Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 15, Issue. 3, March 2026, pg.125 – 133

A NOVEL APPROACH TO DIAGNOSE THE PARKINSON'S DISEASE USING RANDOM FOREST MACHINE LEARNING TECHNIQUE THROUGH VOICE DATASET

D. Karthiga[1]; P. Sumitra[2]

[1]Department of Computer Science and Applications, Vivekanandha College of Arts and Sciences for Women (Autonomous), India
[2]Department of Computer Science and Applications, Vivekanandha College of Arts and Sciences for Women (Autonomous), India
[1] karthigaaa.d@gmail.com; [2] sumitravaradharajan@gmail.com

DOI: https://doi.org/10.47760/ijcsmc.2026.v15i03.014

Abstract: Parkinson's disease is considered as a degenerative disorder that affects humans worldwide. The specialists inspect many symptoms and signs to analyze the condition of the patient's nervous system and diagnose Parkinson's disease by neurological and physical examination. Even with the improved technology early detection of Parkinson's disease remains a challenge. In this paper, a machine learning-based automatic detection model is proposed that assists physicians to treat patients in the early stage. The classification plays a vital role in Parkinson's disease detection in order to save time and improves the treatment. The prior studies indicate various classification techniques that are utilized to obtain greater accuracy. The main issue in the review literature is identifying the efficient classifier. As a part of this paper, we are

using five classifiers that include Random forest, Naive Bayes, K-nearest neighbor, Support Vector machine and Decision tree that was implemented on voice dataset to determine which classifier best suits to detect Parkinson disease .The dataset is collected from the UCI repository. Finally, we conclude that by comparing with other four algorithms and the Random Forest was found to be the suitable classifier that achieved the highest accuracy of 90.78%.

Keywords: Machine Learning, Parkinson's Disease, Classification, Supervised Learning, Random Forest

1. INTRODUCTION

Parkinson's disease is the most common Neurodegenerative disease which affects the human activities including speech [1]. Calne [2] describes that it is the second most common Neurological disorder. Based on the recent study, In worldwide 10 million people are affected by this disease [3]. The reason behind the Parkinson disease is the death of the dopamine producing cells in the substantia nigra which are the parts of the midbrain [4] as in Figure 1.Even though it is incurable, the treatment of early stage will alleviate its effects on patients [5].

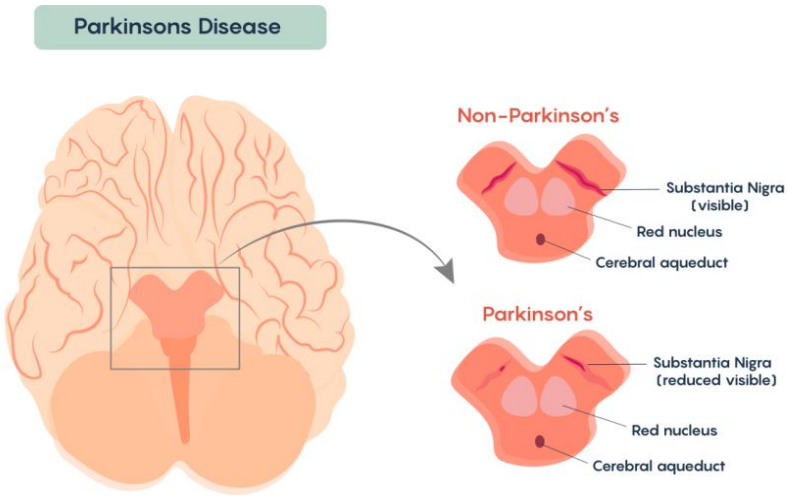

Figure 1: The loss of dopamine-generating cells in Substantia Nigra [9]

The researchers state that environmental factors like head injury, environmental toxin, drinking water, etc. are also the cause of this disease [6]. Table I shows the stages and severity level with Motor and Non-motor symptoms.

Stage	Severity	Main Motor Symptoms	Main Non-Motor Symptoms
1	Very Mild	Tremor on one side, slight stiffness, minor posture changes	Loss of smell, mild sleep problems, fatigue
2	Mild	Tremor on both sides, slow movement, muscle rigidity	Depression, constipation, soft speech
3	Moderate	Balance problems, slow walking, fall risk	Anxiety, mild cognitive issues, fatigue
4	Severe	Severe stiffness, very slow movement, difficulty standing	Memory problems, hallucinations
5	Very Severe	Cannot stand or walk, wheelchair/bed-bound	Advanced dementia, swallowing difficulty, bladder problems

Table I shows Stages, Severity and symptoms of PD

Tremor, slowness of movement, postural instability etc., are the frequent signs and symptoms of the PARKINSON'S DISEASE [7]. Some patients also experience the loss of memory and depression [8]. Parkinson's disease basic symptoms are not easy to diagnose in the primordial stage. But treatment on the advanced level does not control the progression of Parkinson's disease. This situation illustrates that early detection of Parkinson's disease that assists the humans to retain a quality of life. Figure.2 shows a) The Real image

of the Parkinson disease patient brain MRI versus humming bird and glory flower b) a real patient with Parkinson disease with the stages 1-5.

(a)The Real Real image of the Parkinson disease patient brain MRI versus humming bird and glory flower b) a real patient with Parkinson disease with the stages 1-5.

The existing research found that most of the Parkinson's disease patients have voiced disorder (dysphonia) [10] which can be identified by the vocal symptoms [11]. The effect of dysphonia on Parkinson's disease is investigated in [12] through a speech analysis approach. Later research, improved this approach through Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM) classifier to diagnose the various dysphonia [13]. Recent studies focus on vocal symptoms with a machine learning approach to detect the Parkinson's disease [11, 12].

The contribution of the proposed approach is

i. An attempt is made to develop an automatic and efficient Parkinson's disease detection model by using machine learning approaches.
ii. The present model compares the performance of Random Forest with other classifier on the publically available voice dataset. The experiment confirmed that Random Forest improves the classification accuracy on many performance evaluation parameters.

2. RELATED WORKS

The literature shows that various machine learning algorithm has been applied for Parkinson's disease classification based on the features vocal and gait [14]. Achraf Benba et al [15] utilizes the supervised learning classifiers for classifying neurological disorders and discriminate the patient affected by Parkinson's disease and other disorders. Sandhya Joshi et al [16] utilized the Neural Networks classifier for both Alzheimer's disease (AD) and Parkinson's disease detection whereas Carlos Castro et al [17] use to train multiple networks in order to vary the number of neurons for Parkinson's disease detection.

Salama A. Mostafa et al [18] implements the best performing machine learning algorithm Naïve Bayes (NB), Artificial Neural Network(ANN) and Decision tree (DT) on Parkinson's disease dataset and achieved 89.46%, 91.01% and 91.63% accuracy. Tamanna Sood and Padmavati Khandnor [19] applies the classification algorithm to enhance the Parkinson's disease diagnostic accuracy rates with the methods Random Forest, Decision Tree, K-Nearest Neighbors, including deep learning approaches like Convolutional Neural Network (CNN) architecture VGG16 and MobileNet achieved the highest accuracy on Random Forest.

Mandal and sairam [20] improves the accuracy of Parkinson's disease diagnosis with ranker search algorithm and Support Vector Machine for feature selection and classification respectively. This system identifies the Parkinson's disease and also the measures the severity. The model is evaluated on Parkinson's disease dataset.Shreerag Marar et al [21] present the Parkinson's disease detection system in the earlier stage with several machine learning approaches. The author uses the voice dataset from UCI repository and implements the RF, NB and ANN.

Haya Alaskar and Abir Hussain [22] introduce a detection model to identify the Parkinson's disease patient and healthy patients using data mining approaches. The approaches like SVM, Multilayer Neural Network (MNN) and Decision Tree are adopted for the classification. The dataset is collected from 29 individuals. This sample contains both Parkinson's disease patient's records and healthy individual's data. The MNN achieve the highest accuracy of 91.18% compared to other classifier.

3. CLASSIFICATION APPROACHES

The present study adopted the supervised machine learning algorithm to find the best classifier for the Parkinson's disease diagnosis. Algorithms like RF, NB, KNN, DT and SVM. The following section describes the classification algorithms.

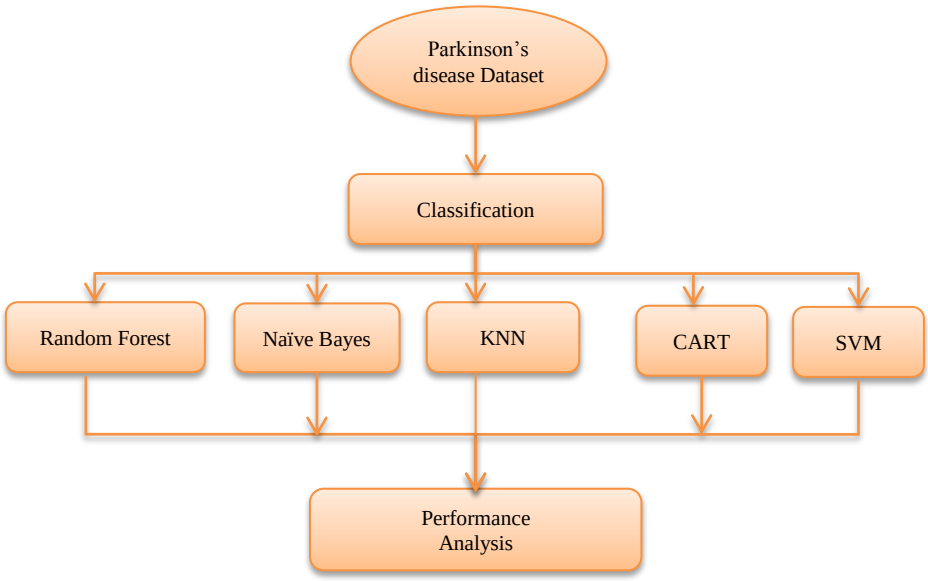

Figure 2. System Architecture

A. Random Forest

Random Forest is the tree based classifier which is the combination of many algorithms. Random forest builds a forest with multiple decision trees. Each tree in the forest is trained based on aggregation and a bootstrap model of the training data. Further a subset of attributes is selected arbitrarily at each
node of each decision tree. To classify the Parkinson's disease, decisions are taken by considering all the trees and select the final output by majority voting of the trees.

Algorithm 1: Random Forest

Input: Training set S=$\{x_1, y_1\}$,…,$\{x_n, y_1\}$,
Output**: Tree with majority voting**
Begin
$RandomForest(P, Q)$
$R \leftarrow \emptyset$
for $i \in 1, \cdots, B$ **do**
$P^{(i)} \leftarrow$ *A bootstrap sample from S*
$r_j \leftarrow Randomizedtreelearn(P^{(i)}, Q)$
$R \leftarrow R \cup (h_j)$
end for
return R
$Randomizedtreelearn(P^{(i)}, Q)$
At each node:
$f \leftarrow$ *small subset of Q*
split on best feature f
return learnedtree
end

B. Naïve Bayes

NB is the efficient classifier for disease prediction. NB belongs to probabilistic classifiers family, which applies the Bayes' theorem for classification. It is easy to implement and provide best result for large datasets. This method is applicable for real time problems [23].

The Bayes' theorem for the Parkinson's disease classification is computed as expressed in equation (1)

$$P(c|x) = \frac{p(x|c)p(c)}{p(x)} \quad (1)$$

Where $p(c)$ and $p(x)$ symbolizes the previous probability of class c and attribute x . $P(c|x)$ denotes the posterior probability and $p(x|c)$ probability of attribute x on class c. The $P(c|x)$ for the independent features $X = \{x_1, x_2, \cdots x_n\}$ are calculated by

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c) \quad (2)$$

In addition, NB provides the best result for multiclass problems.

C. K- Nearest Neighbor

K- Nearest Neighbor is a non-parametric supervised method of learning that can be applied in regression and also in a classification problem. K-NN is applied in classifying Parkinson diseases in this study. K-NN algorithm stores the available data and classifies a new data point depending on the similarity. This implies that when new data is introduced, this data can easily be classified into a well suite category as a result of taking a K - NN algorithm. This procedure offers result depending on the similar observations present in proximity. The proximity is determined by locating the Ed i.e.

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{k}(p_i - q_i)^2} \quad (3)$$

The performance of the present model is evaluated on the Parkinson's disease dataset. The above-mentioned process is repeated and computes performance measures for each k value.

- **Step-1:** select the K value
- **Step-2:** compute the Euclidean distance **as in eqn3**
- **Step-3:** choose the K nearest neighbors as per step 2.
- **Step-4:** with k neighbors, compute the total data points in each category.
- **Step-5:** Allocate the novel data points to that category for which the number of the neighbor is high.

D. Classification and Regression Trees (CART)

CART is a tree based classification developed in the year 1984 to address regression as well as classification problem [24]. CART is an empirical classification technique which uses the final goal, to find the classification and prediction rules. In XY dataset (X, Y), Y is the explained variable, and X = (X 1, X 2, X 3, etc.) is not a random number, but a p-variable that characterizes entities. This method aims at predicting Y values of Xi variables, i x 1,.,p. Both Y and X i can be quantitative thus offering a lot of flexibility to CART since it could be applied in different settings.

E. Support Vector Machine (SVM)

SVM is the powerful linear classification technique that follows the Structural Risk Minimization (SRM) principle. SVM has generalization capability to overcome the problems like over-fitting and Neural Network learning. This method ensures that computational complexity does not increase while mapping the nonlinear problem with high-dimensional space. It also solves the curse of dimensionality issues through kernel function.

For the given data SVM can be summed up in order to resolve the quadratic programming problem.

$$S = \{(p_1, q_1), \cdots, (p_1, q_1)\}, p \epsilon Rn, q \epsilon \{+1, -1\} \quad (4)$$

Consider ξ_i be the relaxation variable, then the formula for the problem is,

$$\Phi(\omega, \xi_i) = \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i \quad (5)$$

F. Dataset

The Parkinson's Data Set is collected from UCI repository which is developed by Max little by the University of Oxford. This dataset contains a variety of biomedical voice measurements collected from 31 individuals, where 23 people have Parkinson's disease. The columns denote the voice measure, 195 voice recording of 35 individuals of people. The goal of the data is to differentiate the healthy people and people with Parkinson

disease in terms of status column (0-healthy, 1- Parkinson disease). Table 2 shows the attributes contained in the voice dataset.

Table-2 Attributes details of the Voice dataset

Attributes	Description
Name	Recording Name
MDVP:Fo(Hz)	Average vocal fundamental frequency
MDVP:Fhi(Hz)	Maximum vocal fundamental frequency
MDVP:Flo(Hz)	Minimum vocal fundamental frequency
MDVP:Jitter(%),MDVP:Jitter(Abs),MDVP:RAP, MDVP:PPQ,Jitter:DDP	Measures of variation in fundamental frequency
MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3, Shimmer:APQ5,MDVP:APQ,Shimmer:DDA	Measures of variation in amplitude
NHR,HNR	Ratio of noise to tonal components in the voice
Status	Parkinson's Healthy
RPDE,D2	Nonlinear dynamical complexity measures
DFA	Signal fractal scaling exponent
spread1,spread2,PPE	Nonlinear measures of fundamental frequency variation

4. EXPERIMENTAL RESULT

The efficacy of the present classification model is evaluated in this section. The implementation was carried out with python platform with Scikit-learn tool for utilizing the Machine Learning algorithms. It is the powerful tool in python programming language for predictive data analysis. The five algorithms such as RF, NB, KNN, CART, and SVM are adopted for Parkinson's disease classification.

The total class distribution details of voice dataset are shown in figure 3. From the graph it is clear that the dataset contains more details on Parkinson patient data compared to healthy patients. This assists the classification model to train on these information and improves the classification accuracy.

The result of various performance metrics of applied machine learning algorithms are listed in table 3. For Parkinson's disease dataset SVM, Random Forest, KNN, CART and Naïve Bayes prove to be better in terms of accuracy, precision, recall, F1-score and RMSE. The graphical representation of these result is shown in different format for better understanding of the performance.

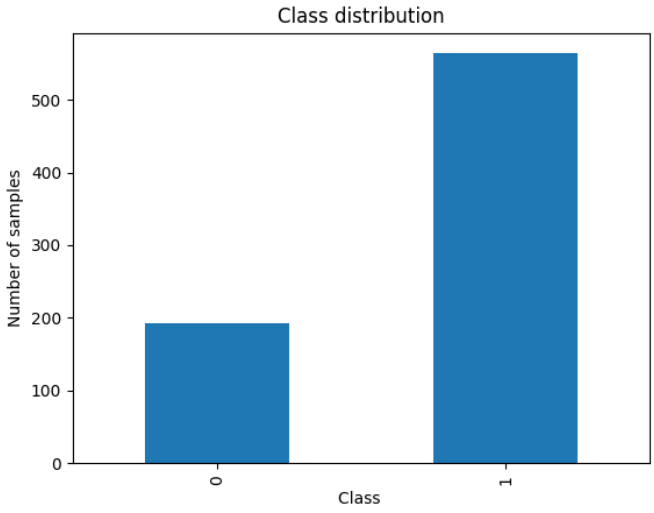

Figure 3. Class distribution of voice dataset

Table 3: Performance measures of applied ML algorithms

Methods	Precision	Recall	F1-score	Accuracy	RMSE
Random Forest	90.46	90.78	90.40	90.78	0.3
SVM	63.36	79.60	70.56	79.6	0.45
KNN	70.85	69.07	69.90	69.07	0.55
NB	78.37	76.97	77.58	76.97	0.47
CART	82.87	80.26	81.20	80.26	0.44

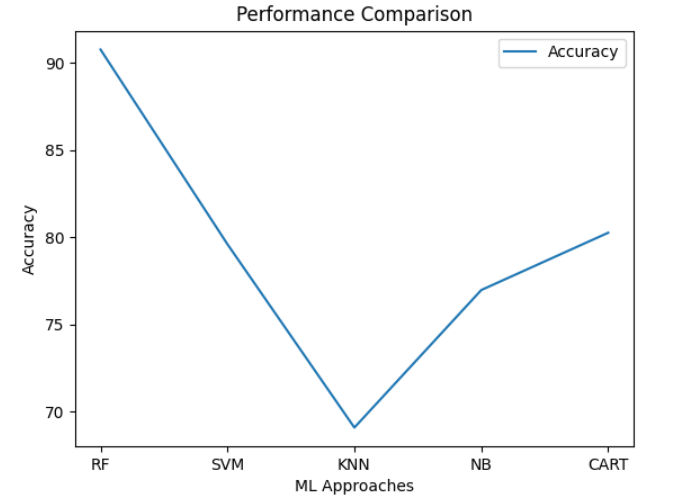

Figure 4. Performance comparison of Accuracy

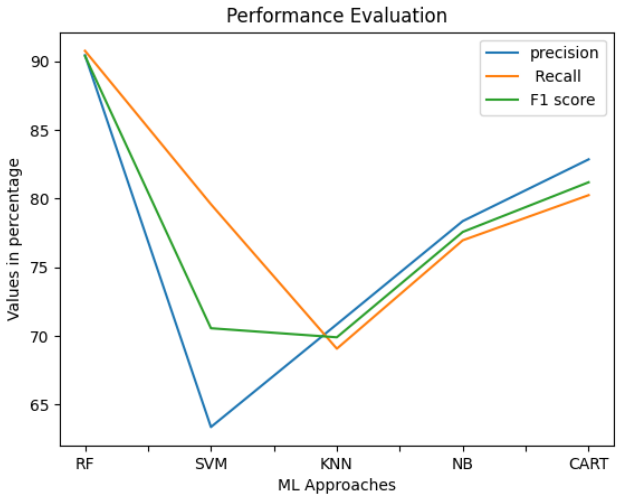

Figure 5 Performance comparison of machine learning approaches

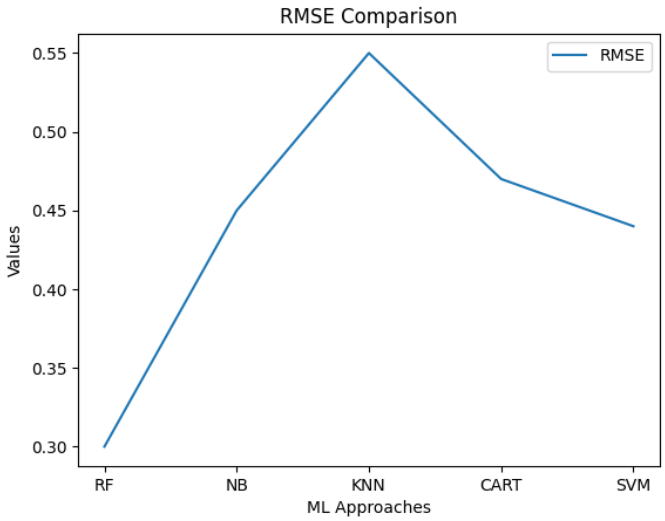

Figure 6. Performance comparison of RMSE

The diagram shows the accuracy of the present model on Parkinson's disease voice dataset. The RF achieves 90.78%, Nb obtains 76.97%, Knn achieve 69.7%, CART obtains 80.26% and SVM score 79.6%. From the plot it is clear that random forest outperformed the other classifier that is examined in this study. Random forest achieves the great accuracy compared to others.

Figure 5 and 6 illustrate the performance of other performance metrics such as precision, recall, fmeasure and RMSE value. The Random forest classifier achieved higher precision recall and fmeasure score and less error rate 0.30 compared to other methods. The efficacy of the system is evaluated with maximum performance metrics and showed in the result section.

5. CONCLUSION

The present paper examines the best performing five machine algorithm on Parkinson's disease dataset to identify the suitable classifier for diagnosing the Parkinson's disease. From the study it is clear that Machine Learning techniques provide greater support for the physicians in diagnosing the Parkinson's disease. This study achieved up to 90.78% accuracy. However to improve the performance of the investigated algorithm, there is a necessity for future improvement in this study. In future the classification accuracy has to be improved through feature selection technique and hybrid classification algorithms.

REFERENCES

[1]. Davie, Charles Anthony. "A review of Parkinson's disease." *British medical bulletin* 86, no. 1 (2008): 109-127.

[2]. D. B. Calne, "Is idiopathic parkinsonism the consequence of an event or a process," *Neurology*, Vol. 44, no. 15, pp. 5–5, 1994.

[3]. Sveinbjornsdottir, Sigurlaug. "The clinical symptoms of Parkinson's disease." *Journal of neurochemistry* 139 (2016): 318-324.

[4]. Ramayya, Ashwin G., Amrit Misra, Gordon H. Baltuch, and Michael J. Kahana. "Microstimulation of the human substantia nigra alters reinforcement learning." *Journal of Neuroscience* 34, no. 20 (2014): 6887-6895.

[5]. Kaya, Ersin, Oguz Findik, Ismail Babaoglu, and Ahmet Arslan. "Effect of discretization method on the diagnosis of Parkinson's disease." *Int. J. Innov. Comput. Inf* 7 (2011): 4669-4678.

[6]. De Lau, Lonneke ML, and Monique MB Breteler. "Epidemiology of Parkinson's disease." *The Lancet Neurology* 5, no. 6 (2006): 525-535.

[7]. Lang, Muriel, Franz MJ Pfister, Jakob Fröhner, Kian Abedinpour, Daniel Pichler, Urban Fietzek, Terry Taewoong Um, Dana Kulić, Satoshi Endo, and Sandra Hirche. "A Multi-Layer Gaussian Process for Motor Symptom Estimation in People with Parkinson's Disease." *IEEE Transactions on Biomedical Engineering* 66, no. 11 (2019): 3038-3049.

[8]. Pfeiffer, Ronald F. "Non-motor symptoms in Parkinson's disease." *Parkinsonism & related disorders* 22 (2016): S119-S122.

[9]. Dr William Ju, "Neuroscience: Canadian 1st Edition", unit 2: Chronic Neurodegenerative Diseases, Creative Commons Attribution 4.0 International License, published using Pressbooks.

[10]. Little, Max, Patrick McSharry, Eric Hunter, Jennifer Spielman, and Lorraine Ramig. "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease." Nature Precedings (2008): 1-1.

[11]. Das, Resul. "A comparison of multiple classification methods for diagnosis of Parkinson disease." *Expert Systems with Applications* 37, no. 2 (2010): 1568-1572.

[12]. Little, Max A., Patrick E. McSharry, Stephen J. Roberts, Declan AE Costello, and Irene M. Moroz. "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection." Biomedical engineering online 6, no. 1 (2007): 23.

[13]. Arjmandi, Meisam Khalil, and Mohammad Pooyan. "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine." *Biomedical Signal Processing and Control* 7, no. 1 (2012): 3-19.

[14]. Nandy, Anup. "Statistical methods for analysis of Parkinson's disease gait pattern and classification." Multimedia Tools and Applications 78, no. 14 (2019): 19697-19734.

[15]. Benba, Achraf, Abdelilah Jilbab, and Ahmed Hammouch. "Discriminating between patients with Parkinson's and neurological diseases using cepstral analysis." *IEEE transactions on neural systems and rehabilitation engineering* 24, no. 10 (2016): 1100-1108.

[16]. Joshi, Sandhya, Deepa Shenoy, Vibhudendra Simha GG, P. L. Rrashmi, K. R. Venugopal, and L. M. Patnaik. "Classification of Alzheimer's disease and Parkinson's disease by using machine learning and neural network methods." In *2010 Second International Conference on Machine Learning and Computing*, pp. 218-222. IEEE, 2010.

[17]. Castro, Carlos, Eunice Vargas-Viveros, Alejandro Sánchez, Everardo Gutiérrez-López, and Dora-Luz Flores. "Parkinson's disease Classification Using Artificial Neural Networks." In *Latin American Conference on Biomedical Engineering*, pp. 1060-1065. Springer, Cham, 2019.

[18]. Mostafa, Salama A., Aida Mustapha, Shihab Hamad Khaleefah, Mohd Sharifuddin Ahmad, and Mazin Abed Mohammed. "Evaluating the performance of three classification methods in diagnosis of Parkinson's disease." In *International Conference on Soft Computing and Data Mining*, pp. 43-52. Springer, Cham, 2018.

[19]. Sood, Tamanna, and Padmavati Khandnor. "Classification of Parkinson's disease Using Various Machine Learning Techniques." In *International Conference on Advances in Computing and Data Sciences*, pp. 296-311. Springer, Singapore, 2019.

[20]. Mandal, Indrajit, and N. Sairam. "Accurate telemonitoring of Parkinson's disease diagnosis using robust inference system." *International journal of medical informatics* 82, no. 5 (2013): 359-377.

[21]. Marar, Shreerag, Debabrata Swain, Vivek Hiwarkar, Nikhil Motwani, and Akshar Awari. "Predicting the occurrence of Parkinson's disease using various Classification Models." In *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*, pp. 1-5. IEEE, 2018.

[22]. Alaskar, Haya, and Abir Hussain. "Prediction of Parkinson disease using gait signals." In *2018 11th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 23-26. IEEE, 2018.

[23]. Dwivedi, Ashok Kumar. "Performance evaluation of different machine learning techniques for prediction of heart disease." *Neural Computing and Applications* 29, no. 10 (2018): 685-693.

[24]. Sathyadevi, G. "Application of CART algorithm in hepatitis disease diagnosis." In *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*, pp. 1283-1287. IEEE, 2011.

