

Parkinson's Disease Detection Using Machine Learning Techniques



P. Anudeep, P. Mourya, and T. Anandhi

Abstract Parkinson disease (PD) is a progressive neuro degenerative disorder that impacts more than 6 Mio. People around the world. Nonetheless, non-specialist physicians still do not have a definitive test for PD, similarly in the early stage of the diseased person where the signs may be intermittent and badly characterized. It resulted in a high rate of misdiagnosis (up to 25% among non-specialists) and many years before treatment, patients can have the disorder. A more accurate, unbiased means of early detection is required, preferably one that individuals can use in their home setting. The proposed system for predictive analytics is a mixture of clustering of K-means and a decision tree used to gain insights from patients. The problem can be addressed with reduced error rate with the application of machine learning techniques. Our proposed system also produces accurate results by combining the spiral drawing inputs of patients impacted by common and Parkinson's. From these drawings, the principal component analysis algorithm (PCA) for extraction of the feature from the spiral drawings and support vector machine is used for classification. UCI machine learning platform voice data collection in Parkinson's disease is used as feedback. Thus, our study results will show early detection of the disorder can promote the therapeutic care of the elderly and increase the chances of their life span and healthier lifestyle living peaceful life.

Keywords Parkinson's disease · PCA · K-means clustering · Decision tree · Machine learning · Data mining

P. Anudeep (✉) · P. Mourya · T. Anandhi
Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, India
e-mail: anudeepasala1998@gmail.com

P. Mourya
e-mail: mouryasagar3810@gmail.com

T. Anandhi
e-mail: anandhitamil2908@gmail.com

1 Introduction

Parkinson's disease is a disorderliness in motion. This effect the nervous organization, and with time, signs become worse. Cerebral palsy, ataxia, and Tourette syndrome are other movement's conditions [1]. We come about when a disturbance in the nervous system influences the capacity of a person to shift or remain still [2, 3]. The National Institutes of Health (NIH) reports that about 60,000 people are identified with Parkinson's disease every year in the United States of America, and about half of the million people are living with the condition [4, 5]. Some signs grow with time, and certain patients may have dementia. Most of the symptoms arise from a reduction of brain dopamine levels [6, 7]. One study, located in France, showed in 2018 that men are 50% more prone to have Parkinson's disease than women in general, but women's risk tends to increase. Symptoms appear in most adults at or above age 60 [8, 9]. They occur sooner in 5–10% of instances, though [10, 11]. If Parkinson's disease progresses before age 50, this is considered Parkinson's disease with "early onset."

The paper is proceeded as follow, Sect. 2 describes the related works; Sect. 3 presents the proposed system; Sect. 4 covers the experimentation results and analysis; Sect. 5 expresses the conclusion [12, 13].

2 Related Work

- Aaswad Sawant et al. [14] studied on the various cancer detection strategies. The device can be used by surgeons and radiologists as a second decision for fast and effective identification of brain tumors.
- Gamal Saad Mohamed et al. [15] presented four forms of classification focused on Naive Bayes, SVM, MLP neural network, and decision trees are used in this paper to identify the PD dataset, and the output of these classification is analyzed when applied on the real PD dataset, distinct PD dataset, and chosen collection of PD dataset attributes. The data set used in this study includes a variety of speech signals from 32 people: 25 with PD and 9 healthy individuals.
- Enes Celik et al. [16, 17], in this analysis, to model Parkinson's disease, similar classification methodologies including logistic regressions, support vector machines, random trees, gradient boostings, and random forest are related. A total of 1200 speech data sets were used in the classification stage, comprising of 26 characteristics gathered from Parkinson's diseased patients and non-patients. Thanks to correlation maps, the features space of the dataset is extended. Such correlation of maps are developed with the features that are collected using the principal component analysis (PCA), information gain (IG), and all features, respectively [18, 19].

Monica Giuliano et al. [20, 21] proposed demographic details, and vocal phonation records/a/ from the accessible mPower database were examined in this study in order

to classify patients with PD. Then, a parsimonious model was identified that achieved a reduction from 62 to 5 characteristics of the phonation, which were considered in addition to sex and age. Neural networks multilayer perceptron (MLP) and logistic regression (LR) were used to achieve a model with strong predictive potential (area below the Receiver operating characteristic's curve, AUC-ROC, over 0.82) [22, 23]. This research leads to the tracking of patients with EP by capturing a few phoning information obtained through a mobile phone [24, 25].

3 Proposed System

We suggest a model and accurate results analyzing data from patients with both speech and spiral painting. Thus, the doctor will infer normality or deviation by comparing both the findings and recommending the drug dependent on the stage affected.

3.1 Voice Data Processing

UCI platform opens the audio file. We used RStudio for analysis of the data. The conceptual architecture for predictive analytics is a variation of the K-means clustering and the decision tree classification method that is used to obtain patient insights. The problem can be solved with reduced error rate by using the machine learning algorithms. The speech dataset of Parkinson's disorder from the UCI machine learning library is used as feedback. While our experimental results show early disease, diagnosis can promote therapeutic care of the elderly and increase the chances of their life span and better lifestyle contributing to peaceful life.

3.2 Spiral Drawing Analysis

We used PyCharm-based python language for data analyzes to process the spiral images. Our proposed system provides reliable results by combining spiral extracting feedback from the patients impacted by usual and Parkinson. From these drawings, the principal component analysis algorithm (PCA) for extraction of the function from the spiral drawings. From the sketches of the spiral: X; Y; Z; Pressure; GripAngle; Timestamp; values of the reference ID are removed. Using machine learning technique (Support vector machine), the extracted values are compared to the trained database and results are obtained (Fig. 1).

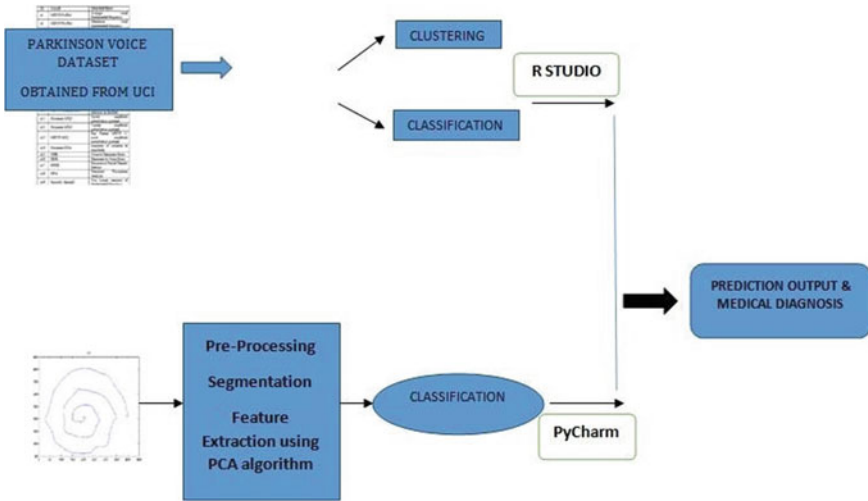


Fig. 1 Overall proposed

3.2.1 Parkinson’s Disease Voice Dataset Analysis

1. Importing data into RStudio

Step 1: Input and arrange the data in Excel.

Arrange the data in an Excel worksheet, so that the first row (Row 1) includes the column names and each following row contains all the information necessary for each data point in the experiment (i.e., Rates of description and measurements).

Step 2: Save your worksheet as a comma-separated file type (.csv).

Save your Excel spreadsheet as usual (default form of file: Excel Workbook); this will be your master file you can always revert to change stuff, add new details, etc. Then, press “Save As...” to create a version of your data to enter in R. A window should open where you can define the filename you want, as well as the sort of file you want.

Step 3: Import data to RStudio.

2. K-Means Clustering

K-means clustering is an unsupervised machine learning algorithm that attempts to clustering data based on similarity between them. Unsupervised machine learning means that no outcomes can be predicted, and the algorithm is simply trying to find patterns in the data. In k-means clustering, the number of clusters we want to divide the data is defined. The algorithm allocates each discovery randomly to a cluster, and determines the centroid of each cluster. The algorithm then iterates through two steps:

- Reassigns data point to the cluster which is closest to the centroid.
- Calculate each cluster with centroid.

Such two measures are replicated until no further reduction of the variability within the cluster is feasible. The variance within the cluster is measured as a percentage of the Euclidean interval between the data points and the centroids concerned.

3.3 Decision Tree

A decision tree is also considered as prediction tree. A decision tree provides a framework to define judgment and result sequences. The aim is to predict an answer or output variable Y , provided the input $X = \{X_1, X_2, \dots, X_n\}$. An input variable is named for every $\{X_1, X_2, \dots, X_n\}$ part of the group. Creating a decision tree with test points and divisions will accomplish the predictions. At each check point, a decision may be taken to pick a single branch and navigate the decision trees in a number of disciplines, such as: on the basis of individual attributes determining whether or not to give a loan to an individual, predicting the rate of return to various investing strategies, predicting whether or not to deliver a direct mail to a prospective client, etc.

A decision tree consists of node, and thus, contains a rooted tree, which implies it is a guided tree with a core node. Root nodes does not have incoming edges, with all other nodes in a decision tree have exactly one incoming edges. An internal node is the node with an incoming edge and outgoing sides. Also known as the check node is an internal node. Nodes with no outgoing edges are classified as terminal nodes or leaves.

3.3.1 Parkinson's Disease Spiral Drawing Analysis

1. Preprocessing

(a) Image Acquisition

Image acquisition is the principal phase of image processing. When opposed to HD images, the videos are obtained with minimal noise. Each package has the key advantages of having images of better clarity, low noise, and distortion.

(b) Image Preprocessing

Image pre-processing is one of the image processing types, an attempt to make identification more evident. Pre-processing of photographs is a way to improve image quality, so that the resulting image becomes better than original. The median filter is a non-linear method, whereas linear is the typical filter. Mean filtering is a fast, intuitive, and quick to apply smoothing images, that is to say reducing the amount of difference in strength between one pixel and the next.

The median filter is usually used in a picture to reduce the salt-and-pepper noise. It also does a better job of maintaining valuable information in the picture than the mean filter. The median is determined by first sorting all the pixel values in numerical order from the surrounding neighborhood, and then, replacing the pixel considered with the center pixel. If there are even numbers of pixels in the area under scrutiny, the sum of the two center pixel values is used. For the reduction of noise, both mean and median filters are used. This pre-processing image is used as the input for image segmentation.

(c) Image Segmentation

The segmentation of images is an important method for most subsequent tasks of image analysis. Segmentation divide an image into its region or artifacts which make up it. The aim of segmentation is to render the portrayal of an image clearer or more readily analyzable in something that is more relevant.

2. Prediction

So, our hybrid model, combining image processing (spiral drawing analysis) utilizing image processing methodology and data analytics (values derived from speech dataset and spiral drawings) using R technology. Data analytics have a larger role to play in healthcare sectors, as these data are diverse and complex in nature, and the Parkinson disease dataset is large in scale, and new opportunities and demands are found, greater complexity is revealed, predictive capacity is improved, and time is productive to adapt to cost-effective measures.

More specifically, that integration helps healthcare organizations to quickly and efficiently evaluate their large data sets. Early detection of any type of disease is an important factor and this results in advance treatment of patients. This system detects the highest classifier precision, and multi-classifier consensus tests are taken to identify the disease sooner and increase PD people's lifespan.

4 Experimental Results

- In our work, we used UCI machine learning repository. The experiments are performed on R studio. The studio consists of enormous quantities of multidimensional details, that are gathered up within different areas including advertising, geo-spatial, and bio-medical areas. And with the help of python, I am generating a pressure graph for both the diseased and un-diseased. With these graphical representation of the data, we can get an idea of how differently both the diseased and non-diseased persons are identified and can be used for easily identification of the person those are effected (Fig. 2).

When this image is passed, the pre-processing takes place here all the features are extracted for further segmentation (Fig. 3).

The features are then trained with a decision tree classifier which is widely used for all classification and regression techniques. Then, as the next step, the RGB

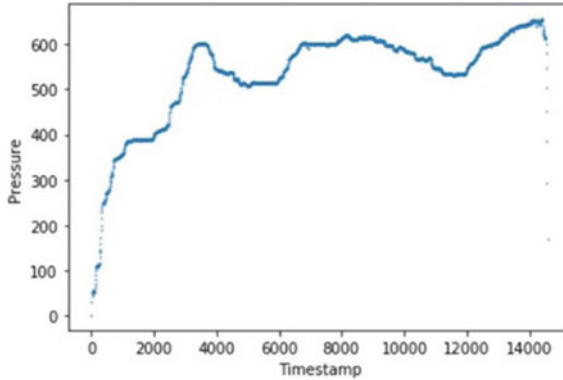


Fig. 2 The pressure graph for diseased person

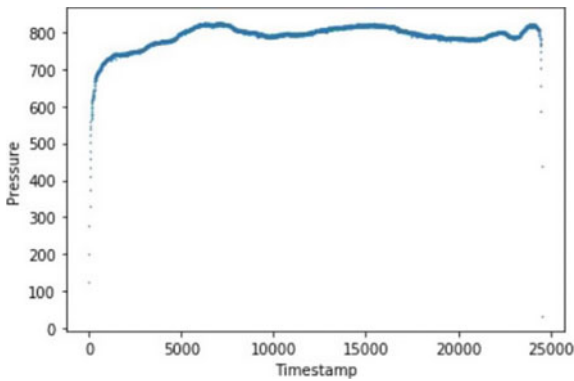


Fig. 3 The pressure graph for un-diseased person

components are extracted from the image and the analysis is done on the number of clusters obtained. Some images are kept for testing and training. The classifier learns [11] the features and successfully classifies when any new image is given to it. It identifies where the image is benign or malignant (Fig. 4).

- Figure 2 shows the efficiency of the proposed model. The accuracy of the classifier is observed with 86.66% whereas 99.9% specificity and 80.48% sensitivity. The segmentation of the prohibited item is extracted from the exact image which gives us various parameters for measuring such as its intensity, volume, and size. This helps in diagnosing and treating the disease more efficiently. In Fig. 3, the transmission speed of the system is depicted (Figs. 5 and 6).

Logistic regression is the technique that also used in the prediction of the diseased persons value as the reference and predict the approximate value for the diseased person (Fig. 7).

```

> summary(predict_seen)
 0  1
 9 21
> table_pred <- table(test_tree$status, predict_seen)
> table_pred
  predict_seen
    0  1
 0  7  2
 1  2 19
> accuracy_Test <- sum(diag(table_pred)) / sum(table_pred)
> print(accuracy_Test)
[1] 0.8666667

```

Fig. 4 Accuracy output

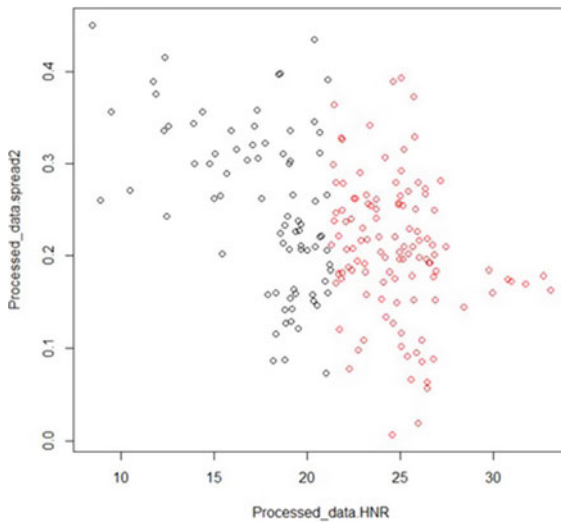


Fig. 5 Scatter plot matrix of Parkinson's data

```

Logistic Regression

In [37]: clf=LogisticRegression()
         clf.fit(train_x, train_y)
         preds=clf.predict(test_x)
         print('accuracy:',accuracy(test_y.tolist(), preds.tolist()), '%')
         print(metrics(test_y.tolist(), preds.tolist()))

accuracy: 70.0 %
{'Precision': 0.5, 'Recall': 0.625, 'F1': 0.5555555555555556}

C:\anaconda\data\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)

```

Fig. 6 Accuracy of logistic regression

Support Vector Machine

```
In [39]: clf=SVC()
         clf.fit(train_x, train_y)
         preds=clf.predict(test_x)
         print('accuracy:',accuracy(test_y.tolist(), preds.tolist()), '%')
         print(metrics(test_y.tolist(), preds.tolist()))

accuracy: 100.0 %
{'Precision': 0.5, 'Recall': 1.0, 'F1': 0.6666666666666666}

C:\anaconda\data\lib\site-packages\sklearn\svm\base.py:193: FutureWarning: The default value of gamma will change from 'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' or 'scale' to avoid this warning.
  "avoid this warning.", FutureWarning)
```

Fig. 7 Accuracy of support vector machine

Decision Tree

```
In [40]: clf=DecisionTreeClassifier()
         clf.fit(train_x, train_y)
         preds=clf.predict(test_x)
         print('accuracy:',accuracy(test_y.tolist(), preds.tolist()), '%')
         print(metrics(test_y.tolist(), preds.tolist()))

accuracy: 100.0 %
{'Precision': 0.5, 'Recall': 1.0, 'F1': 0.6666666666666666}
```

Fig. 8 Accuracy of decision tree

Support vector machines are also helpful in predicting because in these algorithm we generally identify the nearest points of the trained data and based on that we test the data (Fig. 8).

A decision tree is a tree like structure in which one node be the test data and other node will be trained data, and based on the nodes data, we will predict the outcomes (Fig. 9).

K-nearest neighbors, decision tree, SVM, k-means clustering we are going to get the accuracy of different algorithms and will help of all this algorithm we can identify the diseased person with some more accuracy and help the patient to move to the next process.

K-Nearest Neighbors

```
In [41]: clf=KNeighborsClassifier()
         clf.fit(train_x, train_y)
         preds=clf.predict(test_x)
         print('accuracy:',accuracy(test_y.tolist(), preds.tolist()), '%')
         print(metrics(test_y.tolist(), preds.tolist()))

accuracy: 60.0 %
{'Precision': 0.4, 'Recall': 0.5714285714285714, 'F1': 0.47058823529411764}
```

Fig. 9 K-nearest neighbor

5 Conclusion

Past analysis papers provide a comprehension survey for specific modalities in neuro imagination and related analytical techniques proposed for treatment of Parkinson's disease in recent years. Past research articles focused solely on a specific imaging modularity such as MRI or PET, or only on a particular type of dementia such as AD. This study sought to cover the wider range of imaging and machine learning algorithms for diagnosing of mental illness so that field researchers could readily identify the state of the arts in the area. We also emphasize the importance of early detection and prediction of Parkinson's disease, so that patients can be given treatment and support as soon as possible.

References

1. Surendran, R., & Keerthi Samhitha, B. (2014). Energy aware grid resource allocation by using a novel negotiation model. *Journal of Theoretical and Applied Information Technology*.
2. Narmadha, D., & Pravin, A. (2020). An intelligent computer-aided approach for target protein prediction in infectious diseases. *Soft Computing*, 1–14.
3. Jacob, T. P. (2015). Implementation of randomized test pattern generation strategy. *Journal of Theoretical and Applied Information Technology*, 73(1).
4. Mana, S. C., Keerthi Samhitha, B., Jose, J., Swaroop, M. V., & Reddy, P. C. K. (2019). Traffic violation detection using principal component analysis and viola jones algorithms. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(3). ISSN: 2277-3878.
5. Thyagarajan, K. K., & Kalaiarasi, G. (2018). Pulse coupled neural network based near-duplicate detection of images (PCNN–NDD). *Advances in Electrical and Computer Engineering*, 18(3), 87–97.
6. Ramya, D., Antony, A. S. M., Immanuel, D. G., & Nagarajan, G. (2019). Design of interleaved flyback converter. *International Journal of Intelligent Enterprise*, 6(1), 59–76.
7. Selvan, M. P., & Sekar, A. C. (2016, December). Ranking scientific journals based on research author's profile. In *2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)* (pp. 652–654). IEEE.
8. Prasad, K. M., & Dhar, P. (2019, October). Industrial automation with bidirectional visitor counter. In *IOP Conference Series: Materials Science and Engineering* (Vol. 590, No. 1, p. 012012). IOP Publishing.
9. Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019, April). Design and implementing heart disease prediction using Naives Bayesian. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 292–297). IEEE.
10. Divya, S., Vignesh, R., & Revathy, R. (2019). A distinctive model to classify tumor using random forest classifier. In *2019 Third International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, India (pp. 44–47).
11. Nagarajan, G., & Minu, R. I. (2015). Fuzzy ontology based multi-modal semantic information retrieval. *Procedia Computer Science*, 48, 101–106.
12. Venkatakrishna, D., & Ankeyarkanni, B. (2016). A description of content based image retrieval using from block truncation coding and image content description. *Indian Journal of Science and Technology*, 9(21), 1–5.
13. Mala, G. A. (2013, December). Tracking international migration from RFID data using map reduce method. In *2013 Fifth International Conference on Advanced Computing (ICoAC)* (pp. 484–487). IEEE.

14. Sarraf, S., DeSouza, D. D., Anderson, J., & Tofighi, G. (2016). DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. Cold Spring Harbor Laboratory Press.
15. Liu, F., & Shen, C. (2014). Learning deep convolutional features for MRI based Alzheimer's disease classification.
16. Payan, A., & Montana, G. (2015). Predicting Alzheimer's disease: A neuroimaging study with 3D convolutional neural networks.
17. Olanrewaju, R. F., Sahari, N. S., Musa, A. A., & Hakiem, N. (2014) Application of neural networks in early detection and diagnosis of Parkinson's disease. *International Conference on Cyber and IT Service Management*.
18. Shyry, S. P. (2014, July). Performance measurement in selfish overlay network by fuzzy logic deployment of overlay nodes. In *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)* (pp. 717–721). IEEE.
19. Ponraj, A. (2019). Optimistic virtual machine placement in cloud data centers using queuing approach. *Future Generation Computer Systems*, *93*, 338–344.
20. Hadjahamadi, A. H., & Askari, T. J. (2012). A detection support system for Parkinson's disease diagnosis using classification and regression tree. *Journal of Mathematics and Computer Science*, *4*, 257–263.
21. Alemami, Y., & Almazaydeh, L. (2014) Detecting of Parkinson disease through voice signal features. *Journal of American Science*.
22. Little, M. A., McSharry, P. E., Hunter, E. J., & Ramig, L. O. (2008). Suitability of Dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, *56*, 1015–1022.
23. Muhlenbach, F., & Rakotomalala, R. (2015). Discretization of continuous attributes. In J. Wang (Ed.), *Encyclopedia of data warehousing and mining, idea group reference* (pp. 397–402).
24. Yogitha, R., & Mathivanan, G. (2018, April). Performance analysis of transfer functions in an artificial neural network. In *2018 International Conference on Communication and Signal Processing (ICCSP)* (pp. 0393–0397). IEEE.
25. Sangeetha, K., Vishnuraja, P., & Deepa, D. (2016). Stable clustered topology and secured routing using mobile agents in mobile ad hoc networks. *Asian Journal of Information Technology*, *15*(23), 4806–4811.