International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

Parkinson Disease Detection Using Machine Learning Algorithms

Yatharth Nakul[1], Ankit Gupta[2], Hritik Sachdeva[3]

Department of Computer Science Engineering, SRMIST, Uttar Pradesh, India

Abstract: ***Parkinson's disease is a global public health concern. Nerve cells, the building blocks of the nervous system in the brain stop producing when they are damaged. Thus, less dopamine is produced that inhibits motor skills and speech. Voice changes in the first stage before brain cells are affected, hence helps identify Parkinson's disease in the early stages and therefore prevent brain cell damage that can lead to reduced fusion and movement. Introduction of different ML algorithms for classification of Parkinson's disease is presented.***

Keywords: Parkinson's Disease, Machine Learning, Computer Science, KNN, SVM, Regression, RF, Decision Tree

1. Introduction

Nowadays, many neurodegenerative diseases had been determined like Alzheimer, Parkinson disease, Arthritic, Lewy's Dementia, etc. Amongst all the neurodegenerative diseases and disorders, the next most common disease followed by Alzheimer is Parkinson Disease. In initial stages of PD least or no expression might be reflected. Arms may stop swinging while walking; voice becomes soft/fuzzy or stammer. As condition progresses symptoms may worsen over time. Although there is no cure for Parkinson's disease, notable improvements in symptoms might be seen after medication. Medical information is important for patient diagnosis and care. Clinical research provides useful information to facilitate treatment development. Medical information management can be shown as a cycle between research, guidelines, quality indicators, performance measures, results and concepts. For integrating clinical information management, data analysis, and application development, Clinical decision testing is emerging in a new environment to facilitate data management from clinical practice, nursing, health care management. As for the clinic ingenious decision-making, machine learning-based methods are used to gain knowledge as well as research phase on evidence-based analysis of data extracted from research reports, testimonial tables, flow charts, guidelines. There are various researchers looking at Parkinson's disease in several ways. To live with Parkinson Disease is tough, as it brings continuous variations in motor-skills and other nonmotor indications can also be observed, along with depression, difficulty in sleeping and mental dysfunctions, hence indirectly affecting families and careers.

2. Machine-Learning Algorithms

Random Forest

Random Forest is one of the supervised machine-learning techniques i.e., used in the Classification and the Regression type of problems. It's an ensemble learning method, i.e. is the way of combining the multiple classifiers to solve a difficult problem and also to boost the accuracy of the model. It's the algorithm which contains a multitude of the decision trees on the various subsets of a given data & also takes the mean value to increase the accuracy of the prediction from the dataset. Rather than depending on the single decision tree, Random Forest predicts from each and every tree & according to the majority votes of predictions, it gives the final output.

Support Vector Machines (SVM)

SVM is useful in solving classification as well as regression problem statements. In the classification problem, two classes are separated or classified by a hyperplane. SVM creates two marginal lines along with a hyperplane having some distance so that they will be easily linearly separable for both the classification points. Along with the hyperplane, two parallel planes are also created passing to the nearest point of the two classes (support vectors) respectively. The distance between the planes is known as the marginal distance which acts as a cushion to divide a point into classes in a better way. The best hyperplane is selected which has the maximum marginal distance.

Naïve Bayes Classifier

Naïve Bayesian classifier is a machine learning algorithm that falls under the category of supervised machine learning, uses the principle of conditional probability as given by Bayes theorem. Implementation of naive Bayes classifiers is easy, with no complex hyper parameters tuning which makes the algorithm specifically useful for a large dataset. It is a probabilistic classifier, which means it predicts based on the probability of an object.

K-Nearest Neighbours Classifier

K Nearest Neighbours classification is a wonderful method to solve nonlinear classified data points. When a new data point is added, in the Classification used case, a k value is selected which indicates how many nearest neighbours are going to be considered in terms of distance. The distance from that new data point can be calculated using Euclidean Distance.

Euclidean Distance (d) between two points P_1 (x_1, x_2) & P_2 (y_1, y_2)

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Another way to find the distance is Manhattan Distance.

After finding the nearest neighbours the one category having less number of nearest points converts these points to the other category having more number of nearest points.

In Regression used case similar k value is selected and the new data point gets the value of mean which is taken of all the nearest neighbours from the new data point as its predicted point.

Decision Tree
A Decision Tree is a tree-structured classifier that consists of two types of nodes-Decision nodes and Leaf Nodes. It has a starting point from where it branches into two or more outcomes. The leaf nodes give the classification or the value of the input example attribute. The DT Classifier can be used for both classification and regression. However, it is more preferred to be used for classification. The Decision tree starts with the root of the tree and based on the value of the test, it goes to the corresponding branch until it reaches the leaf node. The Decision Tree is important in machine learning as it breaks down a complex problem into a simple one.

Logistic Regression
Logistic regression is a technique used for solving classification problems. It's a supervised machine learning algorithm that uses a sigmoid or logit function to convert the output in the range 0 to 1 to return a probability that can be outlined to two or more distinct classes. The algorithm establishes/handles a relation between categorical dependent variables and 1 or more independent variables.

3. Problem Statement

Diagnosis of Parkinson's Disease commonly demands a neurological record of patient with observations of motor-skills in numerous conditions. It gets more difficult for the clinician at the early stages during the diagnosis when motor effects are not yet severe. A patient needs to revisit the clinic often to track the progress of the disease over time.

The productive screening process does not demand a medical visit and can be more helpful. People who are having Parkinson's Disease show distinctive voice characteristics, therefore voice recordings are considered to be a beneficial tool for the diagnosis. Implementation of machine learning algorithms on the speech dataset for accurate diagnosis of the disease would be a productive screening step before visiting the doctor.

4. Objective

Machine learning predictive models will help to classify the people who are healthy and people who are suffering from Parkinson's Disease through ML based methods/algorithms. Different AI-based techniques for the classification are reasonable for being a good support for the expert. The Machine Learning Classification technique will help to improve the accuracy & result of the model and also the dependability of diagnosis and reduce possible loopholes, hence making the PD classification more time-saving.

5. Preliminary Literature Review

Cam M. et alia (2008) [1] presented a neural network consisting of 2 hidden layers to classify between people having Parkinson's disease and people who aren't suffering. Boosting technique has been used and accuracy achieved is greater than 90 on training and testing data.

Max A. et alia (2009) [2] presented a Support Vector Machine (SVM) algorithm to classify between people having Parkinson's disease and people who aren't suffering with the help of dysphonia detection.

Yadav, G et alia (2009) [3] have presented a classification and Support Vector Machine Classifier (SVC) to distinguish between people having Parkinson's disease & those who are not. This provides an 76% accuracy, 97% sensitivity, 13% specificity.

R. Das et alia (2010) [4] presented a neural network implementation with Data mining neural analysis, regression analysis & decision trees for a comparative study on Parkinson's disease speech data set, achieving the accuracy of 92.9%, 84.3%, 88.6%, and 84.3% respectively.

Chen, H et alia (2013) [5] proposed a comparison between Fuzzy-KNN based system and support vector machines (SVM) to a data set consisting of voice measurements of different people suffering from Parkinson's disease. Maximum accuracy achieved using the FKNN classification model is 96.07% using the cross-validation method with 10 folds.

Indira R. et alia (2014) [6] presented an approach, implementing back propagation to distinguish b/w people suffering from Parkinson's disease and those who are not with the help of ANN deep learning model. Filtering technique was used for boosting and also for data reduction principal component analysis was used.

Shahbakhi et alia (2014) [7] presented a Genetic Algorithm (GA) and SVC algorithm to classify between people having Parkinson's disease and people who aren't suffering (healthy people). Fourteen features of Voice signals are based on pitch, jitter, shimmer, and noise to harmonic ratio, which are vital characteristics of the voice signals. This provides classification accuracy of 94.50, 93.66, and 94.22 per 4, 7, and 9 optimize features respectively.

Yahia A. et alia (2014) [8] presented a comparison between naive Bayes Classifier and KNN algorithm using Parkinson's voice dataset with sound recordings of people having Parkinson's disease and healthy people. The accuracy achieved by the KNN classifier and Naive Bayes algorithm is 80% and 93.3% respectively.

Srishti Grover. et alia (2018) [9] presented a Deep Neural Network (DNN) model using TensorFlow, a deep learning library on Parkinson Voice Dataset. Classification model accuracy achieved is 94.44% and 62.73% for training and testing data respectively.

Author Name	Machine Learning Methods	Dataset Description	Performance (Accuracy)
Cam M. (2008)	PNN	Voice-Recording	92.90%
Max A. (2009)	SVM	Speech Signal	91.40%
Yadav G. (2009)	SVM	Voice Dataset	76%
Das R. (2010)	NN Classifier	Voice-Recording	92.90%
Chen H. (2012)	Nested SVM	Speech Dataset	93.50%
Indira R. (2014)	ANN	Speech-Signal	92%
Shahbakhi (2014)	SVM	Speech-Signal	94.22%
Yahia A. (2014)	KNN	Voice Dataset	93.30%
Srishti Grover (2018)	DNN	Voice Dataset	94.44%

6. Methodology

1) Dataset Description

According to UCI [10]:

"Max Little created the following dataset in collaboration with the National Center of Voice and Speech, Denver, Colorado. The dataset is made of a range of biomedical voice measurements of 31 people out of which 23 were suffering from Parkinson's disease.

The data aims to distinguish between people who are suffering from Parkinson's Disease and healthy ones, according to the status column which depicts "one" for PD and "zero" for healthy."

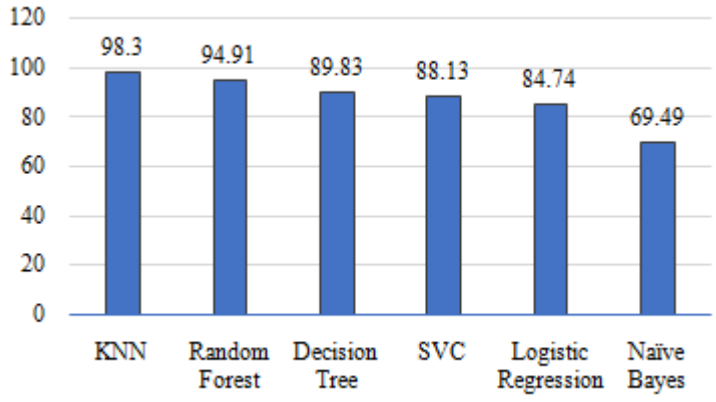
Voice measure	MEANING
MDVP:Fo (Hz)	Average vocal fundamental frequency
MDVP:Fhi (Hz)	Maximum vocal fundamental frequency
MDVP:Flo (Hz)	Minimum vocal fundamental frequency
MDVP:Jitter (%)	Several measures of variation in
MDVP:Jitter (Abs)	fundamental frequency
MDVP:RAP	
MDVP:PPQ	
Jitter:DDP	
MDVP:Shimmer	Several measures of variation in amplitude
MDVP:Shimmer (dB)	
Shimmer:APQ3	
Shimmer:APQ5	
MDVP:APQ	
Shimmer:DDA	
NHR	Two measures of ratio of noise to tonal
HNR	components in the voice
RPDE	Two nonlinear dynamical complexity
D2	measures
DFA	Signal fractal scaling exponent
spread1	Three nonlinear measures of fundamental
spread2	frequency variation
PPE	
status	Health status of the subject: (1) Parkinson's, (0) healthy

2) Machine Learning Classifiers Implementation

The proposed methodology to predict Parkinson's Disease using different machine learning-based algorithms is as follow:

Comparison between 6 machine learning classification algorithms with their accuracy scores is outlined in the table given below. Hyperparameter tuning has been done, iterated over numeric values of the parameters to achieve the maximum accuracy.

	Classifier	Accuracy
1	KNN	0.983051
2	Random Forest	0.949153
3	Decision Tree	0.898305
4	SVC	0.881356
5	Logistic Regression	0.847458
6	Naive Bayes	0.694915



Achieved maximum accuracy of 98.30% using the K Nearest Neighbour classifier by randomly selecting 70% dataset for training and 30% for testing the model.

Classification Report (KNN):

Status	Precision	Recall	f1-Score
0 (healthy)	1	0.38	0.56
1 (PD)	0.85	1	0.92

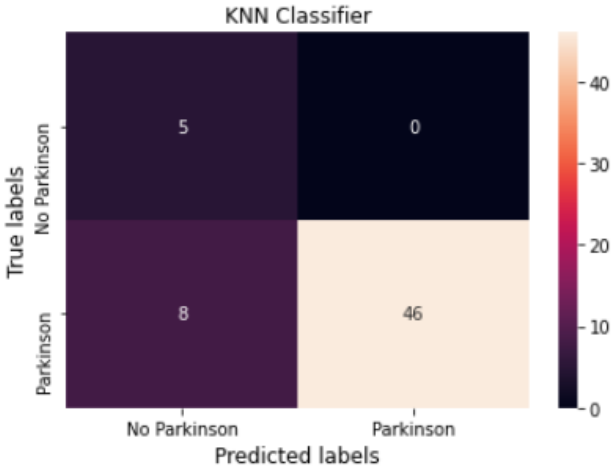
$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$F1 = \frac{Precision * Recall}{Precision + Recall}$$

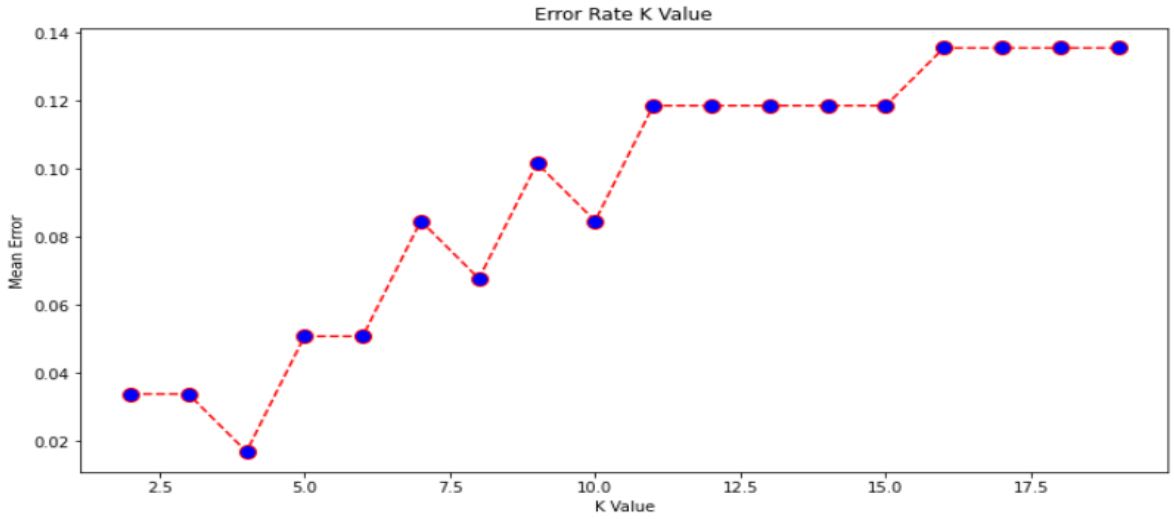
$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

3) Confusion Matrix

A confusion matrix (CM) helps to evaluate the performance of a classifier. The basic ideology is to count the no. of times instances of one class are classified as other. The actual -ve class is represented by the 1st row of matrix and actual +ve class by the 2nd row, 1st column of the matrix denotes the predicted negative class and the 2nd column represents the predicted positive class. Based on the above classification report a confusion matrix is plotted below which clearly describes our model performance on testing data.



Variation in error rate w.r.t. K value in the KNN classifier on the parameter n_neighbour is plotted below.



7. Conclusion

Even today prediction of Parkinson's Disease is one of the most difficult tasks for research engineers and doctors. In this research paper, we compared several ML algorithms for the prediction of Parkinson's disease out of which we found the KNN Classification model with the best accuracy score of 98.30% followed by the Random Forest Classification model with an accuracy score equal to 94.91%.

References

[1] Can, M. (2013). Diagnosis of Parkinson's Disease by Boosted Neural Networks. South East Europe Journal of Soft Computing, 2 (1)

[2] Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N. L., & Yu, W. (2009). Mega SNP Hunter: a learning approach to detect disease predisposition SNPs and high-level interactions in genome wide association study. BMC bioinformatics, 10 (1), 13

[3] Yadav, G., Kumar, Y., & Sahoo, G. A. D. A. D. H. A. R. (2011). Predication of Parkinson's disease using data mining methods: A comparative analysis of tree, statistical, and support vector machine classifiers. Indian journal of medical sciences, 65 (6), 231.

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

[5] 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.

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

[7] 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

[8] Yahia A, Laiali A. (2014). Detection of Parkinson Disease through Voice Signal Features. Journal of American Science 2014; 10 (10), 44-47.

[9] Srishti Grover, Saloni Bhartia, Akshama, Abhilasha Yadav, Seeja K. R. (2018). Predicting Severity Of Parkinson's Disease Using Deep Learning. International Conference on Computational Intelligence and Data Science (ICCIDS 2018)

[10] UCI Machine Learning Repository-Center for Machine Learning and Intelligent System "https://archive.ics.uci.edu/ml/datasets/Parkinsons"