GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

DIAGNOSING PARKINSON'S DISEASE USING RANDOM FOREST TECHNIQUE IN MACHINE LEARNING

A Report for the Evaluation 3 of Project 2

Submitted by

Arpit Banerji

(1613105023)

in partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING WITH SPECIALIZATION OF CLOUD COMPUTING AND VIRTUALIZATION

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

Mr. Lalit Sharma
Astt. Professor

APRIL / MAY- 2020

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report "**DIAGNOSING PARKINSON'S DISEASE USING RANDOM FOREST TECHNIQUE IN MACHINE LEARNING**" is the bonafide work of "**ARPIT BANERJI** (1613105023)" who carried out the project work under my supervision.

SIGNATURE OF HEAD

Dr. MUNISH SHABARWAL, PhD (Management), PhD (CS) Professor & Dean,
School of Computing Science & Engineering

SIGNATURE OF SUPERVISOR

Mr. Lalit Sharma
Astt. Professor
School of Computing Science & Engineerin

TABLE OF CONTENTS

ABSTRACT

Parkinson’s Disease (PD) is a kind of degenerative neurological disorder leading to decrease in dopamine levels inside the brain.

There is no such test, present, to diagnose Parkinson’s Disease. It’s often a very difficult task to diagnose PD, mainly in the early stages when the effects on the body are not yet severe. Observing the advancement of the disease over time requires continuous clinical visits by the patient. In this paper, we have used the power of Machine Learning and Deep Neural Network to build a model for the detection of the disease and also ensemble techniques to improve the prediction accuracy. Also, model is validated using different metrics that are present in like confusion metric and accuracy score.

Keywords: PD, Deep Neural Networks, Machine Learning.

INTRODUCTION

Parkinson's disease is a brain disorder that leads to shaking, stiffness, and difficulty with walking, balance, and coordination.

Parkinson's symptoms usually begin gradually and get worse over time. As the disease progresses, people may have difficulty walking and talking. They may also have mental and behavioural changes, sleep problems, depression, memory difficulties, and fatigue.

Both men and women can have Parkinson’s disease. However, the disease affects about 50 percent more men than women.

One clear risk factor for Parkinson's is age. Although most people with Parkinson’s first develop the disease at about age 60, about 5 to 10 percent of people with Parkinson's have "early-onset" disease, which begins before the age of 50. Early-onset forms of Parkinson's are often, but not always, inherited, and some forms have been linked to specific gene mutations.

What Causes Parkinson's Disease?

Parkinson's disease occurs when nerve cells, or neurons, in an area of the brain that controls movement become impaired and/or die. Normally, these neurons produce an important brain chemical known as dopamine. When the neurons die or become impaired, they produce less dopamine, which causes the movement problems of Parkinson's. Scientists still do not know what causes cells that produce dopamine to die.

People with Parkinson's also lose the nerve endings that produce norepinephrine, the main chemical messenger of the sympathetic nervous system, which controls many automatic functions of the body, such as heart rate and blood pressure. The loss of norepinephrine might help explain some of the non-movement features of Parkinson's, such as fatigue, irregular blood pressure, decreased movement of food through the digestive tract, and sudden drop in blood pressure when a person stands up from a sitting or lying-down position.

Many brain cells of people with Parkinson's contain Lewy bodies, unusual clumps of the protein alpha-synuclein. Scientists are trying to better understand the normal and abnormal functions of alpha-synuclein and its relationship to genetic mutations that impact Parkinson's disease and Lewy body dementia.

Although some cases of Parkinson's appear to be hereditary, and a few can be traced to specific genetic mutations, in most cases the disease occurs randomly and does not seem to run in families. Many researchers now believe that Parkinson's disease results from a combination of genetic factors and environmental factors such as exposure to toxins.

Symptoms of Parkinson's Disease

Parkinson's disease has four main symptoms:

- Tremor (trembling) in hands, arms, legs, jaw, or head
- Stiffness of the limbs and trunk
- Slowness of movement
- Impaired balance and coordination, sometimes leading to falls

Other symptoms may include depression and other emotional changes; difficulty swallowing, chewing, and speaking; urinary problems or constipation; skin problems; and sleep disruptions.

Symptoms of Parkinson’s and the rate of progression differ among individuals. Sometimes people dismiss early symptoms of Parkinson's as the effects of normal aging. In most cases, there are no medical tests to definitively detect the disease, so it can be difficult to diagnose accurately.

Early symptoms of Parkinson's disease are subtle and occur gradually. For example, affected people may feel mild tremors or have difficulty getting out of a chair. They may notice that they speak too softly, or that their handwriting is slow and looks cramped or small. Friends or family members may be the first to notice changes in someone with early Parkinson's. They may see that the person's face lacks expression and animation, or that the person does not move an arm or leg normally.

People with Parkinson's often develop a parkinsonian gait that includes a tendency to lean forward, small quick steps as if hurrying forward, and reduced swinging of the arms. They also may have trouble initiating or continuing movement.

Symptoms often begin on one side of the body or even in one limb on one side of the body. As the disease progresses, it eventually affects both sides. However, the symptoms may still be more severe on one side than on the other.

Many people with Parkinson’s note that prior to experiencing stiffness and tremor, they had sleep problems, constipation, decreased ability to smell, and restless legs.

Diagnosis of Parkinson’s Disease

A number of disorders can cause symptoms similar to those of Parkinson's disease. People with Parkinson's-like symptoms that result from other causes are sometimes said to have parkinsonism. While these disorders initially may be misdiagnosed as Parkinson's, certain medical tests, as well as response to drug treatment, may help to distinguish them from Parkinson's. Since many other diseases have similar features but require different treatments, it is important to make an exact diagnosis as soon as possible.

There are currently no blood or laboratory tests to diagnose nongenetic cases of Parkinson's disease. Diagnosis is based on a person's medical history and a neurological examination. Improvement after initiating medication is another important hallmark of Parkinson's disease.

Treatment of Parkinson’s Disease

Although there is no cure for Parkinson's disease, medicines, surgical treatment, and other therapies can often relieve some symptoms.

Medicines for Parkinson's Disease

Medicines prescribed for Parkinson's include:

- Drugs that increase the level of dopamine in the brain
- Drugs that affect other brain chemicals in the body
- Drugs that help control nonmotor symptoms

The main therapy for Parkinson's is levodopa, also called L-dopa. Nerve cells use levodopa to make dopamine to replenish the brain's dwindling supply. Usually, people take levodopa along with another medication called carbidopa. Carbidopa prevents or reduces some of the side effects of levodopa therapy—such as nausea, vomiting, low blood pressure, and restlessness—and reduces the amount of levodopa needed to improve symptoms.

People with Parkinson's should never stop taking levodopa without telling their doctor. Suddenly stopping the drug may have serious side effects, such as being unable to move or having difficulty breathing.

Other medicines used to treat Parkinson’s symptoms include:

- Dopamine agonists to mimic the role of dopamine in the brain
- MAO-B inhibitors to slow down an enzyme that breaks down dopamine in the brain
- COMT inhibitors to help break down dopamine
- Amantadine, an old antiviral drug, to reduce involuntary movements
- Anticholinergic drugs to reduce tremors and muscle rigidity

Deep Brain Stimulation: For people with Parkinson’s who do not respond well to medications, deep brain stimulation, or DBS, may be appropriate. DBS is a surgical procedure that surgically implants electrodes into part of the brain and connects them to a small electrical device implanted in the chest. The device and electrodes painlessly stimulate the brain in a way that helps stop many of the movement-related symptoms of Parkinson's, such as tremor, slowness of movement, and rigidity.

PROPOSED SYSTEM AND METHODS

A. Dataset

Link for Dataset:

https://archive.ics.uci.edu/ml/datasets/parkinsons

The PD data-set which is used in the work is consisting of vocal measurements of 31 people, 23 with Parkinson's disease (PD). Each of the column in the data is a kind of vocal data with some voice measurements, and each row is corresponding to one of 195 voice recordings from these individuals. The aim of the dataset which is used in the work is to differentiate healthy from those suffering with Parkinson's Disease and this is denoted in the "status" column which is having a value of 0 for healthy and 1 for denoting PD. The dataset considered is in a form of CSV format. The entries(rows) of the Comma Separated file(CSV) contains an instance of a suspect which corresponds to a voice recording. The names of every patient is recorded in the first column and there are a total of around six records per patient.[3]

B. Exploratory Data Analysis

The data is loaded into the data-frame and proper EDA is done on the dataset to get the insights of the data. We saw the correlation of the data-points with each other. We plot the data visually so that we can get further insights into the dataset. The histograms give a proper data distribution of the data. As we can see that there is an imbalance in the data-set but we are giving the model variance with making the class balance so that it can perform well with the given dataset.[4]

We have used box-plot to check the presence of outliers and we came to the conclusion that there were very few outliers so we didn't treat those outliers.

We have checked the correlation between the data and the target variable to check the correlation between the status and other independent variables and we found that all the variables contributed some correlation to the status, some positive and some negative so we are dropping no columns.

C. Model Development

We have split the data in 70:30 ratio using train test split from the Scikit-Learn library, where, we are using 70 percent of the data in training the model and 30 percent of the data is used for the testing of the model.

• The first model we have implemented is a Decision Tree Classifier[5]. We have set the hyper-parameters to default and we have using the criterion of 'entropy'.

The model gives an accuracy of 89.8 percent. F1-score of 92 percent.
Confusion matrix is as follows: Figure 1

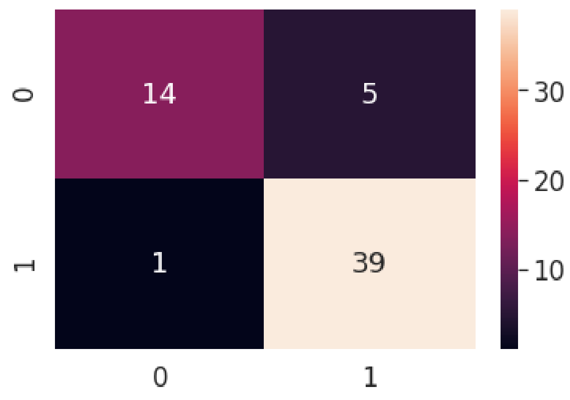


Figure 1

• The second model we used is a Sckit-Learn model name as Decision Tree Classifier with a max depth: 5 and max leaf nodes: 2.
It gave an accuracy of 81 percent

F1 -score of 86 percent and confusion matrix is as follows:

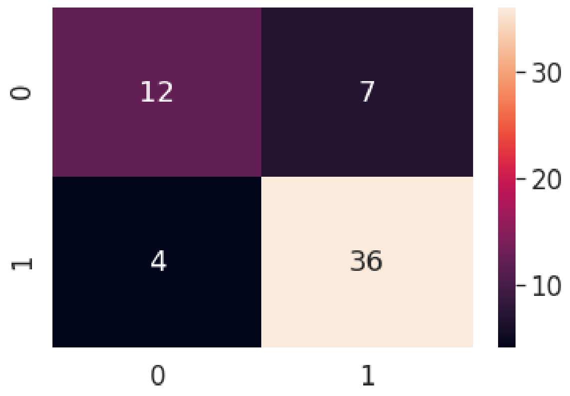


Figure 2

• The third model we used is a decision tree [5] with max depth and leaf nodes 5.

It gave an accuracy of 86 percent.

F1 score of 90 percent and confusion matrix as follows:

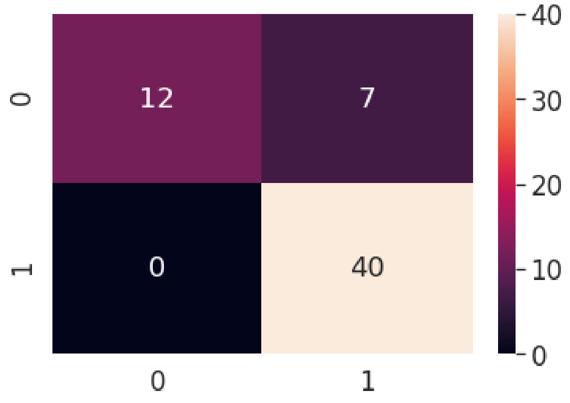


Figure 3

As you can see from the above models, the ideal no of leaf nodes is between 10 to 15. The ideal no of leaf nodes and depth of the tree can be found out by running these model iteratively by tuning the hyperparameters. In the above case, I have showcased how accuracy values change for few cases. It would be a right call to use Grid search or ROC curve (which you will be learning later as a part of this program) to find out the optimal hyperparameters.

Now we'll use Ensemble techniques i.e. Random Forest classifier.

- Random forest [6] gave an accuracy of 89 percent.

  F-1 score of 92 percent approx. Confusion matrix is as follows:

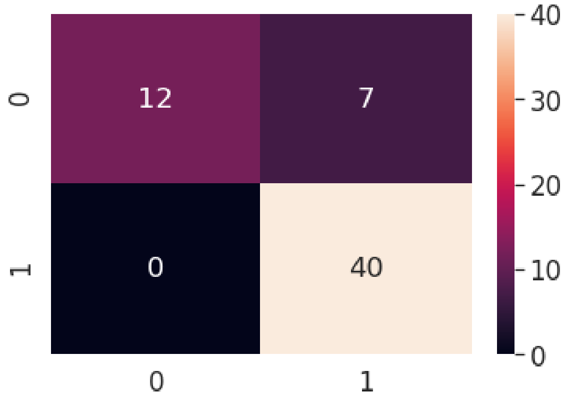


Figure 4

- Next model is with Neural Networks [7]. We fed the data to the neural-networks and every time it gave an accuracy of 76 percent approx.

  F1- score of 83 percent.

Confusion Matrix as follows:

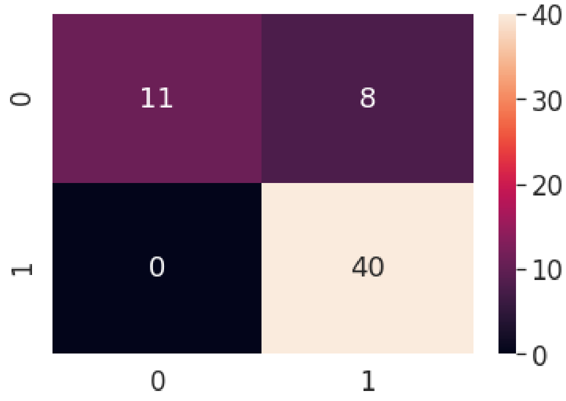


Figure 5

Loss (Figure 7) and Accuracy (Figure 7) curves for the DNN model:

Figure 6

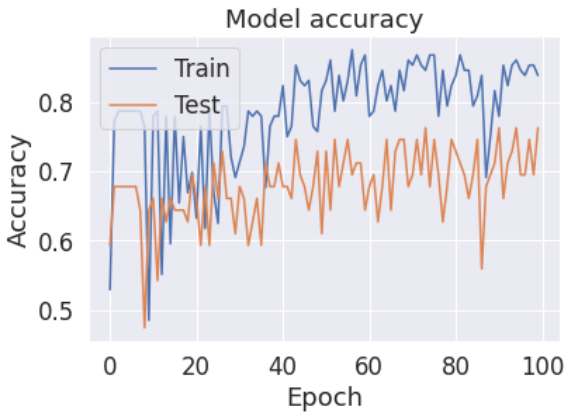


Figure 7

With all the models we have seen, the best performing models are Random Forest and Decision Tree.

RESULTS

A. *Further Discussion.*

Neural Networks did not perform well with this dataset, the reason behind this may be the data of similar type i.e. all the data is vocal recording measurements. Neural networks try to find features inside the given features and that is not useful with similar kind of features. If a variety of features is present neural networks have performed better than this.

When we are training the data for long the training loss decreases and the validation loss is constant at the given time this shows that the model is getting overfitted. That is the reason why we have trained the model to 50 epochs only. We used dropout to create some generalization in the data but still, the accuracy was reliable.

Model comparison is given below:

Model	***Scores***
Decision Tree (Default) Accuracy, F1-score	*89, 92*
Decision Tree (max depth:5, max leaf nodes: 2) Accuracy, F1-score	*81, 86*
Decision Tree (max depth: 5, max leaf	*86, 90*

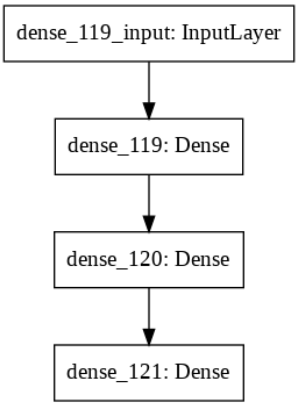
nodes: 5) *Accuracy, F1-score*	
Decision Tree (max depth: 5, max leaf nodes: 10) *Accuracy, F1-score*	*86, 90*
Decision Tree (max depth: 5, max leaf nodes: 15) *Accuracy, F1-score*	*86, 90*
Random Forest (max depth: 5, max leaf nodes: 15) *Accuracy, F1-score*	89, 92
DNN Sequential 	76, 83

Figure 1

CONCLUSION

With all the models the best score achieved is 92 percent with the decision tree classifier and Random Forest Classifier so it should be used in case of classification of PD with the dataset used.

It gives the best prediction that whether the person who is tested is affected from Parkinson's Disease or not in a very early stage.

When the data is more complex in terms of adding more sensor data, Neural networks can perform better but as far as we are concerned with similar data with sensor information of similar types Classical Machine Learning Algorithms can perform better.

We can use synthetic data generator SMOTE [8] to reduce the class imbalance to further improve the scores. In a case where data is having more number of outlier, we can treat them using Z-score for better treatment of variance in the data as far as generalizing the model for better understanding the data.

REFERENCES

[1] "Parkinson's Foundation: Better Lives. Together." https://www.parkinson.org

[2] Prof Thomas Davenport, "The potential for artificial intelligence in healthcare". Future Healthc J. 2019 Jun; 6(2): 94–98 10.7861/futurehosp.6-2-94

[3] "UCI Machine Learning Repository: Parkinson's Data Set." https://archive.ics.uci.edu/ml/datasets/parkinsons

[4] "Exploratory Data Analysis", https://en.wikipedia.org/wiki/Exploratory_data_analysis

[5] "Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study" International Journal of Computer Applications (0975 – 8887) Volume 52 – No. 6, August 2012

[6] "A Random Forest based predictor for medical data classification using feature ranking" Alam, Md & Rahman, M. & Rahman, Mohammad. (2019). A Random Forest based predictor for medical data classification using feature ranking. Informatics in Medicine Unlocked. 15. 100180. 10.1016/j.imu.2019.100180.

[7] "Applications of artificial neural networks in health care organizational decision-making: A scoping review" Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PLoS One*. 2019;14(2):e0212356. Published 2019 Feb 19. doi: 10.1371/journal.pone.0212356

[8] "SMOTE for high-dimensional class-imbalanced data." Blagus, R., Lusa, L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics* 14**,** 106 (2013). https://doi.org/10.1186/1471-2105-14-106