GALGOTIAS UNIVERSITY

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

ANALYSIS OF PARKINSON'S DISEASE USING RANDOM FOREST

A Report for the Evaluation of Project 2

Submitted by

Yash Warrdhan Gautam

(1613105142/16SCSE105050)

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. SHUBHAM KUMAR

APRIL / MAY- 2020

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **“ANALYSIS OF PARKINSONS DISEASE USING RANDOM FOREST TECHNIQUE IN MACHINE LEARNING ”** is the bonafide work of **“YASH WARRDHAN GAUTAM (1613105142)”** 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

Dr. SANJEEV KUMAR PIPAL, M.Tech., Ph.D.,
Professor

School of Computing Science & Engineering

ABSTRACT

Parkinson's disease (PD) is a nerve disorder in which the dopamine level of the brain decreased that leads to shaking, stiffness, difficulty in walking, balance and coordination. In the early stages of Parkinson's disease, you may little or no expression. Your arm may not swing when you walk, it will affect your speech, writing style and the rest of all the automatic movements of the body and these symptoms gets worse over time.

Traditional diagnosis of Parkinson's disease involves regular seating in a clinic and observing the motor skills of the patient in various situations. Since there is no definitive laboratory to test, diagnose of Parkinson's disease, therefore it's difficult to diagnose the Parkinson's disease at early stages when motor effects are not yet serve. Monitoring progression of the disease requires repeated clinic visits by the patient. An effective screening process that doesn't require a clinic visit, would be beneficial. Since Parkinson's disease patients exhibits characteristics of vocal feature, voice recordings are useful and non-invasive tool for diagnosis. 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

TABLE OF CONTENTS

INTRODUCTION

Various classification methods and algorithms have been developed for resolving machine learning problems, including statistical models, decision and regression trees, rules, connectionist networks, probabilistic networks. Supervised classification is a principle core of what has been recently called the data mining. The applications of supervised classification in real life are very vast, like automatic speech treatment, face detection, signature recognition, customer discovery, spam detection, systems biology etc. Many decision making problems in a variety of domains such as medical science, engineering, human sciences and management science can be considered as classification problems.

In the present time, the number of aging people in Thailand is raising every day. Parkinson's disease is one of afflictions found in these people. Parkinson's disease is a progressive neurodegenerative disorder. Parkinson's disease is both debilitating and negatively impacting on quality of life in later years and age is the most consistent risk factor. The essential neuropathological changes in Parkinson's Disease are a loss of melanin containing dopaminergic neurons in the substantia nigra pars compacta. This results in a dysfunction of the basal ganglia circuitrym, which mediates motoric and cognitive functions. The severity of patients depends on their Parkinson's disease state. As Parkinson's disease affects motoric function, the major problems caused by Parkinson's disease over the course of their illness are about their voice or speech. These problems include reduced prosodic pitch inflection; breathy, harsh, or hoarse voice quality; reduction of vocal loudness; imprecise and reduced range of articulatory movements; short rushes of speech, voice tremor; sometimes lead by a decay in vocal loudness and articulatory movements toward the end of a phrase or a sentence; audible or inaudible dysfluencies; and overall reduction in speech intelligibility. Though its cause is related to the lack of cells of substantia nigra, which contain the neurotransmitter dopamine, but for any particular case, the cause is unknown. Some scientists believe that a change in a specific gene, may be the reason while other experts think that it could be something in the environment that causes the damage, such as pesticides or other chemicals. Although the role that heredity plays isn't completely understood, about 1 million people in the United States have Parkinson's disease. At present, there is no cure for Parkinson’s disease, but a variety of medications like a combination of levodopa and carbidopa provide dramatic relief from the symptoms.

The diagnosis of the disease is difficult in some cases due to its overlapping symptoms, but vocal impairment is considered one of the earliest indications of the onset of this disease. Amongst the various vocal tests conducted, sustained phonation is one. One major challenge in data analysis is due to the problems of size where the number of the characteristics far exceeds the number of observations. In addition, a major problem which may affect each study in data analysis is the high correlations among attributes. That may be efficient to learn from data when we overcome the complex interactions among characteristics

The National Institute of Neurological Disorders and Stroke (NINDS) conducts Parkinson’s disease research in laboratories at the National Institutes of Health (NIH) and promotes the advancement of research directed to the understanding, treatment and eventual cure of Parkinson’s disease.

LITERATURE REVIEW

Previous studies have shown the implementation of Neural networks in assessment of the risk factors associated with Parkinson's disease. Factors such as genes, age, stroke and diabetes were found out to have higher correlation with the outcome of the prediction.

Feature relevance analysis has also been carried out previously using Parkinson's voice dataset. Tensor flow deep learning library has also been used to detect the increase in disease severity in case of Parkinson's disease

Subtype identification of PD has also been done before using both supervised and unsupervised machine learning methods. The progression rate of the disease can be predicted using these learning systems and can be classified as: highly progressive, moderately progressive and mild progressive state of Parkinson's Disease. This has been utilized effectively in order to improve the quality of the counselling of patients, improve clinical trial design and using resources available in a better way.

Work has also been carried out on MRI dataset of the brain for identification of medical image related biomarkers. Supervised Machine Learning classifiers such as Support Vector Machine has been utilized in order to understand the progression of the disease and role of the biomarkers in identifying the progression behaviour.

Gait classification in case of patients suffering from PD has also been studied using Artificial Neural Networking (ANN) and Support Vector Machine classifiers. Kinematic, kinetic and spatiotemporal gait are the 3 basic parameters, which are to be taken under consideration before performing the classification task.

METHODS AND MATERIALS

Subject

In this study, Data and attributes are taken from an open source, Anaconda platform along with libraries like Pandas and Numpy are used for data manipulation and analysis. First the data is imported from a source and stored in the form of pdf file, in a structured format. From the above structured format then EDA [Exploratory data analysis] is done, the attributes are identified and Bivariate analysis is used for analysis because it's one simplest forms of quantitative analysis. It involves the analysis of two variables, for the purpose of determining the empirical relationship between them.

We have used different models of machine learning and combined them all to obtain better predictive performance than could be obtained from any of the constituent learning algorithms. This process of combing different machine learning models in one is called ensemble methods.

Method that can be followed to achieve the goal is given below:

1) **Pre-processing**: Pre-processing of data is the mandatory step which is followed in machine learning before applying any machine learning algorithm. Exploratory data analysis is an approach to analysing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA

is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task.

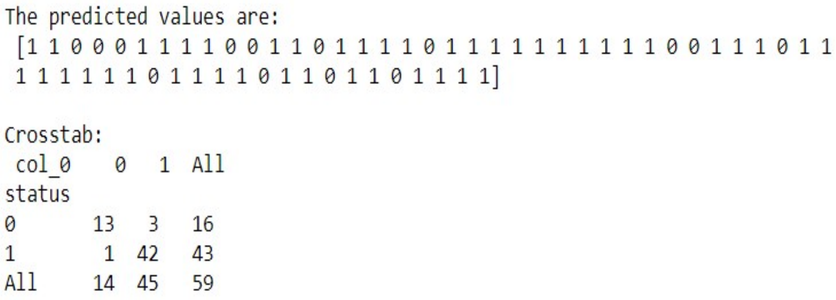
2) **Visualization:** Second step is the visualization of the data the data is analysed visually using pair plot Fig [1]. Pair plot allows us to see both distribution of single variables and relationships between two variables. Pair plots are a great method to identify trends for follow-up analysis.

a. Fig [1]

3) **Training of data:** Data is divided in a ratio of 70:30. Major part of data is training the model using machine learning algorithm**.** The rest of data is used later for testing of the model and to verify whether the trained model is trained properly or not**.**

4) **Decision tree**: Decision tree algorithm is used for identifies ways to split a data set based on different conditions. The model is used to predict the data and check the accuracy score of the data.

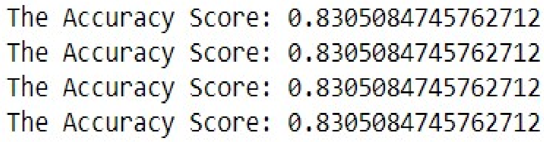
    a. Then finally the crosstab () function is used to compute a simple cross tabulation of two (or more) factors.as shown Fig [2]



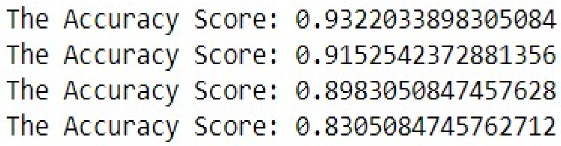
```
The predicted values are:
 [1 1 0 0 0 1 1 1 1 0 0 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 0 1 1
 1 1 1 1 1 1 0 1 1 1 1 0 1 1 0 1 1 0 1 1 1 1]

Crosstab:
 col_0    0   1  All
status
0       13   3   16
1        1  42   43
All     14  45   59
```

RESULTS

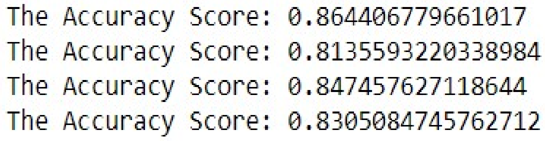
In the result of classification using ensemble methods in Fig [3], we can see the details of accuracy measure by class. Here true positive rate and false negative rate for cross validation fold values is shown



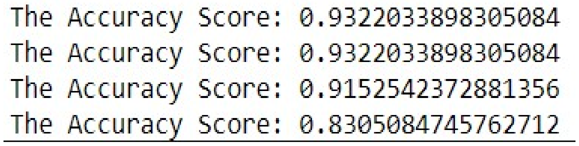
```
The Accuracy Score: 0.8305084745762712
The Accuracy Score: 0.8305084745762712
The Accuracy Score: 0.8305084745762712
The Accuracy Score: 0.8305084745762712
```



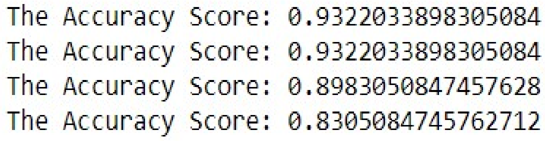
```
The Accuracy Score: 0.9322033898305084
The Accuracy Score: 0.9152542372881356
The Accuracy Score: 0.8983050847457628
The Accuracy Score: 0.8305084745762712
```



```
The Accuracy Score: 0.864406779661017
The Accuracy Score: 0.8135593220338984
The Accuracy Score: 0.847457627118644
The Accuracy Score: 0.8305084745762712
```



```
The Accuracy Score: 0.9322033898305084
The Accuracy Score: 0.9322033898305084
The Accuracy Score: 0.9152542372881356
The Accuracy Score: 0.8305084745762712
```



```
The Accuracy Score: 0.9322033898305084
The Accuracy Score: 0.9322033898305084
The Accuracy Score: 0.8983050847457628
The Accuracy Score: 0.8305084745762712
```

Fig 3

From the above result we can infer which patient are suffering from Parkinson’s disease and the accuracy of our result is shown Fig [3].

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 over fitted. 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

CONCLUSION

This study designed for analyse speech characteristics of Parkinson’s disease and to verify the that the person is suffering from Parkinson’s Disease or not. The Parkinson’s disease state can also be compared to find the changing of speech characteristic along the increasing of Parkinson’s disease state. Once we confirm that, what is the stage of patient whether it’s the beginning stage of it’s the last stage, we can do the therapy accordingly. These analyses can use for speech assessment in speech therapy for Parkinson’s disease patients.

We can conclude that If machine learning algorithms could be applied to a voice recording dataset to accurately diagnose Parkinson’s disease, this would be an effective screening step patient don’t need appointment with a clinic. Hence the analysis can be done using voice of the person.

FUTURE WORKS

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) Buder E.H., Kent R.D., Kent J.F., Milenkovic P., and Workinger M.S. FORMOFFA: An automated formant, moment, fundamental frequency, amplitude analysis of normal and disordered speech. Clinical Linguistics and Phonetics, 10, 1996, 31–54.
2) Dunham, W. Heron's Formula for Triangular Area. Journey through Genius: The Great Theorems of Mathematics, New York: Wiley, 1990, 113–132.
3) Harel B. T. et al. Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment. Journal of Neurolinguistics, 17, 2004, 439– 453
4) Zhou, J. Xu, W. and Zhu, L. 2015. Robust estimating equation-based sufficient dimension reduction. J. Multivar. Anal. 134, 99–118. DOI= http://dx.doi.org/10.1016/j.jmva.2014.10.006.
5) Albright, S. C. Winston, W. L. and Zappe, C. J. 2011. Data Analysis and Decision Making. 4th ed. South-Western. Cengage Learning.
6) Cho, H. and Fryzlewicz, P. 2012. High dimensional variable selection via tilting. J. R. Stat. Soc. Ser. B. 74(3), 593-622
7) Harel, B. Cannizzaro, M. and Snyder, P. J. 2004. Variability in fundamental frequency during speech in prodromal and incipient Parkinson’s disease: A longitudinal case study. Brain Cogn. 56, 24–29
8) Campenhausen, A. S. Bornschein, B. Wick, R. Bötzel, K. Sampaio, C. Poewe, W. Oertel, W. Siebert, U. Berger, K. and Dodel R. 2005. Prevalence and incidence of Parkinson's disease in Europe. Eur. Neuropsychopharmacol. 15, 473-490
9) Little M. A., McSharry P.E., Hunter E.J., Ramig L.O. (2008), Suitability of dysphonia measurements for telemonitoring of Parkinson's disease ser IEEE Transactions on Biomedical Engineering

10) Marius Ene Neural network-based approach to discriminate healthy from those with Parkinson's disease, Annals of the University of Craiova, Math. Comp. Sci. Ser. Volume 35, 2008, Pages 112{116 ISSN: 1223-6934}.

11) JACQUES COHEN, Bioinformatics—An Introduction for Computer Scientists, Brandeis University. ACM Computing Surveys, Vol. 36, No. 2, June 2004, pp. 122–158.

12) http://www.ninds.nih.gov/disorders/parkinsons_disease/parki nsons_disease.htm

13) http://kidshealth.org/kid/grownup/conditions/parkinson