EARLY DETECTION OF PARKINSON'S DISEASE USING MACHINE LEARNING

Anitha R[1], Nandhini T[2], Sathish Raj S[3], Nikitha V[4]

[1] *Assistant Professor, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Tamil Nadu, India*
[2] *UG Student, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Tamil Nadu, India*
[3] *UG Student, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Tamil Nadu, India*
[4] *UG Student, Department of Computer Science and Engineering, SRM Valliammai Engineering College, Tamil Nadu, India*

ABSTRACT

Parkinson's disease (PD) is a neurodegenerative movement disease where the symptoms gradually develop start with a slight tremor in one hand and a feeling of stiffness in the body and it became worse over time. It affects over 6 million people worldwide. At present there is no conclusive result for this disease by non-specialist clinicians, particularly in the early stage of the disease where identification of the symptoms are very difficult in its earlier stages. The proposed predictive analytics framework is a combination of K-means clustering and Decision Tree which is used to gain insights from patients. By using machine learning techniques, the problem can be solved with minimal error rate. Voice data sets obtained from the UCI Machine learning repository if given as the input for voice data analysis. Also our proposed system provides accurate results by integrating spiral drawing inputs of normal and Parkinson's affected patients. From these drawings Random forest classification algorithm is used which converts these drawings into pixels for classification and the extracted values are been matched with the trained database to extract various features and results are produced with maximum accuracy. Also OpenCV (Open Source Computer Vision Library) a library of programming functions mainly aimed at real-time computer vision was built to provide an infrastructure for computer vision applications and to accelerate the use of machine perception in the real time. Thus our output will showcase the early detection of the disease and can be able to increase the lifespan of the diseased patient with proper treatments and medications leads to peaceful life.

Keywords: *- k-means clustering, decision tree, Random forest classification.*

1. INTRODUCTION

Parkinson's disease symptoms can be different for everyone. Early signs are mild that goes unnoticed. Symptoms usually begin on one side of your body and gets worsen on that side, afterwards it affects both the sides. Parkinson's symptoms may include

- Tremor
- Slowed movement
- Rigid muscles.
- Impaired posture and balance.
- Loss of automatic movements
- Speech changes
- Writing changes

The Parkinson's disease is due to a loss of neurons that produce a chemical messenger in the brain called dopamine. when there is a decrease in level of the amino acid named dopamine it leads to the abnormal brain activity, which leads to Parkinson's disease. The cause of Parkinson's disease is still a question mark, but several factors appear to play a role, including:

- Genes
- Environmental
- Triggers

As a result people suffer from this disease for many years before diagnosis. The estimated results have shown that there are 7-10 million people are affected by parkinson's disease worldwide. People with age above 50 are the one's who has the higher possibility of getting parkinson's disease but still an estimated 4 percentage of people who are under the age 50 are diagnosed with parkinson's disease. There is no cure or prevention for PD. However, the disease can be controlled in early stage. The data mining techniques is used as a effective way for early detection and diagnosis of the disease. Data mining techniques in medicine is a research area that combines sophisticated representational and computing techniques with the insights of expert physicians to produce tools for improving healthcare.Data mining is a statistical method for finding hidden patterns in datasets by constructing predictive or cla ssification models that can be learned from past experience and applied in future cases, so there is a need for a more accurate, objective means of early detection, ideally one which can be used by individuals in their home setting.

2. EXISTING SYSTEM

In existing system, PD is detected at the secondary stage only (Dopamine deficiency) which leads to medical challenges. Also doctor has to manually examine and suggest medical diagnosis in which the symptoms might vary from person to person so suggesting medicine is also a challenge. Thus the mental disorders are been poorly characterized and have many health complications. PD is generally diagnosed with the following clinical methods as,

- **MRI or CT scan -** Conventional MRI cannot detect early signs of Parkinson's disease
- **PET scan -** is used to assess activity and function of brain regions involved in movement
- **SPECT scan -** can reveal changes in brain chemistry, such as a decrease in dopamine

This results in a high misdiagnosis rate (up to 25% by non-specialists) and many years before diagnosis, people can have the disease. Thus existing system is not effective in early prediction and accurate medicinal diagnosis to the affected people

3. PROPOSED METHOD

By using machine learning techniques, the problem can be solved with minimal error rate. The voice dataset of Parkinson's disease from the UCI Machine learning library is used as input. Also our proposed system provides accurate results by integrating spiral drawing inputs of normal and Parkinson's affected patients. We propose a hybrid and accurate results analyzing patient both voice and spiral drawing data's. Thus combining both the results, the doctor can conclude normality or abnormality and prescribe the medicine based on the affected stage.

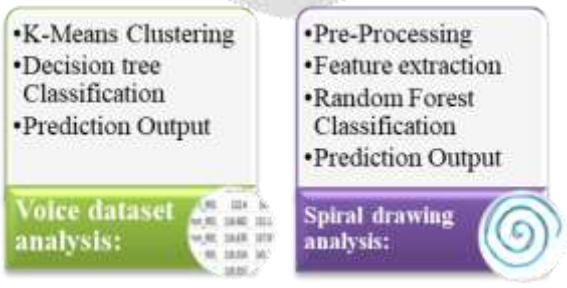

Fig -1: Proposed methodology

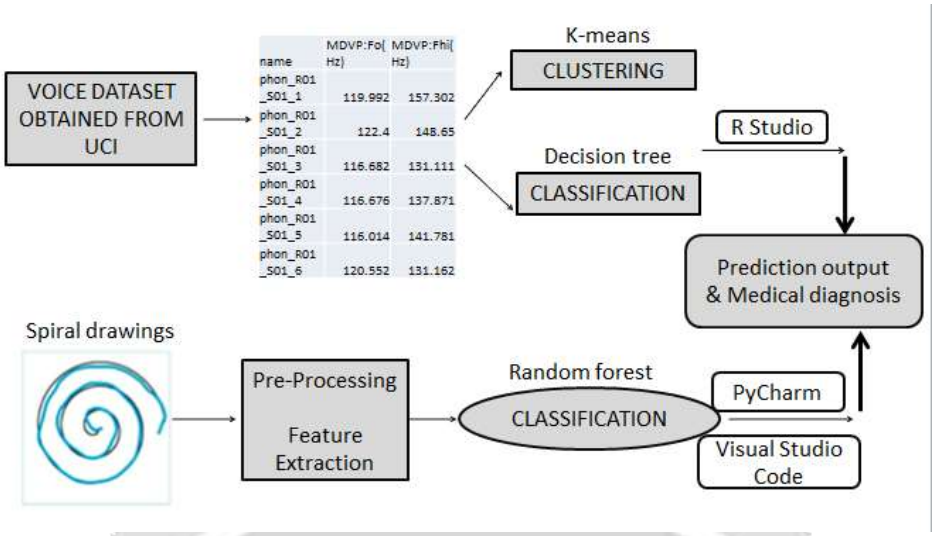

Fig -2: Architecture diagram

4. MODULE DESCRIPTION

4.1 PARKINSON'S DISEASE VOICE DATASET ANALYSIS

PD voice dataset is collected from UCI machine learning repository and these are stored into the RStudio environment as Testing and Training datasets. These are stored into the RStudio environment as Testing and Training datasets. R is a programming language and software environment for statistical analysis, graphics representation, data analysis and as well as machine learning. It involves the following steps and procedures,

1. Importing data to RStudio - organize the data in an Excel worksheet to include column names in the first row (i.e. person's voice collected at various time zones) and each subsequent row contains all the information (i.e. set of 22 parameter is taken into consideration and the person's voice range for those parameters is tested and then noted), finally the status column shows two values 0 (healthy) and 1(affected). Import data into RStudio, using the "Import data..." feature.

2. Clustering (k-means) - An unsupervised learning algorithm that tries to cluster data based on their similarity, and just tries to find patterns in the data. Here, we have to specify the number of clusters we want the data to be grouped into and then the algorithm randomly assigns each observation to a cluster, and finds the centroid of each cluster and then, it iterates by reassigning data points to the cluster whose centroid is closest and calculates new centroid of each cluster.

3. Classification (Decision Tree) – It is also called a prediction tree. It uses a structure to specify sequences of decisions and consequences, the goal is to predict a response or output.The forecast can be accomplished by creating a decision tree with test points and branches.At each check poi nt, a decision is made to pick a particular branch and cross the trees and can be used in a variety of disciplines, on the basis of individual characteristics

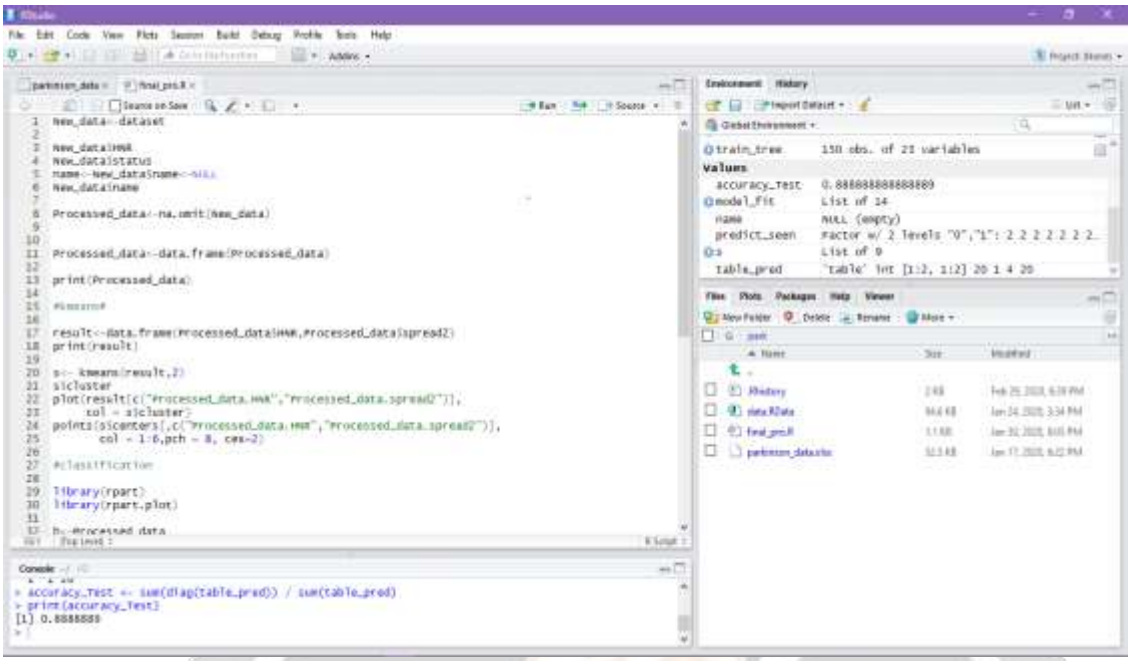

Fig -3: Implementation of code in RStudio

4. Predicted Output - The predicted output for voice data analysis based on clustering and classification is with an accuracy of 88%

4.2 PARKINSON'S DISEASE SPIRAL DRAWING ANALYSIS:

Spiral drawing datasets of PD affected and unaffected patients collected by neurologists are obtained from Machine Learning repository. These are stored into the python environment as Testing and Training datasets and imported usingnecessary packages. Python is an open source dynamic, high level, free and interpreted programming language. This supports objectoriented programming and procedural programming.Python is currently the most popular progra mming language for Machine Learning research and development.PyCharm is an integrated development environme nt (IDE) primarily for the Python language, used in computer programming. Microsoft Visual Studio is a development environment by Microsoft. It is used to develop computer programs, websites, web applications, web services, and mobile apps. 36 different programming languages are supported by Visual Studio which includes C#, C++, etc., and allows the code editor and debugger to support nearly any programming language, provided a language-specific service exists. Built-in languages include C, C++, C++/CLI, Visual Basic .NET, C#, JavaScript, Typescript, XML, HTML, and CSS.

1. Importing datasets into PyCharm/Visual studio code - Spiral drawing datasets of PD affected and unaffected patients collected by neurologists are obtained from Machine Learning repository. These are stored into the python environment as Testing and Training datasets and imported using necessary packages.

2. Pre-Processing – It involves image acquisition, pre-processing and segmentation. Preprocessing image is a way to improve image quality, so that the resulting image is better than the original one.
The goal of image acquisition is to collect images having low noise when compared to HD images. The main advantage of this module is to have images with better clarity, low noise and distortion.The aim of segmentation is to make the representation of an image simpler or more easily analyzable.

3. Feature extraction - In this project, mean filter and median filter are presented for processing of selecting the images. The median filter is a non-linear tool, while linear is the average filter. Mean filtering of smoothing images is fast, intuitive and easy to implement i.e. reduces the amount of variation in intensity between one pixel and the next. The median filter is normally used in a picture to reduce salt-and-pepper noise. It often does a better job than maintaining useful 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 area and then replacing the pixel that is considered with the middle pixel value. If there are even number of pixels in the neighborhood under consideration the sum of the two middle pixel values is used. Both mean and median filters are used to remove noise. This is used as the input for further analysis.

4. OpenCV library function - OpenCV (Open Source Computer Vision Library) was developed to provide an interface for computer vision applications and to facilitate the use of machine perception in the real time

5. Classification (Random Forest) – It is a supervised learning algorithm used for classification. Random forest algorithm builds decision trees on data samples, then obtains the prediction from each and finally sel ects the best solution by voting.It is an ensemble approach that is better than a single decision tree, as it eliminates o verfitting by averaging the outcome. Where we can find the confusion matrix with the help of confusion_matrix() function of sklearn, which is nothing but a table with two dimensions viz. "Actual" and "Predicted" and furthermore, both the dimensions have "True Positives (TP)", "True Negatives (TN)", "False Positives (FP)", "False Negatives (FN)",which calculates accuracy, specificity and sensitivity

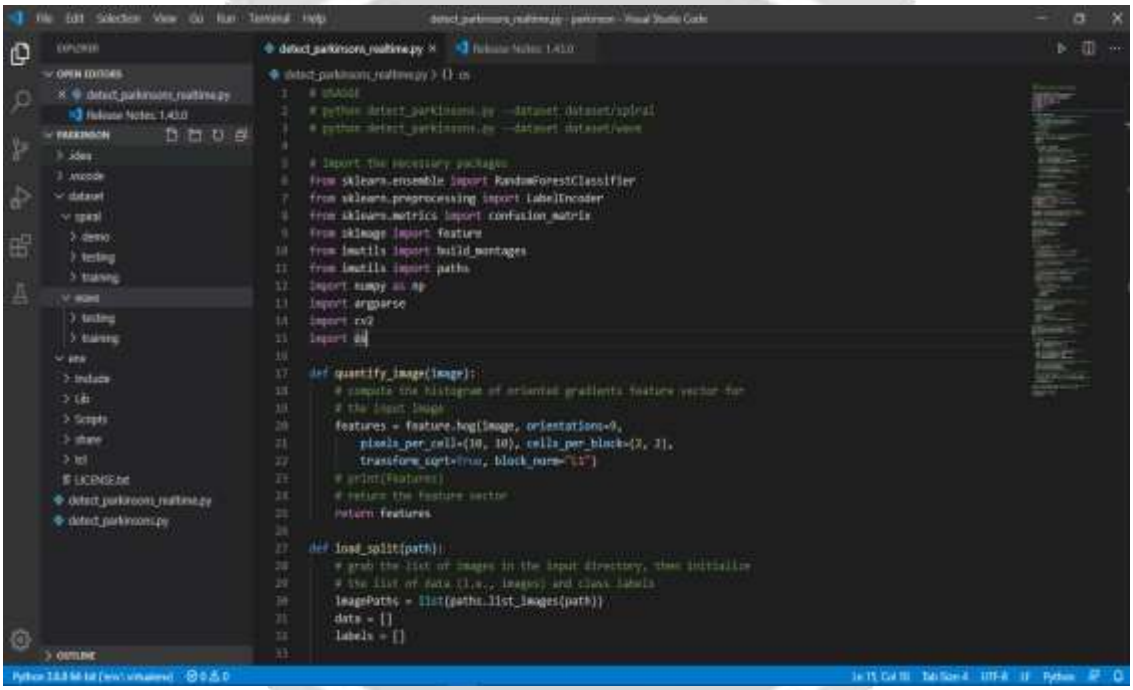

Fig -4: Implementation of code in Visual Studio Code

6. Predicted Output - Thus our hybrid architecture, integrating image processing (spiral drawing analyzing) using image processing technique, the predicted output based on Random forest Classification and confusion matrix is with an accuracy of 83%. Also it produces real-time accurate results by giving a person's spiral drawing as an input to the OpenCV function, that indicates whether a person is healthy or affected by Parkinson's.

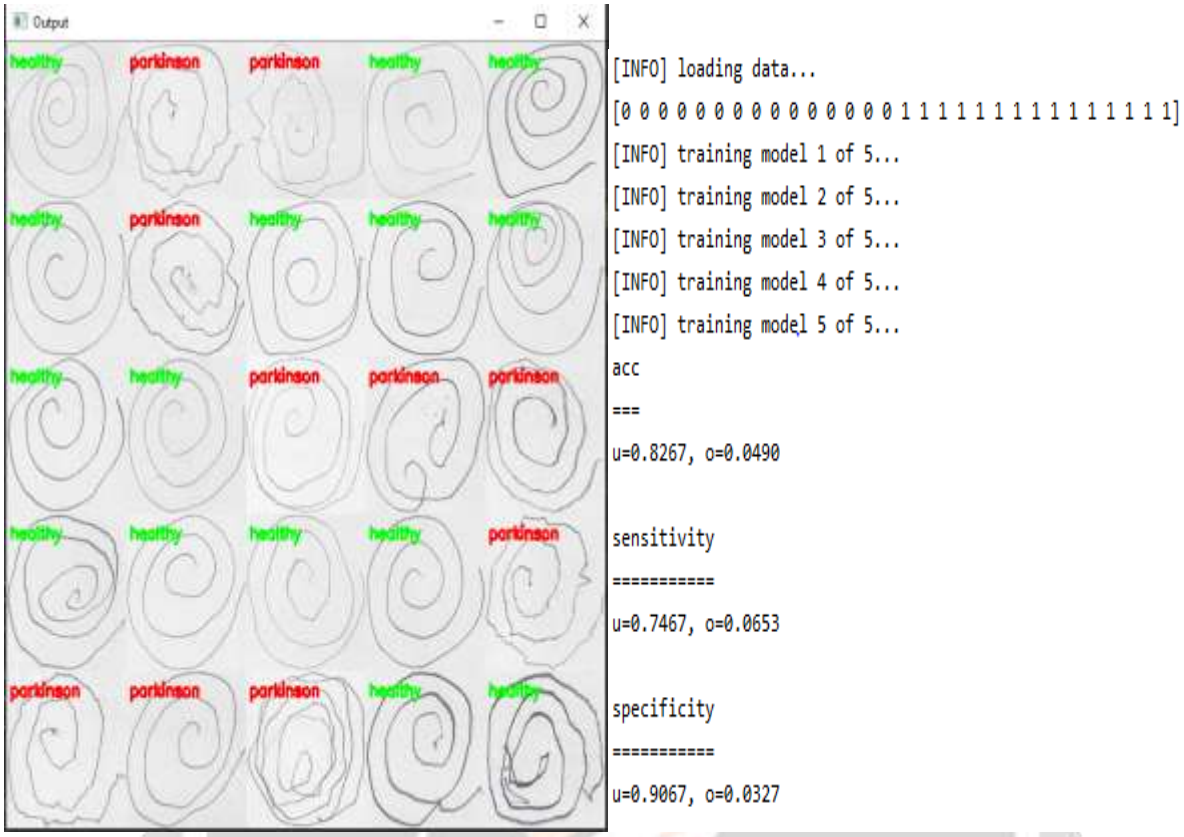

Fig -5: Tested output for spiral drawings

The real-time accurate result by giving a person's spiral drawing as an input to the above tested function, that produces results and indicates whether a person is healthy or affected by Parkinson's.

Fig -6: healthy

Fig -5: Parkinson's affected

5. CONCLUSION

Previous review papers provides a comprehensive survey of relevant neuroimaging modalities and associated analysis techniques presented in the recent years for diagnosing Parkinson's disease. Previous review papers have focused only on a particular imaging modality such as MRI or PET, or on one specific type of dementia only such as AD. This project aimed to cover a broader space of imaging and machine learning technologies for mental illness diagnostics such that researchers in the field could readily identify the state of the art in the domain. Moreover, we emphasize the importance of early detection and prediction of Parkinson's disease, such that treatment and support can be provided to patients as soon as possible.

6. FUTURE WORK

In future work, we can focus on different techniques to predict the Parkinson disease using different datasets. In this research, we using binary attribute (1- diseased patients, 0-non-diseased patients) for patient's classification. In the future we will use different types of attributes for the classification of patients and also identify the different st ages of Parkinson's disease.

7. APPLICATIONS

- Used to detect Dementia at early stage.
- Used to detect neurodegenerative disorders.
- Used for clinical diagnosis for patients above 50 years.

8. REFERENCES

[1] Adrien Payan, Giovanni Montana, Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks.

[2] Alemami, Y. and Almazaydeh, L. (2014) Detecting of Parkinson Disease through Voice Signal Features. Journal of American Science,

[3] Fayao Liu, Chunhua Shen, Learning Deep Convolutional Features for MRI Based Alzheimer's Disease Classification.

[4] Hadjahamadi, A.H. and 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.

[5] Little, M.A., McSharry, P.E., Hunter, E.J. and Ramig, L.O. (2008), Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's disease. IEEE Transactions on Biomed ical Engineering, 56, 1015-1022.

[6] Muhlenbach, F. and Rakotomalala, R. (2015) Discretization of Continuous Attributes. In: Wang, J., Ed., Encyclopedia of Data Warehousing and Mining, Idea Group Reference, 397-402.

[7] Olanrewaju, R.F., Sahari, N.S., Musa, A.A. and Hakiem, N. (2014) Application of Neural Networks in Early Detection and Diagnosis of Parkinson's Disease. International Conference on Cyber and IT Service Management.

[8] Saman Sarraf, Danielle D. DeSouza, John Anderson, Ghassem Tofighi, DeepAD: Alzheimer′s Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI, Cold Spring Harbor Laboratory Press.