ZKG INTERNATIONAL

ISSN: 2366-1313

A EFFICIENT PARKINSON'S DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

M. RAHUL[1], S. SINDU[2], P. HARI KRISHNA[3], T. HARSHITA[4], T. PRAVEEN[5].

[1] Associate Professor, CSE,Chalapathi Institute of Technology,Guntur, India

[2]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

[3]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

[4]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

[5]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

ABSTRACT: In various data repositories, there are large medical datasets available that are used to identifying the diseases. Parkinson's is considered one of the deadliest and progressive nervous system diseases that affect movement. It is the second most common neurological disorder that causes disability, reduces the life span, and still has no cure. Nearly 90% of affected people with this disease have speech disorders. In real-world applications, the information is been generated by using various Machine Learning techniques. Machine learning algorithms help to generate useful content from it. To increase the lifespan of elderly people the machine learning algorithms are used to detect diseases in the early stages. Speech features are the main concept while taking into consideration the term 'Parkinson's'.

The Dataset which is used in this project is taken from the Kaggle website and the dataset is verified by the doctors. The Parkinson's disease is progressive neuro degenerative disorder that affects a lot only people significantly affecting their quality of life. It mostly affect the motor functions of human. The main motor symptoms are called "parkinsonism" or "parkinsonian syndrome" From the whole data 60% is used for training and 40% is used for testing. The data of any person can be entered in db to check whether the person is affected by Parkinson's disease or not. There are 24 columns in the data set each column will indicate the symptom values of a patient except the status column. The status column has 0's and 1's those values will decide the person is effected with Parkinson's disease. I's indicate person is effected, 0's indicate normal conditions.

1. INTRODUCTION

The recent report of the World Health Organization shows a visible increase in the number and health burden of Parkinson's disease patients increases rapidly. In China, this disease is spreading so fast and estimated that it reaches half of the population in the next 10 years. Classification algorithms are mainly used in the medical field for classifying data into different categories according to the number of characteristics. Parkinson's disease is the second most dangerous neurological disorder that can lead to shaking, shivering,

Stiffness, and difficulty walking and balance. It caused mainly due by the breaking down of cells in the nervous system. Parkinson's can have both motor and non-motor symptoms.

The motor symptoms include slowness of movement, rigidity, balance problems, and tremors. If this disease continues, the patients may have difficulty walking and

ISSN: 2366-1313

talking. The non-motor symptoms include anxiety, breathing problems, depression, loss of smell, and change in speech. If the above-mentioned symptoms are present in the person then the details are stored in the records.

In this project, the author considers the speech features of the patient, and this data is used for predicting whether the patient has Parkinson's disease or not. Neurodegenerative disorders are the results of progressive tearing and neuron loss in different areas of the nervous system. Neurons are functional units of the brain. They are contiguous rather than continuous. A good healthy looking neuron as shown in fig 1 has extensions called dendrites or axons, a cell body, and a nucleus that contains our DNA. DNA is our genome and a hundred billion neurons contain our entire genome which is packaged into it. When a neuron gets sick, it loses its extension and hence its ability to communicate which is not good for it and its metabolism becomes low so it starts to accumulate junk and it tries to contain the junk in the little packages in little pockets. When things become worse and if the neuron is a cell culture it completely loses its extension, becomes round and full of vacuoles.

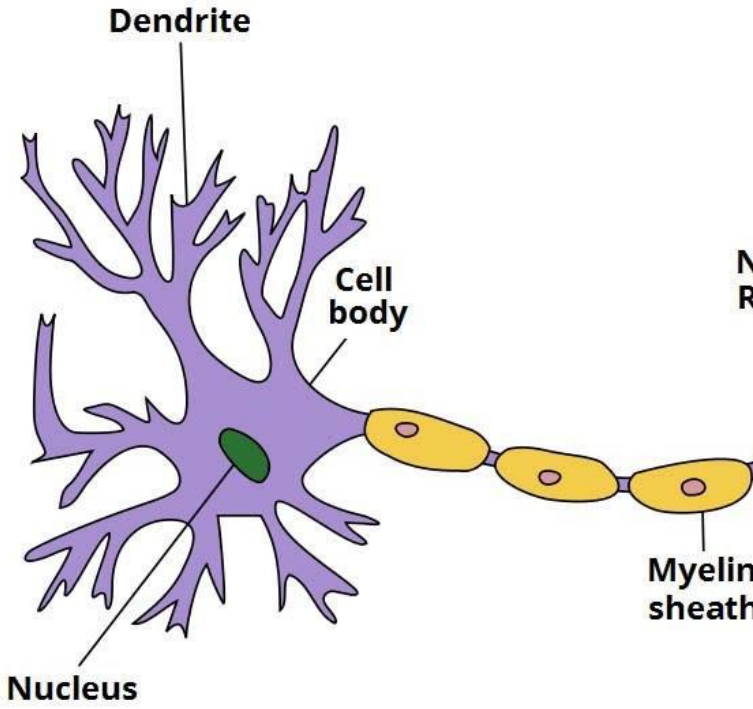

Fig-1 Structure of Neuron

This work deals with the prediction of Parkinson's disorder which is now a day is tremendously increasing incurable disease. Parkinson's disease is a most spreading disease which gets its name from James Parkinson who earlier described it as a paralysis agita's and later gave his surname was known as PD. It generally affects the neurons which are responsible for overall body movements. The main chemicals are dopamine and acetylcholine which affect the human brain. There is a various environmental factor which has been implicated in PD below are the listed factor which caused Parkinson's disease in an individual.

Environmental factors: Environment is defined as the surroundings or the place in which an individua lives. So the environment is the major factor that will not only affects the human's brain but also affects all the living organism who lives in the vicinity of it. Many types of research and evidence have proved that the environment

ISSN: 2366-1313

has a big hand in the development of neurodegenerative disorders mainly Alzheimer's and Parkinson's. There are certain environmental factors that are influencing neurodegenerative disorder with high pace are:

Exposure to heavy metals (like lead and aluminum) and pesticides.

Air Quality: Pollution results in respiratory diseases.

Water quality: Biotic and Abiotic contaminants present in water lead to water pollution.

Unhealthy lifestyle: It leads to obesity and a sedentary lifestyle.

Psychological stress: It increases the level of stress hormone that depletes the functions of neurons.

Brain injuries or Biochemical Factors: The brain is the control center of our complete body. Due to certain trauma, people have brain injuries which leads some biochemical enzymes to come into the picture which provides neurons stability and provides support to some chromosomes and genes in maintenance.

Aging Factor: Aging is one of the reasons for the development of Parkinson's disease. According to the author in India, 11,747,102 people out of 1, 065, 070, 6072 are affected by Parkinson's disease.

Genetic factors: Genetic factor is considered as the main molecular physiological cause which leads to neurodegenerative disorders. The size, depth, and effect of actions of different genes define the status or level of neurodegenerative disease which increases itself gradually over time. Mainly the genetic factors which lead to Neurodegenerative disorders are categorized into pharmacy co-dynamics and pharmacokinetics.

Speech Articulation factors: Due to the condition associated with Parkinson's disease (rigidity and bady kinesia), some speech-language pathology such as voice, articulation and swallowing alterations are found. There are various ways in which Parkinson's disease (PD) might affect the individual. The voice get breathy and softer

The person finds difficulty in finding the right words due to which speech becomes slower.

2. LITERATURE SURVEY

"**Diagnosis of Parkinson's disease using Artificial Neural network" by Anila M and Dr G Pradeepini**

The main objective of this paper is that the detection of the disease is performed by using the voice analysis of the people affected with Parkinson's disease. For this purpose, various machine learning techniques like, Naïve Bayes, XG Boost are used to classify the best model, error rates are calculated, and the performance metrics are evaluated for all the models used. The main drawback of this paper is that it is limited to ANN with only two hidden layers. And this type of neural networks with two hidden layers are sufficient and efficient for simple datasets. They used only one technique for feature selection which reduces the number of features.

"Machine Learning-based Approaches for Prediction of Parkinson's Disease" by **Arvind Kumar Tiwari**

In this paper, minimum redundancy maximum relevance feature selection algorithms were used to select the most important feature among all the features to predict Parkinson diseases. Here, it was observed that the random forest with 20 number of features selected by minimum redundancy maximum relevance feature selection algorithms provide the overall

ISSN: 2366-1313

accuracy 90.3%, precision 90.2%, Mathews 12 correlation coefficient values of 0.73 and ROC values 0.96 which is better in comparison to all other machine learning based approaches such as bagging, boosting, random forest, rotation forest, random subspace, support vector machine, multilayer perceptron, and decision tree based methods. There are some limitations to this paper namely:

They used the validation set only to investigate the model performance during the training and this reduced the number of samples in the training set .

RNN training is too slow and this is not flexible in practice work.

Disconnecting and resource exhaustion: working with cloud services like Google Collaborators causes many problems like disconnecting suddenly. And because it is shareable service by the world zones, this leads to resource exhaustion error many times.

3. 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 – It 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.

DRAWBACKS OF EXISTING SYSTEM:

TIME TAKING: In the existing system the multiple scan must be taken in order to detect the disease.so, the existing system is a time taking procedure.

DEPENDENCEY ON SCANNING: The existing system is quite dependent on SPECT, MRI and PET scan.

NOT SO ACCURATE: The accuracy of the existing system is not up to the mark.

Thus, Existing System is not effective in early prediction and accurate medicinal diagnosis to the affected people.

4. PROPOSED SYSTEM:

The machine learning model can be implemented to significantly improve diagnosis method of Parkinson disease. In this study it indicates that the ensemble techniques XG- Boost classification (Extreme gradient boosting) algorithm achieved the high test accuracy rate (95%) compared to other classification algorithm.

This architecture diagram describes the high-level overview of major system components and important working relationships. It represents the flow of execution.

Speech Dataset 2. Pre-processing data 3. Training data 4. Apply Machine Learning Algorithms.

ISSN: 2366-1313

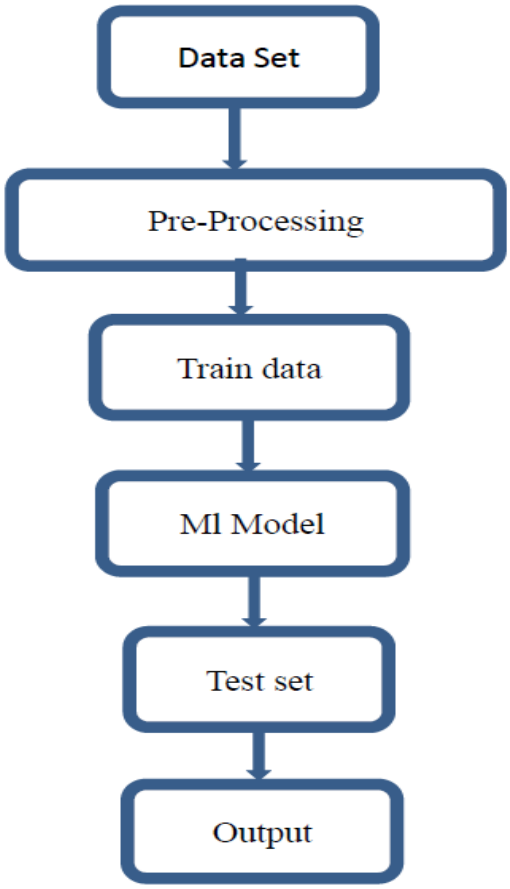

Fig: Proposed System

5. IMPLEMENTATION

There are several modules that could be included in parkinson's disease project. some of the modules include:

MODULE 1: Data Collection
MODULE:2: Preprocessing:
MODULE 2: Training and testing of data.
MODULE 3: Apply XGBoost algorithm.
MODULE 4: Cod completion.

Parkinson's disease is a progressive disorder that affects the nervous system and the parts of the body controlled by the nerves. Symptoms are also not that sound to be noticeable. Signs of stiffening, tremors, and slowing of movements may be signs of Parkinson's disease.

But there is no ascertain way to tell whether a person has Parkinson's disease or not because there are no such diagnostics methods available to diagnose this disorder.

Parkinson Disease Prediction using Machine Learning in Python Importing Libraries and Dataset

Python libraries make it very easy for us to handle the data and perform typical and complex tasks with a single line of code.

Pandas – This library helps to load the data frame in a 2D array format and has multiple functions to perform analysis tasks in one go.

NumPy – NumPy arrays are very fast and can perform large computations in a very short time.

Matplotlib/Seaborn – This library is used to draw visualizations.

Sklearn – This module contains multiple libraries having pre-
Implemented functions to perform tasks from data preprocessing to model development and evaluation.

XG-Boost – This contains the extreme Gradient Boosting machine learning algorithm which is one of the algorithms which helps us to achieve high accuracy on predictions.

Imblearn – This module contains a function that can be used for handling problems related to data imbalance.

The dataset we are going to use here includes 755 columns and three observations for each patient. The value's in these columns are part of some other diagnostics which are generally used to capture the difference between a healthy and affected person. Now, let's load the dataset into the panda's data frame.

Data Cleaning:

The data which is obtained from the primary sources is termed the raw data and required a lot of preprocessing before we can derive any conclusions from it or to some modeling on it. Those preprocessing steps are known

ISSN: 2366-1313

as data cleaning and it includes, outliers removal, null value imputation, and removing discrepancies of any sort in the data inputs.

These many features only indicate that they have been derived from one another or we can say that the correlation between them is quite high. In the below code block a function has been implemented which can remove the highly correlated features except for the target column.

So, from a feature space of 755 columns, we have reduced it to a feature space of 287 columns. But still, it is too high as the number of features is still more than the number of examples or data points. Reason behind this statement is the same as that behind the curse of dimensionality problem as the feature space grows the number of examples required to generalize on the dataset becomes difficult and the model's performance decreases.

So, let's reduce the feature space up to 30 by using the **chi-square**

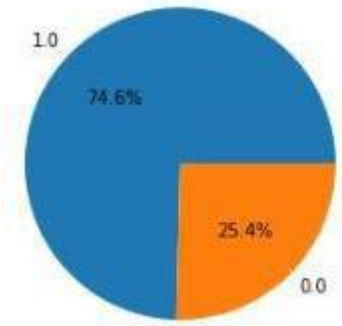

Fig: Pie chart for the distribution of the data within two class

Oh! the data imbalance. We will have to deal with this problem otherwise the model trained on this dataset will have a harder time predicting positive classes which is our main objective here.

Model Training:

Now we will separate the features and target variables and split them into training and the testing data by using which we will select the model which is performing best on the validation data.

Handling the data imbalance problem by using the over-sampling method on the minority class.

The dataset has been already normalized in the data cleaning step we can directly train some state-of-the-art machine learning models and compare them which fit better with our data.

Model Evaluation:

From the above accuracies, we can say that Logistic Regression and SVC() classifier perform better on the validation data with less difference between the validation and training data. Let's plot the confusion matrix as well for the validation data using the Logistic Regression model.

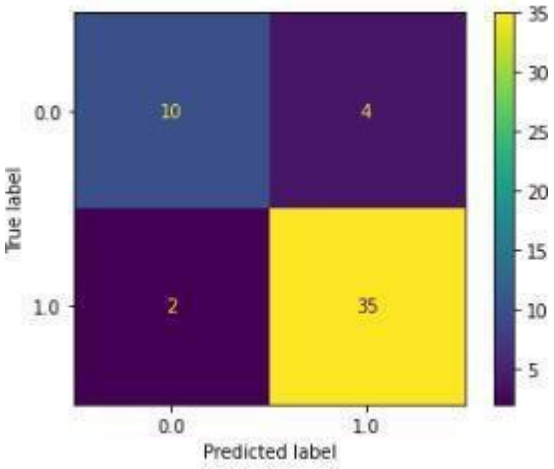

Fig: Confusion matrix for the validation data

Overall ,these modules can help develop an accurate and effective Parkinson's disease prediction system that can help diagnose patients early, monitor symptom progression ,and provide more personalized treatment plans.

ISSN: 2366-1313

7. SYSTEM ARCHITECTURE

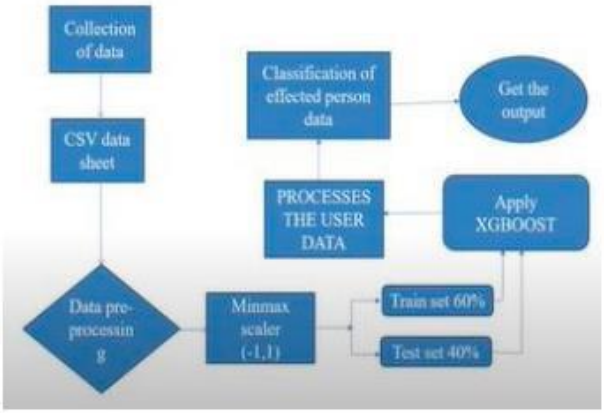

Fig: System Architecture

6. SCREEN SHORT

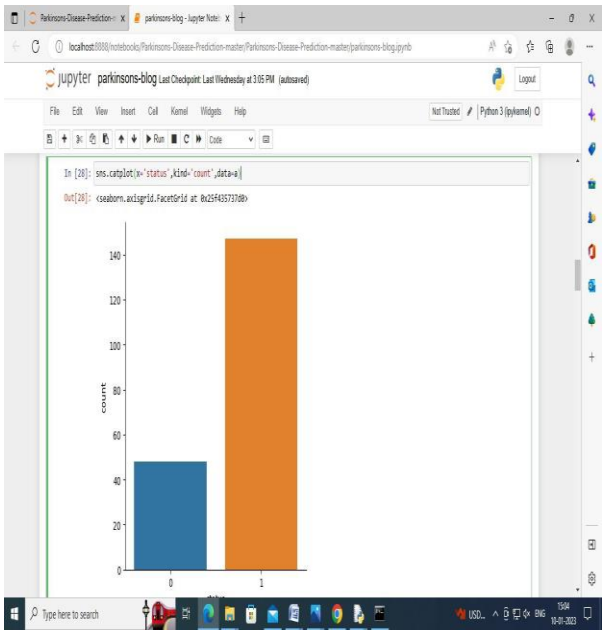

Fig : Screen Short

7. CONCLUSION

In conclusion, Parkinson's disease prediction projects are an important and promising area of research in the field of healthcare. Through the use of machine learning algorithms and data analysis techniques, i.e. XG Boost. This projects have the potential to improve the accuracy of diagnosis, enable personalized treatment plans, monitor disease progression over time, provide remote monitoring options, and help researchers better understand the disease and its causes. With further development and implementation, Parkinson's disease prediction projects have the potential to significantly improve patient outcomes and ultimately lead to the development of more effective treatments and a cure for this debilitating disease. This project gives above 90% accuracy.

FUTURE SCOPE

In future, these models can be trained with different datasets that have best features and can be predicted more accurately. If the accuracy rate increases, it can be used by the laboratories and hospitals so that it is easy to predict in early stages. This models can be also used with different medical and disease datasets. In future the work can be extended by building a hybrid model that can find more than one disease with an accurate dataset and that dataset has common features of two diseases.

REFERENCES

1.A. Ozcift, “SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease” Journal of medical systems, vol-36, no. 4, pp. 2141-2147, 2012.

1.Arvind Kumar Tiwari, “Machine Learning based Approaches for Prediction of Parkinson’s Disease” Machine Learning and Applications: An International Journal (MLAU) vol. 3, June 2016.

2.CarloRicciardi, et al, “Using gait analysis’ parameters to classify Parkinsonism: A data mining approach” Computer Methods and Programs in Biomedicine vol. 180, Oct. 2019.

3.Dr. Anupam Bhatia and RaunakSulekh, “Predictive Model for Parkinson’s Disease

ISSN: 2366-1313

through Naive Bayes Classification" International Journal of Computer Science & Communication vol-9, Dec. 2017, pp. 194-202, Sept 2017 - March 2018.
4.Dr. R.GeethaRamani, G.Sivagami, ShomonaGraciajacob "Feature Relevance Analysis and Classification of Parkinson's Disease TeleMonitoring data Through Data Mining" International Journal of Advanced Research in Computer Science and Software Engineering, vol-2, Issue 3, March 2012.
5.DraganaMiljkovic et al, "Machine Learning and Data Mining Methods for Managing Parkinson's Disease" LNAI 9605, pp. 209-220, 2016.
6.FarhadSoleimanianGharehehopogh, PeymenMohammadi, "A Case Study of Parkinson's Disease Diagnosis Using Artificial Neural Networks" International Journal of Computer Applications, Vol-73, No.19, July 2013.
7.Heisters. D, "Parkinson's: symptoms, treatments and research". British Journal of Nursing, 20(9), 548–554. doi:10.12968/bjon.2011.20.9.548, 2011.
8. M. Abdar and M.Zomorodi-Moghadam, "Impact of Patients' Gender on Parkinson's disease using Classification Algorithms" Journal of AI and Data
Mining, vol-6, 2018.
9. M. A. E. Van Stiphout, J. Marinus, J. J. Van Hilten, F. Lobbezoo, and C. De Baat, "Oral health of Parkinson's disease patients: a case-control study" Parkinson's disease, vol- 67 2018, Article ID 9315285, 8 pages, 2018.
10. Md. Redone Hassan, et al, "A Knowledge Base Data Mining based on Parkinson's Disease" International Conference on System Modelling& Advancement in Research Trends, 2019.