2021 2nd International Conference for Emerging Technology (INCET)
Belgaum, India. May 21-23, 2021

Prediction of Parkinson's disease and severity of the disease using Machine Learning and Deep Learning algorithm

Dr. Pooja Raundale
Prof. and Head, MCA ,
Sardar Patel Institute of Technology
Mumbai, India
pooja@spit.ac.in

Chetan Thosar
Student. MCA Department
Sardar Patel Institute of Technology
Mumbai, India
chetan.thosar@spit.ac.in

Shardul Rane
Student. MCA Department
Sardar Patel Institute of Technology
Mumbai, India
shardul.rane@spit.ac.in

2021 2nd International Conference for Emerging Technology (INCET) | 978-1-7281-7029-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/INCET51464.2021.9456292

***Abstract*—Parkinson's disease is a neurodegenerative disease which worsens over time. People have trouble vocally, writing, strolling, or completing other simple tasks when dopamine-generating neurons in parts of the brain become impaired or expire. These symptoms worsen over time, increasing the severity of the condition in patients. We have suggested a methodology in this article for the prediction of Parkinson's disease severity using deep neural networks on UCI's Parkinson's Telemonitoring Vocal Data Set of patients. We have created a neural network to predict the severity of the disease and a machine learning model to detect the disorder. Classification of Parkinson's Disease is done by Neural network, Random Forest Classifier.**

Keywords— Parkinson's Disease, Prediction, Deep Learning, Machine Learning, XGBoost, Neural Network, Tensorflow & Keras

I. Introduction

Parkinson's disease (a neurodegenerative disorder) that causes the patients' motor abilities to degrade over time due to the damage caused to the dopamine-generating brain cells. Shaking, trouble moving, behavioral disorders, dementia, and depression are some of the results of this disorder. The primary motor conditions are referred to as ”Parkinsonism,” or a ”Patient with Parkinson's Disease.” One of the most common symptom that can be recognized by studying the patients' voice data is changes in their voice. The patient's speech stutters and becomes increasingly impacted as the disease progresses. Deep learning has risen in importance as a method for analysing unstructured data such as speech and audio signals. Multiple layers of neurons are often used in deep neural networks, these layers are stacked as a single unit for classification and feature selection models. Deep learning is being used in this paper to classify the patient's voice data into ”extreme” and ”not severe” categories. The two UPDRS (Unified Parkinson's Disease Rating Scale) scores - total UPDRS and motor UPDRS - were used as assessment criteria in this study. The motor UPDRS assesses the patient's motor capacity in the scale of 0-108, while the total UPDRS assesses the patient's overall ability and its score range from 0-176.

II. Literature Review

Much studies have been done to predict Parkinson's sickness, but less work has been reported to predict the severity of Parkinsonism. Various machine learning models were used in these works. In a study conducted by Das et al. [1] on the use of different classification strategies in the diagnosis of Parkinson's disease (PD), neural networks were found to be the most systematic classifier and regression algorithm compared to machine learning regression algorithm and decision tree. There are numerous study done to create a prediction classifier for Parkinson's disease. In most scientific papers, features extracted from voice signal have been used to forecast the severity of PD. Genain et al.[2] used Bagged Decision Trees to estimate PD severity from patient audio recordings and found an improvement in accuracy of 2%. Maleket al.[3] used the 40-feature dataset and recognised the 9 best features while using LLBFS (Local Learning Based Feature Selection) for the class to partition PD subjects into four classes (Healthy, Early, Intermediate and Advance), on the basis of their UPDRS score. In Seeja K.R. et al[4] a two class classifying neural network has been created using acoustic dataset.The neural network created has multiple input and multiple output. The neural network is used for classification rather than regression. The classification produces an output with accuracy of 79%. This research paper has been used mainly referred in the study. The objective of this research was to predict if the victim has Parkinson's disease or not using a keyboard typing test dataset, and if he have then we also predicted the severity of the Parkinson' disease using voice impairment of the patient . For the detection part we have referred a study [5]”High-accuracy detection of early Parkinson's Disease using multiple characteristics of finger movement while typing”. In this study different machine learning algorithms have been implemented and compared to detect the Parkinson's disease and find the most suitable model for it.

A study named ”Dynamical Learning and Tracking of Tremor and Dyskinesia From Wearable Sensors” [6] implemented and tested several evolving machine-learning algorithms capable of tracking the prevalence and intensity of tremor and dyskinesia at 1-second resolution by analysing the signals gathered from Parkinson's disease (PD) Patients wearing a significant subset of sensors with a 3-D accelerometric (ACC) as well as a surface electromyographic (EMG) modes. The algorithms by 8 PD patients and 4 healthy subjects who completed unrehearsed and unbridled daily living activities in a house environment. Results show that the performance of our machine learning algorithms against independent clinical citations of disorder prevalence and severity shows that, despite their different approaches to dynamic pattern classification, dynamic neural networks (DNN), dynamic support vector machines (DSVM) and hidden Markov models (HMM) were similarly efficient in maintaining dynamic pattern monitoring failure rates below 10%. In this research paper we create a machine learning technique to detect the Parkinson's disease and a deep learning Neural Network for foreseeing severity of the disease.

III. Methodology

The proposed system (Fig(1)) classifies process from data collection to model selection in a flowchart.

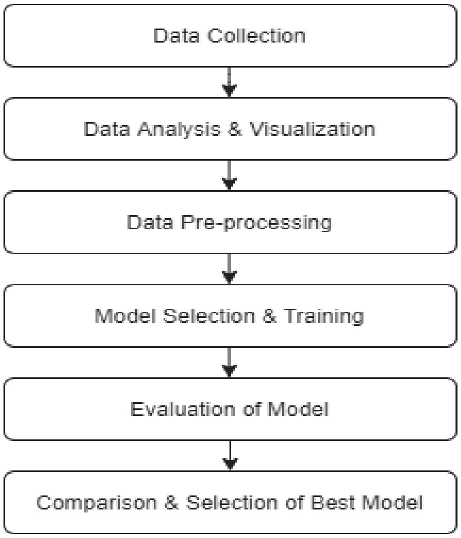

Data Collection
Data Analysis & Visualization
Data Pre-processing
Model Selection & Training
Evaluation of Model
Comparison & Selection of Best Model

Fig 1. Proposed Methodology.

A. Data Collection

For the detection of Parkinson's disease and to predict the severity of the disease , two different datasets are used. The dataset[7] for detection, it contains the typing data of the participants. Participants from the United States, Canada, the United Kingdom and Australia checked the website of the project and agreed to participate in the study. The research was permitted by the Human Research Ethics Committee of Charles Sturt University, Australia, protocol number H17013.

Almost all data file gathered includes the timing of the typing activity when participants utilised their numerous Windows applications (e.g. email, word processing, web searches, etc.). The keystroke acquisition software ('Tappy') generated the timing accuracy of the key presses and only produced timestamps after a few milliseconds.

The provided data files are divided into two sub-folders:
Folder 1: Archived users, which contains information on the participants' details (gender, year of diagnosis, whether the participant has tremors, etc.)
Folder 2: Tappy Data, which contains keystroke statistics from specific participants (hold time, current hand, previous hand, etc.)
The columns in the combined dataset are - 'BirthYear', 'Gender', 'Parkinsons', 'Tremors', 'DiagnosisYear', 'Sided', 'UPDRS', 'Impact', 'Levadopa', 'DA', 'MAOB', 'Other'

The dataset[8] for the severity test was developed by Athanasios Tsanas and Max Little of Oxford University, in close cooperation of 10 medical centres in the US and Intel Corporation, which created a telemonitoring device to document speech signals. The experimental procedure used a variety of linear and nonlinear regression algorithms to predict the symptom score for Parkinson's disease on the UPDRS scale.
The columns in telemonitoring dataset includes:
subject# - Integer that uniquely identifies each subject
age - Subject age
sex - Subject gender '0' - male, '1' - female
test_time - Time since recruitment into the trial. The integer part is the number of days since recruitment.
motor_UPDRS - Clinician's motor UPDRS score, linearly interpolated
total_UPDRS - Clinician's total UPDRS score, linearly interpolated
Jitter- Several measures of variation in fundamental frequency
Shimmer- Several measures of variation in amplitude
NHR,HNR - Two measures of ratio of noise to tonal components in the voice
RPDE - A nonlinear dynamical complexity measure
DFA - Signal fractal scaling exponent
PPE - A nonlinear measure of fundamental frequency variation

B. Data Analysis & Visualization

The type and kind of data collected plays a key role in deciding which algorithm to use. So it is very important to know data. The author understands the data through data visualization technique Python libraries like Matplotlib, Seaborn etc. Visualization helps detect relevant relationships between variables or class.

C. Data Processing

The data that we collected, can't be used directly for performing the analysis as this data may be unorganized and may contain a lot of missing vales, duplicates, noisy data and extreme values. The author performs data pre-processing and tuning that involves dealing with missing values and these values are replaced by 'NaN', removing the duplicates, correcting outliers if present and handling data, etc. This phase actually involves removing useless and incomplete data. Also to avoid overfitting, the author reduces the dimension of data that involves reducing the count of features present in the dataset.

D. Selection & Training the Model

This step serves as a baseline. Since the output of the prediction of Parkinson's disease is a class, it is a classification problem. For the detection part we are calculating the UPDRS value and the concluding the severity of the disease, so its a regression problem. Study is carried on the below mentioned classifier and regression algorithm which are as follows:

1) XGBoost
2) Neural Network for regression

In machine learning, the dataset is partitioned into three subsets namely training set, testing set, and validation set. The author trains the classifiers using 'training dataset' and tune the parameters using 'validation dataset'.The performance testing of the classifier is checked on the previously unseen 'test dataset'.
Usually the dataset is split in the ratio 8:2 as train to test dataset.

Train Data
80%
Test Data
20%
Full Dataset

Fig 2. Train Test Split.

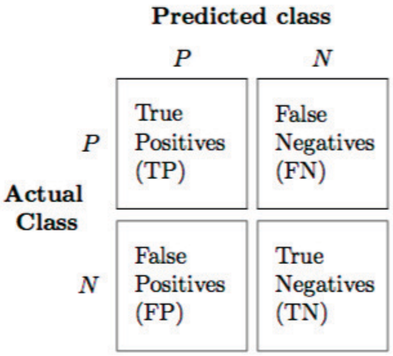
E. Evaluation of Model

There are various ways to check the performance of the machine learning algorithms. The author evaluates the model based on its accuracy and the confusion matrix. A confusion matrix is a summary of prediction in table format that is used to describe the performance of the model. It is a table with combination of predicted and actual values. Following is the confusion matrix:

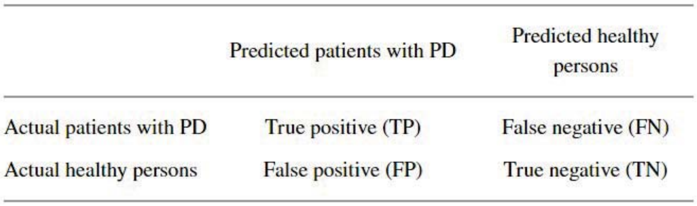
IV. Results

The complete dataset for detection of Parkinson's disease comprised 217 participants, however, only some of these included the following analysis, including:

- The records with minimum 2000 keystrokes.
- Out of those with PD, just a record of 'mild' severity (since the study was about the detection of Parkinson's Disease in its early stage).
- Those who do not end up taking levodopa (Sinemet® and the like) to remove any effect of the medication on the typing test.



		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)



	Predicted patients with PD	Predicted healthy persons
Actual patients with PD	True positive (TP)	False negative (FN)
Actual healthy persons	False positive (FP)	True negative (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

This resulted in a group of 53 participants (including both PD and non-PD). Other columns of the dataset included birth date, Diagnosis year, Gender and the medicines the patients have been prescribed.
This dataset for classifying severity of Parkinson's disease consists of various biomedical audio measurements from 42 people having early-stage Parkinson's disease who have been asked for a six-month telemonitoring device for remote symptom progression monitoring.

The table's columns include subject number, subject age, subject gender, time interval from baseline recruitment date, motor UPDRS, total UPDRS, and 16 biomedical voice measurements. Each record contains one of 5,875 audio recordings from the selected participants. The primary goal of the data is to predict the motor and total UPDRS scores ('motor UPDRS' and 'total UPDRS') of the 16 voice metrics.

The generated information resulted in CSV data (ASCII). The rows of the CSV file have an instance corresponding to one voice/audio recording. Approximately 200 records are documented per patient as well as the patient's subject number is identified in the first column.

The above figure(Fig. 3) shows the distrbution of age among Parkinon's patients and healthy people. The plot determined

Fig(3) shows age group suffering from Parkinson’s and healthy people

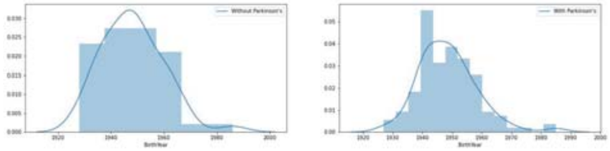


Fig 3. Age group distribution of people suffering from Parkinson’s and healthy people

the fact that most of the Parkinson’s patient are between the age of 50 to 60 years of age. So, the likelihood for a young person to have Parkinson’s is negligible. This information acted an important factor while creating the model.

Fig(4) shows the number of people suffering from Parkinson’s in each gender.

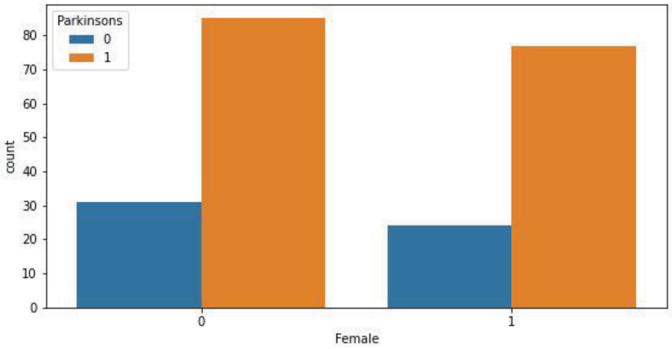


Fig 4. Gender Distribution

The gender distribution bar plot(Fig. 4) helped us to determine that the number of females having Parkinson’s Patient is more than the number of males suffering from Parkinson’s. As stated in the data collection stage that this dataset has was collected for a different study where participants were chosen randomly, we can determine that this distribution shows a trend which will help our study.

The above boxplots in Fig.5, visualize distributions of different time data (hold time, latency time, and flight time) between participants with and without Parkinsons’s. Each subplot contains data in a specific typing switch type–for example, the top left subplot contains typing data when participants go from a left-hand key to another left-hand key (denoted as LL above the subplot), while the top right one contains data when participants switch from a left-hand key to a space (LS). There are 9 such switches in total - LL, LR, LS ,RL, RR, RS, SL, SR ,SS
where,
L is left key

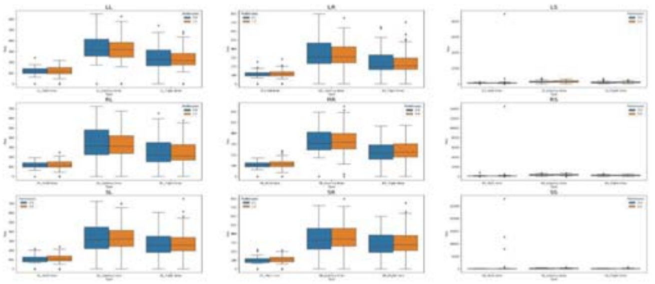


Fig 5. Hold Time, Latency Time, Flight Time from one key to another while typing.

R is right key
S is Space bar.
The reason for plotting these different times is that they act as important factors while distinguishing a healthy person from a Parkinson’s patient.

The optimal features were analyzed to see their significance by the technique of feature importance. A correlation graph was created to determine features affecting the severity of the Parkinson’s disease.

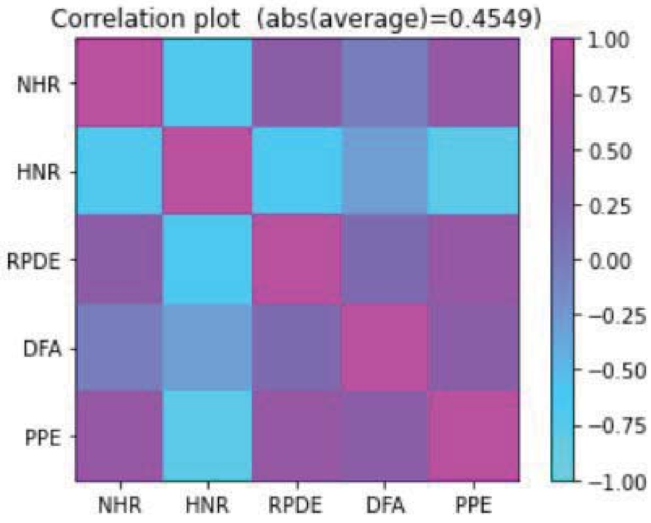


Fig 6. A correlation plot determining important features required for predicting severity of Parkinson’s Disease.

The below correlation plot in Fig 6. has been used to determine the important features for the feature selection step. Along with EDA methods like correlation plot, PCA was also used to select important features that largely affect our prediction. We concluded that all the voice metrics including Jitter(frequency variation from cycle to cycle of the sound wave) , Shimmer(amplitude variation from cycle to cycle of the sound wave), NHR(Noise to Harmonic Ratio), HNR(Harmonic to Noise Ratio) etc. are all extremely important and non negligible features for a proper result for our study.

Two models were used in this study one for determining if the patient has Parkinson's Disease or not and the other one determined severity of Parkinson's Disease.
For detection purpose, XGBoost was used

TABLE I: Evaluation Metrics

Algorithm	Accuracy
XGBoost	0.95
Artificial Neural Network	0.85

V. Conclusion

Detection of Parkinson's Disease

1) The research area for Parkinson's Disease is significant, early stage detection of it can improve patient's health
2) This solution was capable of differentiating among early stage Parkinson disease subjects and controls with a tolerance of 92 to 100 %, a specificity of 95 to 100 % and an AUC(Area Under Curve) in the range of 0.97 and 1.00.
3) It was found that Parkinson's disease was detected positive in people above the age of 55 years.
4) According to the study, females are more likely to have Parkinson's than males

Fig(7) Comparison of proposed model with other popular models used in referred studies

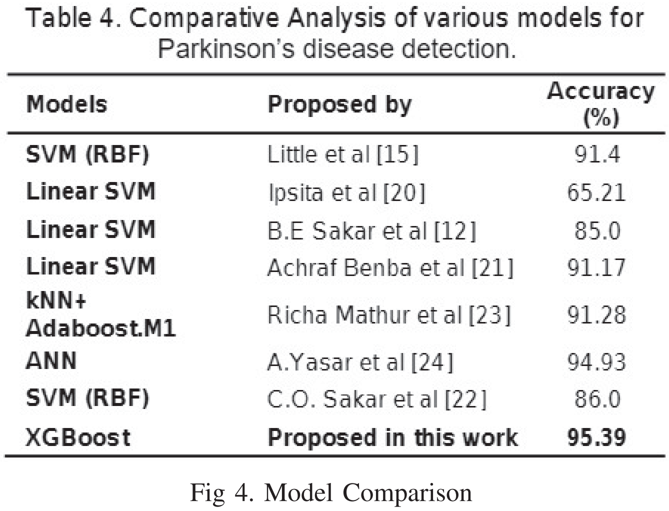


Table 4. Comparative Analysis of various models for Parkinson's disease detection.

Models	Proposed by	Accuracy (%)
SVM (RBF)	Little et al [15]	91.4
Linear SVM	Ipsita et al [20]	65.21
Linear SVM	B.E Sakar et al [12]	85.0
Linear SVM	Achraf Benba et al [21]	91.17
kNN+ Adaboost.M1	Richa Mathur et al [23]	91.28
ANN	A.Yasar et al [24]	94.93
SVM (RBF)	C.O. Sakar et al [22]	86.0
XGBoost	**Proposed in this work**	**95.39**

Fig 4. Model Comparison

Predicting Severity of Parkinson's Disease

1) Neural Network is the most efficient algorithm for the study
2) All audio features are important dimensions for the prediction.
3) Motor and total UPDRS values determine the severity of disease

Thus it can be concluded that the prognostication of Parkinson's Disease is very complex and is dependent on many variable factors which keeps on changing. If the features are properly selected we can get an optimized and efficient model which can get proper severity and extent of the spread of disease in the patient.

VI. Future Scope

The study was done to detect and severity of Parkinson's disease using a single model for each purpose. The study can be further extended by using other models and comparing the results to find the most optimised and efficient models for detection of disease and to determine the severity of the disease in the patient.

VII. Acknowledgment

This research paper was supported by the Sardar Patel Technology Institute. We like to extend our gratitude towards the faculty members whose comments have greatly influenced the implementation of the work.

We thank our faculties who provided insight and expertise that greatly assisted the research. Also we thank the students of Sardar Patel Institute of Technology and also our friends for helping us with our survey.

References

[1] Das R. "Comparison of multiple classification methods for diagnosis of Parkinson disease". *Expert Systems With Applications*, 37:1568–1572, 2010.

[2] Genain N, Huberth M, and Vidyashankar R. "Predicting Parkinson's Disease Severity from Patient Voice Features." 2014.

[3] Benmalek E, Elmhamdi J, and Jilbab A. "UPDRS tracking using linear regression and neural network for Parkinson's disease prediction." *International Journal Of Emerging Trends and Technology In Computer Science*, 4:189–193, 2015.

[4] Seeja K.R., Srishti Grover, Saloni Bhartia, Akshama, and Abhilasha Yadav. "Predicting Severity Of Parkinson's Disease Using Deep Learning". *Procedia Computer Science*, 132:1788–1794.

[5] Warwick R. Adams. "High-accuracy detection of early Parkinson's Disease using multiple characteristics of finger movement while typing". 2017.

[6] Cole B, Roy S, De Luca C, and Nawab S. "Dynamical Learning and Tracking of Tremor and Dyskinesia From Wearable Sensors." *IEEE Transactions On Neural Systems And Rehabilitation Engineering*, 22:982–991, 2018.

[7] Tappy keystroke dataset.

[8] Parkinson's telemonitoring dataset.

[9] Google colaboratory.

[10] Keras api reference.

[11] Python machine learning course.

[12] Deep learning course.

[13] Multiple output neural networks.