p_i * \log p_i$$

where p_i is the ratios of elements of each label in the set.The information gain of the dataset S for an attribute A is calculated using the formula:

$$IG(S, A) = H(S) - \sum \frac{|S_v|}{|S|} * H(S_v)$$

where H(s) = entropy (of the set S)

iv. Moving Average Algorithm: Moving average is also called running average, rolling average or moving mean. It creates a series or range of mean of various subsets of the whole data set and then analyses the data points through certain calculations. Moving average algorithm is mainly used for time series data and financial data but it has been used in the detection of Parkinson's disease through tremors.

v. Support Vector Machine: A Support Vector Machine comes under supervised learning algorithms. SVM is supported for models with extreme cases. It is used to segregate two classes by drawing a decision boundary also known as a hyperplane. The training data is mapped on this hyperplane and the points that are close to the opposing class form the support vectors. The margins are set based on these support vectors. The points in SVM are referred to as vectors.

A linearly separable dataset uses Linear Support Vector Machine but SVM can also be used in multi-dimensional datasets. In order to decrease the computation cost for converting a low dimensional space to a high dimensional space a kernel trick or kernel function is used. SVM find applications in medical imaging, studying air quality in urban areas, medical classification, image interpretation, time series prediction, and financial data analysis. Although SVM has some disadvantages too, it gives poor performance for the datasets with the number of features greater than the number of samples and SVMs do not provide probability estimates which then have to be calculated using expensive k-fold cross-validation.

Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020

Algorithm	Parameters	Data set
Multiple Instance Learning (MIL) Algorithm	-Dyskinesia -Tremors in hands	ADNi database Medpix
K-means clustering, Decision tree classification algorithm	-Dyskinesia -movement characteristics	100forparkinsons, Clinical trials
Moving Average Algorithm	-Tremors	100forparkinsons, Medline
Artificial Intelligent Algorithms	-movement characteristics -direction changes	Kaggle, Medpix
Support vector machine (SVM) algorithm	-rigidity -tremors -handwriting markers	handwriting samples from 37 PD patients and 38 sex- and age-matched controls

Table 1.1. Survey Table

II. EXISTING SYSTEMS

P. Bonato, D.G. Standaert, D.M. Sherrill, S.S. Salles, M. Akay proposed “data mining techniques to detect motor fluctuations in Parkinson's disease”. They used accelerometer (ACC) and surface electromyographic (EMG) signals as their algorithms in which the main focus is on specific clinical application. This data mining technique can be used in the analysis of huge data sets derived from wearable sensors.F. Widjaja, P. Poignet, W. T. Ang, C. Y. Shee, and W. L. Au, in their paper“Towards a sensing system for quantification of pathological tremor", proposed an algorithm that involved the use of sEMG system and accelerometers in order to tracktremors in the patients’ upper limb. The data received through the system can be used to model the tremors for engineering purpose. Samarjit Das, Breogan Amoedo, Fernando De la Torre, Jessica Hodgins proposed “detection of Parkinsons' symptoms in uncontrolled home environments using a multiple instance learning approach”. They used a weakly supervised learning framework to study and detect the symptoms in patients at home and not in a controlled laboratory.

Yi Liu, Chonho Lee, Martin J. McKeown, Bu-Sung Lee, James K.R. Stevenson proposed “the analysis of visually guided tracking performance in Parkinson's disease”. According to their paper, there are notable differences in the motor movement of PD patients with dyskinesia and without dyskinesia. Dyskinesia is the difficulty in voluntary muscle movements. They have used Decision tree classification and K-means clustering algorithms to track the performance of PD patients through visual guidance. K-means gave better accuracy as compared to decision tree. Thus, data mining techniques can be used to observe the differences between the patients who have dyskinesia and the ones who don’t.

Ziqian Dong, N. Sertac ArtanU Kit Pun, and Huanying Gu stated the “use of avisualization tool for detecting PD and classification of gait data”. They have followed a statistical and graphical approach using

various data mining techniques. The classification process includes data selection, features selection, visualization, and formula integration.

Irena Rektorová, Lucia Masarová, Zdeněk Smékal, Peter Drotár, Jiří Mekyska, Marcos Faundez-Zanuy proposed "a decision support framework for PD based on handwriting markers using Support vector machine algorithm". Since PD affects various motor movements, they have used these aspects as parameters in each task. These parameters were then fed to the SVM for diagnosis. The results showed an accuracy of over 88%, thus proving that handwriting can be used as a valuable marker for the diagnosis of PD.

Nancy Huggins, David Standaert, John Growdon, Shyamal Patel, Richard Hughes, Jennifer Dy, Paolo Bonato proposed "the use of wearable sensors to predict the severity of symptoms and motor complications in late stage Parkinson's Disease". Their analysis was based on readings received from wearable sensors worn by PD patients. They have used Support vector machine algorithm for the implementation of the above approach.

J. Synnott, G. Moore, L. Chen, C.D. Nugent have used a computer vision-based approach in their paper "Assessment and visualization of Parkinson's disease tremor". Moving average algorithm is used to calculate the amplitude of tremor and an easy to use tool is developed to assess PD through 3D techniques.

Cristian F. Pasluosta, Jochen Klucken, Bjoern M. Eskofier Heiko Gassner, and Juergen Winkler have tried to shift the approach from data mining to internet-of-things in their paper "An Emerging Era in the Management of Parkinson's Disease: Wearable Technologies and the Internet of Things". The data collected from wearable tools in PD patients is fed to different artificial learning algorithms. IoT and artificial intelligence can prove to be a boon for the healthcare field.

The paper "Exploiting Nonlinear Recurrence and Fractal Scaling Properties forVoice Disorder Detection" by Declan AE Costello Patrick E McSharry, Stephen J Roberts and Irene M.Moroz states the use of DFA(Deterministic Finite Algorithm) algorithm to calculate the dimensions of correlation, successfully separate normal from disordered subjects. They have used bootstrap resampling for validation of their study.

Research in Parkinson's disease in India written by Pratibha Surathi, Pramod Kumar Pal, Ketan Jhunjhunwala, and Ravi Yadav stated that India's condition is very different from the rest of the world in terms of PD, be it epidemiology or genetics or response to treatment. Therefore, more research is needed to be encouraged, to understand the characteristics of the disease among the Indian population. And hence the further research is still going on.

Andreas Kuhner, Isabella Katharina Wiesmeier, Volker Arnd Coenen, Wolfram Burgard, Tobias Schubert, Massimo Cenciarini, Cornelius Weiller, Christoph Maurer proposed the use of Random forest algorithm for differentiating PD patients from healthy people in their paper "Correlation between motor systems across different Motor tasks, quantified via Random Forest Feature classification in Parkinson's Disease". It involved various tasks that involved motor movements, the algorithm was applied to the data set obtained from each task and accuracies were compared. It was noted that standing up had the highest accuracy.

Bastiaan R. Bloem, Alice Nieuwboer, Bart Post, Evžen Růžička, Christopher Götz, Johan Marinus, Quincy J. Almeida, Leland Eric Dibble, Glenn T. Stebbins, Pablo Martínez-Martín, Anette Schrag stated "measurement instruments to assess posture, gait, and balance in Parkinson's disease: Critique and

recommendations". They surveyed different journal papers to find out various tools that measure important symptoms of PD. The tools or instruments were given scores on the basis of clinical results and different impact factors such as UPDRS. The scores divided the instruments into three categories, recommended, suggested, listed.

"Factors associated with freezing of gait in patients with Parkinson's disease" is written by Byeong C. Kim, Sung Eul Choi, Hyunjean Jung, Geum-Jin Yoon. According to them, Freezing of gait (FOG) is a common and debilitating problem in PD patients. They estimated the prevalence of FOG and identified the independently contributing factors to FOG in PD patients.

III. DRAWBACKS

Individual analysis of every symptom has some drawback attached to it such as handwriting samples can be affected by various other factors that can influence the motor movement of the fingers and hand, in speech recognition additional steps such as noise removal and speech segmentation are required, making it a complex process, and using breath samples has proved to fail to meet clinically relevant results.

IV. CONCLUSIONS

The existing systems include the use of wearable technologies through the implementation of Internet of things, handwriting as a marker for the diagnosis of PD using support vectormachine achieving the accuracy of 88.13%, providing an easy to use tool fordiagnosing Parkinson's through three dimensional visualization techniques, using data mining techniques to track performance of PD patients through visual guidance and using voice and speech data to detect Parkinson's.

The proposed system aims at achieving an accuracy of above 90% by using two different symptoms i.e. voice and finger tapping keystroke. Because of the unavailability of datasets with multiple symptoms, the model is based on the assumption that both the symptoms are of the same patient.

The voice dataset is created by "Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado", who recorded the speech signals. The primary study is based on the methods for feature extraction for basic speech disorders. This dataset comprises of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each field in the table is a particular voice measurement, and each row is equivalent to one of 195 voice recording from these individuals. The data aims at discriminating people diagnosed with Parkinson's from the healthy ones on the basis of "status" column which contains the value 0 for healthy and 1 for PD patients. The data is in ASCII CSV format. Each row of the CSV file contains an entry that corresponds to one voice recording. The first column specifies the name of the patient. There are around six recordings per patient. Other columns give values of various attributes such as jitter, shimmer, variations in fundamental frequency etc. The second dataset gives information about the multiple characteristics of finger movement while typing. The dataset comprises of keystroke values acquired from around 200 subjects, with and without Parkinson's Disease (PD), as they normally typed on their own computer in the absence of any supervision, for a period of months or

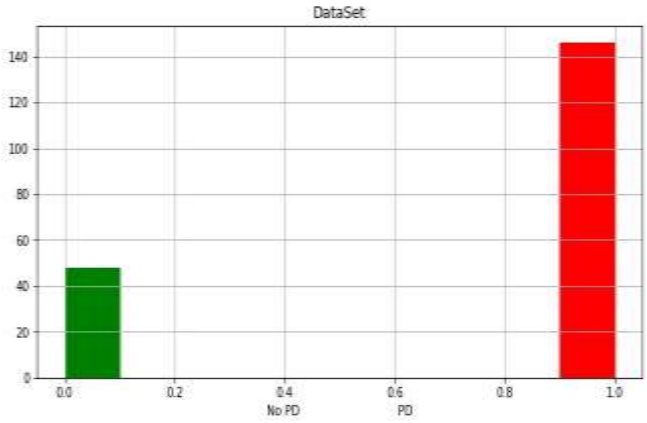
weeks (through the means of a custom keystroke recording app, Tappy). The columns in the dataset are described in the table below.

COLUMN	DESCRIPTION
Birth Year	Year of Birth
Gender	M/F
Parkinson's	If diagnosed with PD or not [true/false]
Tremors	Presence of tremors [true/false]
Diagnosis year	When was PD first diagnosed
Sided	If there is sidedness of movement [left/right/none]
UPDRS	The Unified Parkinson's Detection Rating Scale(UPDRS) score, if known [1-5]
Impact	Impact of PD on their daily life [mild/medium/severe]
Levodopa	If they are using Sinemet and the like [yes/no]
Dopamine agonist (DA)	If they are using a dopamine agonist [yes/no]
Monoamine Oxidase B (MAOB)	If they are using an MAO-B inhibitor [yes/no]
Other	If they are taking any other PD medication [yes/no]
User key	10-character code for the user
Date	Date when the reading was taken [YYMMDD]
Timestamp	The time when the reading was taken HH:MM: SS.SSS
Hand	Left or right key pressed [L/R]
Hold time	The time between press and release for the current key in milliseconds
Direction	Previous to current [LL/LR/RL/RR/S] (S for a space key)
Latency Time	Time between pressing the previous key and pressing the current key (in milliseconds)
Flight time	Time between the release of previous key and press of the current key(in milliseconds)

Table 2. Keystroke data set fields and description

The datasets have been merged on the basis of the status (0-healthy, 1- Parkinson's) field in both the datasets. The final dataset consists of a total of 195 entries and 40 attributes. Merging is followed by data pre-processing which includes converting the categorical data and dropping the missing data.

Fig 4.1. The graph compares the number of patients with PD andwithout PD. The x-axis depicts the status of disease and y-axis depicts the no. of entries in the dataset.


DataSet
140
120
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
No PD
PD

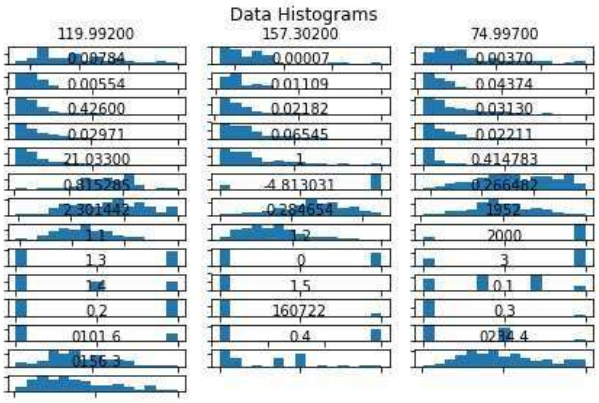


Fig 4.2 The histogram depicts the values of each attribute in a graphical form.

This dataset will be analysed using Support vector machine (SVM) algorithm and Logistic Regression. A Support Vector Machine comes under supervised learning algorithms. SVM is supported for models with extreme cases. It is used to segregate two classes by drawing a decision boundary also known as a hyperplane. The training data is mapped on this hyperplane and the points that are close to the opposing class form the support vectors. The margins are set based on these support vectors. The points in SVM are referred to as vectors. A linearly separable dataset uses Linear Support Vector Machine but SVM can

also be used in multi-dimensional datasets. In order to decrease the computation cost for converting a low dimensional space to a high dimensional space a kernel trick or kernel function is used. A kernel function takes vectors in the original space as input and returns the dot product of the vectors as output. Thus, a kernel function is used to transform a non- linear space into a linear space. Some of the popular kernel types include Polynomial kernel, Radial basis function kernel, Sigmoid kernel etc. SVM has many advantages such as, it is a very effective machine learning algorithm for high dimensional spaces, it is useful when the number of dimensions is greater than the number of samples and it is memory efficient. Also, it provides different kernel functions for various decision functions and more complex hyperplanes can be achieved by adding the kernel functions.

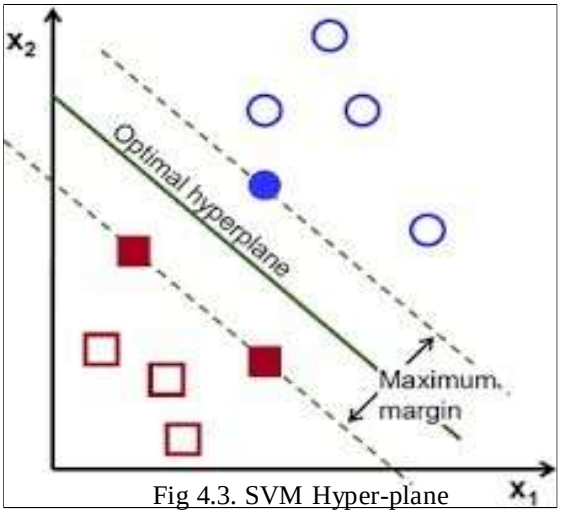


Fig 4.3. SVM Hyper-plane

Logistic regression also comes under classification algorithms. Logistic regression uses logistic function also known as the sigmoid function or logit function. It can be used to fit complex non-linear datasets. More complex decision boundaries can be built by putting complex parameters by using higher order polynomials. The coefficients of the logistic function can be estimated using stochastic gradient descent. Logistic regression is very efficient as compared to other algorithms such as decision trees, it does not require too many computational resources and is affordable. Also, logistic regression is less prone to overfitting as compared to decision tree. This model will serve as a base model for future studies related to Parkinson's disease detection with multiple symptoms. The number of symptoms can be increased from two to four in order to include the four main symptoms of Parkinson's abbreviated as TRAP i.e. tremors, rigidity, akinesia, and postural instability. The presence of these symptoms confirms the diagnosis of Parkinson's.

REFERENCES

1. https://pn.bmj.com/content/15/1/14
2. “Data mining techniques to detect motor fluctuations in Parkinson's disease”;2005; P. Bonato; D.M. Sherrill; D.G. Standaert; S.S. Salles; M. Akay;IEEE.
3. “Towards a sensing system for quantification of pathological tremor", F. Widjaja; C. Y. Shee; W. L. Au; P. Poignet; W. T. Ang; 2007 ;IEEE.
4. “Detecting Parkinsons' symptoms in uncontrolled home environments, A multipleinstance learning approach”; Samarjit Das; Breogan Amoedo; Fernando De laTorre; Jessica Hodgins; 2012; IEEE.
5. “Analysis of visually guided tracking performance in Parkinson's disease”; Yi Liu; Chonho Lee; Bu-Sung Lee; James K.R. Stevenson; Martin J. McKeown; 2014; IEEE. [6] “Classification and visualization tool for gait analysis of Parkinson's disease”; U Kit Pun; Huanying Gu; Ziqian Dong; N. Sertac Artan; IEEE.
6. “Decision Support Framework for Parkinson’s Disease Based on Novel Handwriting Markers”; Peter Drotár; Jiří Mekyska; Irena Rektorová; Lucia Masarová; Zdeněk Smékal; Marcos Faundez-Za; 2015
7. “Using wearable sensors to predict the severity of symptoms and motor complications in late stage Parkinson's Disease” Shyamal Patel; Richard Hughes; Nancy Huggins; David Standaert; John Growdon; Jennifer Dy; 2008; IEEE.
8. “Assessment and visualization of Parkinson's disease tremor”; J. Synnott; L.Chen; C.D. Nugent; G. Moore; 2011;IEEE.
9. “An Emerging Era in the Management of Parkinson's Disease: Wearable Technologies and the Internet of Things”; Cristian F. Pasluosta; Heiko Gassner; Juergen Winkler; Jochen Klucken; Bjoern M. Eskofier;2015; IEEE.
10. 'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007); BMC geriatrics.
11. http://archive.ics.uci.edu/ml/datasets/Parkinsons

12. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation101(23):e215-e220 [Circulation Electronic Pages].
13. http://circ.ahajournals.org/content/101/23/e215
14. Baken RJ, Orlikoff RF: "Clinical Measurement of Speech and Voice. 2nd edition. San Diego: Singular Thomson Learning"; 2000.
15. "Correlation between motor systems across different Motor tasks, quantified via Random Forest Feature classification in Parkinson's Disease"; Andreas Kuhner, Tobias Schubert, Massimo Cenciarini, Isabella Katharina Wiesmeier, Volker Arnd Coenen, Wolfram Burgard, Cornelius Weiller, Christoph Maurer.;2007; Front.
16. "Assessment of fall-related self-efficacy and activity avoidance in people with Parkinson's disease"; Maria H Nilsson, Anna-Maria Drake, and Peter Hagel; 2010 in BMC Geriatrics.
17. "Factors associated with freezing of gait in patients with Parkinson's disease"; SunEul Choi, Hyunjean Jung, Byeong C. Kim, Geum-Jin Yoon; Neurological Sciences in2018 [19]https://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svmhtml