# Classification of Parkinson's disease

**Suresh D[2], Sarath M[3], Nimeesha K[4], Sravani G[5], Sai Venkata Sriram Chowdary K[6].** Students

**P.AnnanNaidu** Asst.Professor[1,]

Department of Computer Science and Engineering Aditya Institute of Technology and Management (A)-Tekkali, Andhra Pradesh.
sriramchaowdary@gmail.com

**Abstract:**

Parkinson's disease is a progressive and chronic neurodegenerative disorder. As the dopamine-generating neurons in parts of the brain become damaged or die, people begin to experience difficulty in speaking, writing, walking, or completing other simple tasks. These symptoms grow worse over time, thus resulting in the increase of its severity in patients. In this paper, we have proposed a methodology for the prediction. Machine learning based method is used to classify between healthy people and people with Parkinson's disease (PD). We are using five machine learning models for the detection of Parkinson disease datasets. SVM, KNN, RF, K-means, NB and DT were utilized for the forecast of Parkinson Disease.

**Keywords:** Parkinson disease, Classification, Random forest, Decision tree, Support vector machine.

## I. Introduction

Parkinson disease (PD) is a neurological disorder based on dopamine receptors. Parkinson disease mostly causes problems in moving around. It can cause a person to move very slowly. Parkinson is a progressive neurological condition, which is characterized by both motor (movement) and non-motor symptoms. Apart from many common symptoms each person will experience and demonstrate an individual presentation of the condition. A person with Parkinson disease appears stiff or rigid. At times, a person with Parkinson disease may appear to suddenly "freeze up" or be unable to move for a short period of time. Parkinson disease is a progressive neurodegenerative condition resulting from the death of the dopamine containing cells of the substantia nigra. There is no consistently reliable test that can distinguish Parkinson disease from other conditions that have similar clinical presentations. The diagnosis is primarily a clinical one based on the history and examination.Although Parkinson disease is predominantly a movement disorder, other impairments frequently develop, including psychiatric problems such as depression and dementia. Autonomic disturbances and pain may later ensue, and the condition progresses to cause significant disability and handicap with impaired quality of life for the affected person. Family and carers might get affected indirectly.

**Symptoms:** Parkinson's disease signs and symptoms can be different for everyone. Early signs may be mild and go unnoticed. Symptoms often begin on one side of your body and usually remain worse on that side, even after symptoms begin to affect both sides.

Parkinson's signs and symptoms may include:

- **Tremor.** A tremor, or shaking, usually begins in a limb, often your hand or fingers. You may rub your thumb and forefinger back and forth, known as a pill-rolling tremor. Your hand may tremble when it's at rest.
- **Slowed movement (bradykinesia).** Over time, Parkinson's disease may slow your movement, making simple tasks difficult and time-consuming. Your steps may become shorter when you walk. It may be difficult to get out of a chair. You may drag your feet as you try to walk.
- **Rigid muscles.** Muscle stiffness may occur in any part of your body. The stiff muscles can be painful and limit your range of motion.
- **Impaired posture and balance.** Your posture may become stooped, or you may have balance problems as a result of Parkinson's disease.

- **Loss of automatic movements.** You may have a decreased ability to perform unconscious movements, including blinking, smiling or swinging your arms when you walk.
- **Speech changes.** You may speak softly, quickly, slur or hesitate before talking. Your speech may be more of a monotone rather than have the usual inflections.
- **Writing changes.** It may become hard to write, and your writing may appear small.

**Risk factors**: Risk factors for Parkinson's disease include:

- **Age.** Young adults rarely experience Parkinson's disease. It ordinarily begins in middle or late life, and the risk increases with age. People usually develop the disease around age 60 or older.
- **Heredity.** Having a close relative with Parkinson's disease increases the chances that you'll develop the disease. However, your risks are still small unless you have many relatives in your family with Parkinson's disease.
- **Sex.** Men are more likely to develop Parkinson's disease than are women.
- **Exposure to toxins.** Ongoing exposure to herbicides and pesticides may slightly increase your risk of Parkinson's disease.

## II. Literature Survey

In this paper[1] the final idea is to provide the neurologist with a decision support system, based on the analysis of motor data acquired by inertial sensors during the performance of a structured protocol, applying machine learning techniques aiming to help him in objective clinical diagnosis of PD patients, since the early stage. This paper [2] primary objective is study to inspect the exhibition of three supervised algorithms for improving Parkinson disease analysis by detection. SVM, KNN, and LR were utilized for the forecast of Parkinson Disease. LR achieved the second-highest classification accuracy of 97%. Also, as far as precision for dissecting Parkinson illness datasets, KNN acquired the worst performance (i.e. 60%). In this SVM obtained the highest performance for analyzing the Parkinson datasets. This perusal has emphasized the current of Parkinson research aptitude and scope in connection to clinical research fields by machine learning techniques. That will be the viable effect in the Parkinson's disease. This paper [3], primary objective in doing so was to improve the performance and the accuracy of the model and also to reduce the computational cost of classification task. In this study developed a FS based decision support system by using the features extracted from speech signals of PD patients and healthy people. Different FS methods were applied to different classifiers and the one with the highest performance was determined. Three FS methods were used in this study and they showed different performance for different classifiers. The best combination of FS method and classification method was determined and used for the diagnosis of PD. It provided about 13% performance improvement for SVM, about 11% for ANN, and about 5% improvement for CART. In addition, classification performance also depends on the parameters of classifiers. In this paper[4],Different machine learning approaches were implemented in these studies to distinguish between two groups (i.e., healthy subjects and PD patients) with prevalence of Support Vector Machine (SVM),2,6,18,21,33,34 Linear Discriminant Analysis (LDA),6,25,27 and Random Forest (RF). The performance of the classifiers evaluated in terms of recall, Specificity, Precision, Accuracy, F-measure. Concerning the use of features coming from both feet, among the three classifiers, the best results were achieved with RF and NB when considering the 2C60 dataset (accuracy and F-measure both equal to 0.97 for RF and to 0.95 for NB), while RF gave better performances when the other two datasets were given as input to the machine learning algorithms (accuracy and F-measure both equal to 0.77 for 3C90 and to 0.98 and 0.97 for 2C90, respectively). In all cases, SVM generally appeared as the worst classifier to be used for this purpose, considering the acquired data.

| S.NO | TITLE | METHODOLOGY USED | PERFOMANCE MATRIX | DRAW BACK |
|---|---|---|---|---|
| 1 | **Upper limb motor pre-clinical assessment in Parkinson's disease using machine learning** | • Random Forest | Accuracy 79 | The present method provides the diagnosis of PD using voice dataset through machine learning algorithms. |
| 2 | **A Comparative Analysis Of Parkinson Disease Prediction Using Machine Learning Approaches** | • Support vector machine | Accuracy 90.26 | The application will be able to detect in Parkinson disease in very few minutes and notify the dangerous probability of having the disease |
| 3 | **Early Diagnosis of Parkinson's Disease Using Machine Learning Algorithms** | • Support vector machine | Accuracy 93.84 | The brief introduction of various computational intelligence techniques based approaches used for the prediction of Parkinson diseases are presented |
| 4 | **Comparative Motor Pre-clinical Assessment in Parkinson's Disease Using Supervised Machine Learning Approaches** | • Random Forest | Accuracy 97.8 | Major challenge is the ability to accurately interpret results generated by the algorithms |

Table-1. Literature review

## III. Methodology

### A. Data collection

This dataset is collected from UCI Machine Learning Repository. The data used in this study were gathered from 188 patients with PD (107 men and 81 women) with ages ranging from 33 to 87 (65.1Â±10.9) at the Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University. The control group consists of 64 healthy individuals (23 men and 41 women) with ages varying between 41 and 82. During the data collection process, the microphone is set to 44.1 KHz and following the physician's examination, the sustained phonation of the vowel /a/ was collected from each subject with three repetitions. We extracted features from the Pd_speech_features datasets. Then, we picked the 754 columns and 755 entries of data and conducted several experiments to checking missing values, redundant values.We had split our existing data into train and test with 70% and 30%, in train data we got 529 rows and 754 columns and in test data we got 227 rows and 754 columns. For Normalizing data we used standard scaler. Standard Scaler follows Standard Normal Distribution (SND). Therefore, it makes mean = 0 and scales the data to unit variance.

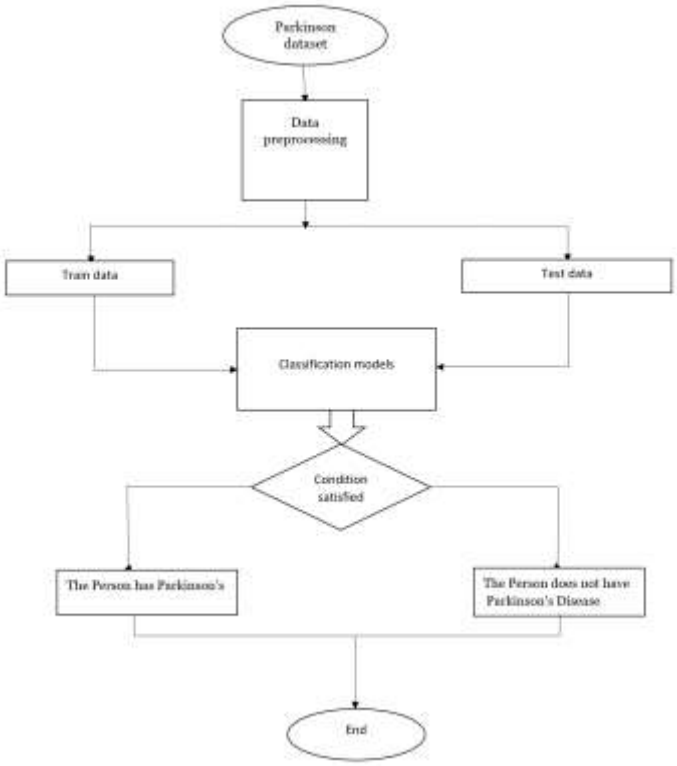


Fig-1: Flow of Machine learning processing steps

We had collected dataset from UCI Machine Learning Repository. We checked for null values as there are no null values presented in the dataset we did normalizing and split the data into training data and testing data. We perform various models on our training data which will be discussed further and with the highest accuracy model we predicted is the person normal or the person effected with Parkinson's disease and that will be the end of the flow chart.

**B. Random Forest:**

Random forest is supervised learning technique that can be used for both classification and regression problems. A process is a combination of multiple classifiers to solve a complex problem and improve the performance of the model is called ensemble learning. It is used by random forest algorithm. Random forest is a classifier which is having number of decision trees on where is subject of the given data set and take the average to improve the predictive accuracy of the data set. The random tree take the prediction from each tree and based on maximum votes off predictions and it predicts the final output. Random forest having more number of tree than it will have high accuracy and it avoids the problem of overfitting.

**C. Decision tree:**

Decision tree is based on supervised learning technique. It can be solved both regression and classification problem but we mostly used for classification problem. Decision tree having two nodes. They are leaf node and decision node. Decision node is used to take the decision. There are multiple branches means it contains multiple decisions. The leaf node is the output of these decisions. Leaf node does not have any other branches. Decision tree ask questions and by the answer will be split. The tree is in to the subtree.

**D. Support vector machine:**

SVM is a simple supervised machine learning algorithm. We can solve classification/regression but we mostly used for classification. SVM having hyperplane which is act as boundary between the types of data. Here each data item in the dataset is plot into N-dimensional space. Where N is the number of attributes in the data. We find the hyperplane to separate the data. SVM can perform binary classification only. If there is multiple classifications then we can use SVM for multiclass problem.

**E. Naive Bayes:**

Naive Bayes is a classification algorithm for Binary and multiclass classification problems. The Naive Bayes classifiers are a collection of classification algorithms based on Bayes theorem. It is not a single algorithm but a family of algorithm where all of them share a common principle it assumes that the presence of a feature in a class is not related to any other teacher so every pair of future being classified is independent of each other.

Formula:

P(A/B)=P(B/A)*P(A)/P(B)

**F.K-Nearest Neighbors-Classification:**

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

**Algorithm**

A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

**Distance functions**

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

It should also be noted that all three distance measures are only valid for continuous variables. In the instance of categorical variables the Hamming distance must be used. It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

**Hamming Distance**

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

| X | Y | Distance |
|---|---|---|
| Male | Male | 0 |
| Male | Female | 1 |

**G. Evaluation Criteria**

In this work, we used three supervised learning strategies for the identification of Parkinson disease. Therefore, the performance measurements of the classifiers are evaluated by various measurable methods. For example, Recall, Precision, f1- measure, etc. Hence, the calculation technique for the estimation considerations are as pursues

Accuracy = (TP + TN) / (TP + FP + TN + FN)

Recall or sensitivity = TP / (TP + FN)

Precision = TP / (TP + FP)

f1 = 2 ∗ (recall ∗ precision)/ (recall + precision)

## IV. Results and Discussion

In this segment, we directed different experiments to assess the 5 machine learning supervised algorithms for recognition of Parkinson Disease. The investigation of five classification techniques was evaluated for the exposure of Parkinson disease data.

**Decision Tree**

Decision tree is a supervised learning technique that can be used for both classification and regression problems but mostly it is preferred for solving classification problems. It is a tree structured classifier where internal nodes represent the features of a data set, branches represent the decision rules and each leaf node represents the outcome. which results in accuracy rate of 80.61.

Table-2. Performance metrics of Decision Tree

| Name Of The Models | Accuracy | Precision | Recall | F1_Score |
|---|---|---|---|---|
| DT_Test | 76.2 | 83.7 | 83.7 | 83.7 |
| DT_Train | 100.0 | 100.0 | 100.0 | 100.0 |

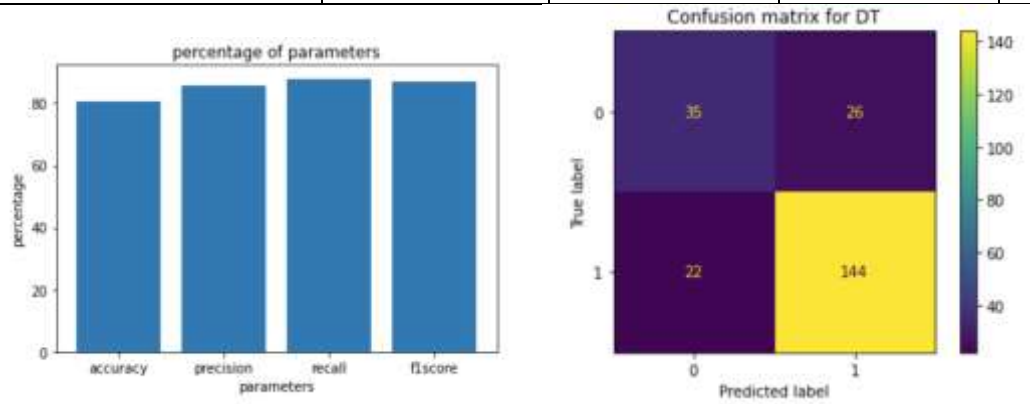


Fig-2: Visualization of Decision Tree Performance

**Support Vector Machine**

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The support vector clustering algorithm created by Hava Siegelman and Vladimir Vapnik applies the statistics of support vector machines algorithm, to categorize unlabeled data, and is one of the most widely used clustering algorithms in applications. Compared to other machine learning methods, SVM has better generalization. SVM has a solid theoretical base and provides more accurate results in many applications than other algorithms. which results in accuracy rate of 83.7%.

Table-3.Performance metrics of SVM

| Name Of The Models | Accuracy | Precision | Recall | F1_Score |
|---|---|---|---|---|
| SVM_Test | 83.7 | 82.1 | 99.4 | 89.9 |
| SVM_Train | 92.2 | 91.2 | 99.2 | 95 |

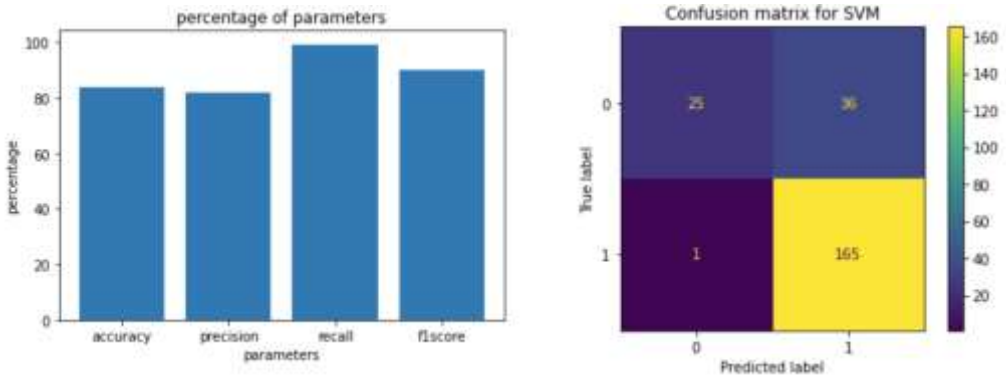


Fig-3. Visualization of SVM Performance

**K-Nearest Neighbour (KNN)**

KNN has no model other than storing the entire dataset, so there is no learning required. Efficient implementations can store the data using complex data structures like k-d trees to make look-up and matching of new patterns during prediction efficient. KNN is one of the simplest machine learning algorithms based on supervised learning technique. This results in accuracy rate of 87.2%.

Table-4.Performance metrics of KNN

| Name Of The Models | Accuracy | Precision | Recall | F1_Score |
|---|---|---|---|---|
| KNN_Test | 87.2 | 85.5 | 99.4 | 91.9 |
| KNN_Train | 91.3 | 90.9 | 98.2 | 94.4 |

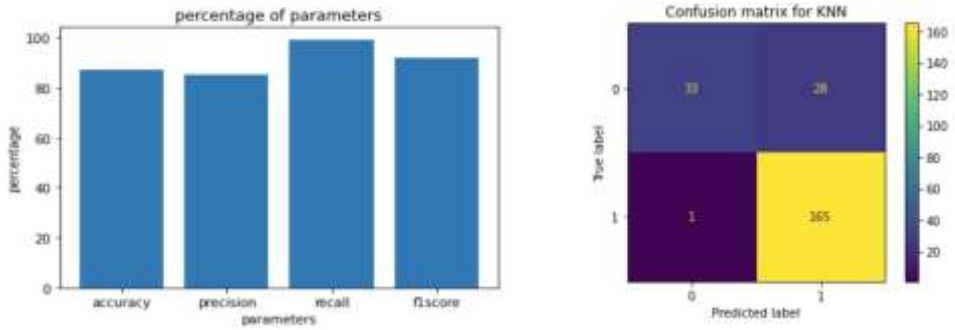


Fig-4.Visualization of KNN Performance

**Random Forest**

It is based on concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest(RF) is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive

accuracy of the dataset. This results in accuracy rate of 85.9%.

Table-5. Performance metrics of RF

| Name Of The Models | Accuracy | Precision | Recall | F1_Score |
|---|---|---|---|---|

| Name Of The Models | Accuracy | Precision | Recall | F1_Score |
|---|---|---|---|---|
| RF_Test | 85.9 | 85.6 | 96.9 | 90.9 |
| RF_Train | 100.0 | 100.0 | 100.0 | 100.0 |
| NB_Test | 76.6 | 86.4 | 80.7 | 83.5 |
| NB_Train | 82.6 | 90.4 | 85.9 | 88.1 |

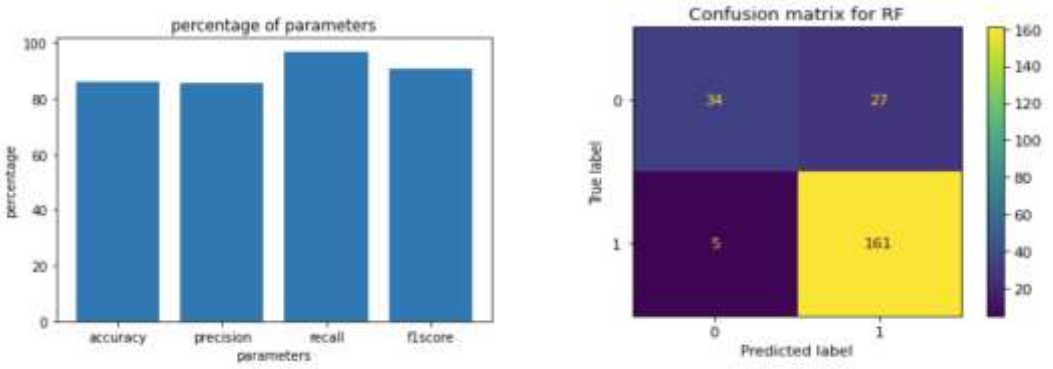


Fig-5.Visualization of RF Performance

**Naive Bayes**

Naive Bayes (NB) is a classification algorithm for Binary and multiclass classification problems. Naive Bayes classifiers are a collection of classification algorithms based on Bayes theorem. It is not a single algorithm but a family of algorithm where all of them share a common principle it assumes that the presence of a feature in a class is not related to any other teacher so every pair of future being classified is independent of each other. This results in the accuracy of 76.6%.

Table-6.Performance metrics of Naive Bayes

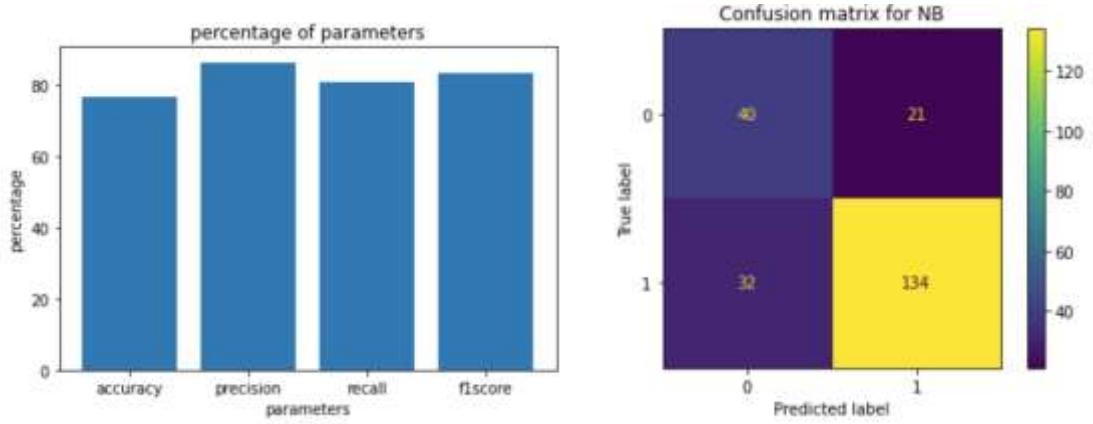


Fig-6.Visualization of NB Performance

| Name Of The Models | Accuracy | Precision | Recall | F1_Score |
|---|---|---|---|---|
| KNN | 87.2 | 85.5 | 99.4 | 91.9 |
| SVM | 83.7 | 82.1 | 99.4 | 89.9 |
| RF | 85.9 | 85.6 | 96.9 | 90.9 |
| NB | 76.6 | 86.4 | 80.7 | 83.5 |
| DT | 76.2 | 83.7 | 83.7 | 83.7 |

Table-7.Comparasion of different machine learning model performance

According to the performance measurements of five classification algorithms are displayed in below figure.

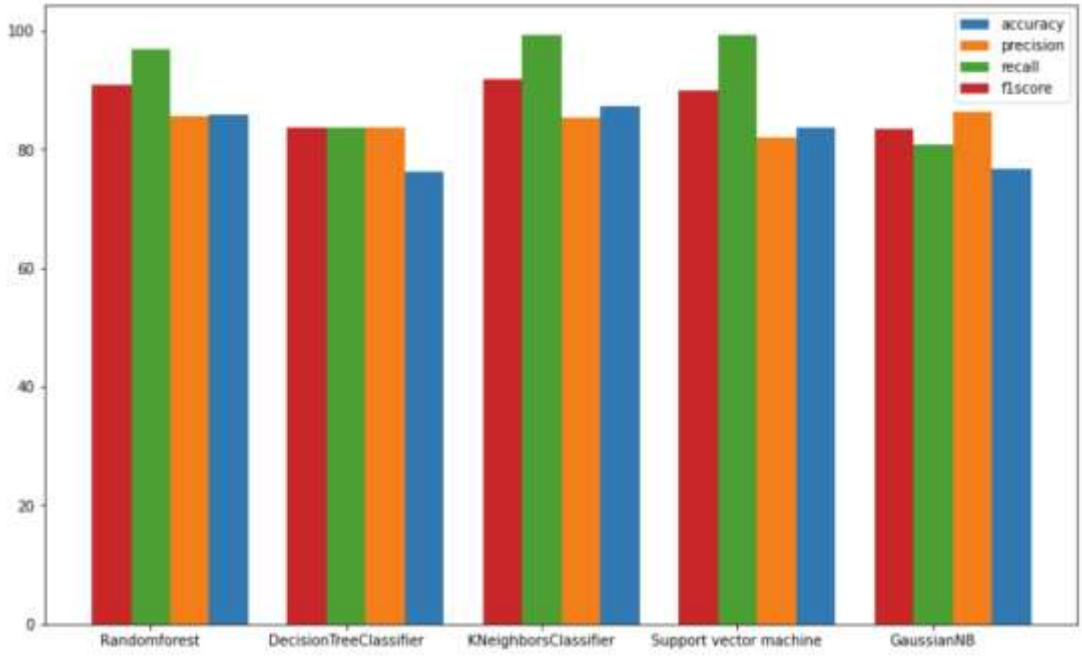


Fig-7. Visualization of Models performance

The results evidently show that the SVM and KNN reached the maximum recall (99.4%) and least recall is NB (80.7%). KNN reached maximum accuracy (87.2%) and least accuracy is DT (76.2%). NB reached maximum precision (86.4%) and least precision is SVM (82.1%). KNN reached maximum f1_score (91.9%) and least f1_score is NB (83.5%).

## Conclusion

This paper describes machine learning models based on Parkinson diseases dataset and also presented a comprehensive work for the prediction of Parkinson disease by using machine learning based approaches. The brief introduction of various Machine learning techniques and approaches used for the prediction of Parkinson diseases are presented. The summary of results obtained by various researchers available in evaluation criteria to predict the Parkinson diseases is also presented.

## References

1. Cavallo F, Moschetti A, Esposito D, Maremmani C, Rovini E. Upper limb motor pre-clinical assessment in Parkinson's disease using machine learning. Parkinsonism Relat Disord. 2019 Jun;63:111-116. doi: 10.1016/j.parkreldis.2019.02.028. Epub 2019 Feb 22. PMID: 30826265.

2. Shamrat, F.M. & Asad, Md.Asaduzzaman & Rahman, A. K. M. & Tusher, Raja Tariqul & Tasnim, Zarrin" A Comparative Analysis Of Parkinson Disease Prediction Using Machine Learning Approaches"https://www.researchgate.net/publication/338924577 1.2576-2580."2019.

3. Karapinar Senturk, Zehra. (2020). Early Diagnosis of Parkinson's Disease Using Machine Learning Algorithms. Medical Hypotheses. 138. 109603. 10.1016/j.mehy.2020.109603." 2020.

4. Rovini E, Maremmani C, Moschetti A, Esposito D, Cavallo F. "Comparative Motor Pre-clinical Assessment in Parkinson's Disease Using Supervised Machine Learning Approaches" Ann Biomed Eng. 2018 Dec;46(12):2057-2068. doi: 10.1007/s10439-018-2104-9. Epub 2018 Jul 20. PMID: 30030773.2018.