International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.202

Predict the Beginning and Course of PD with Machine Learning and Deep Learning Algorithms

J. Ezhilarasi
Dept. of Computing Technologies
SRMIST,
Kattankulathur, Chennai, TamilNadu, India
ej1362@srmist.edu.in

Dr.T.Senthil kumar
Associate Professor Department of Computing Technologies
SRMIST,
Kattankulathur
Chennai, India,
senthilt2@srmist.edu.in

***Abstract*—A progressive neurological disorder, Parkinson's disease(PD).People have trouble speaking, writing, walking, and doing other primary duties when particular brain regions' dopamine-producing neurons are harmed or die. These signs worsen with time, variable degrees of sickness aggravation in each person. In this research, we provide a method for calculating the Parkinson's Speech dataset from GitHub, deep neural networks were used to determine the severity of the condition. To forecast the severity of the illness and identify the condition, we have developed neural networks and machine learning algorithms. Parkinson's disease is categorized using both a random forest classifier and a neural network.**

Keywords—Parkinson's Disease (PD), UPDRS (unified Parkinson's disease rating scale, EMG, ACC, DN, RNN.

I. INTRODUCTION

The dopamine-producing brain cells are gradually destroyed by Parkinson's disease, a neurological illness., gradually affecting the sufferers' ability to move. This condition's symptoms include trembling, trouble moving, strange behaviour, dementia, and melancholy. The term "Parkinsonism" or "a Patient with Parkinson's Disease" is used to describe the main motor issues. One of the most frequent symptoms that may be picked up by listening to the patient's speech data is voice changes. The patient's speech starts to stutter as their condition worsens. As a method of analysing unstructured data, such as speech and voice signals, deep learning is gaining popularity.To create feature selection and classification models, deep neural networks usually use many layers of neurons. The speech data of the patient is categorised in this article into "extreme" and "not severe" categories using deep learning.

The motor and overall UPDRS (Unified Parkinson's Disease Rating Scale) scores were used in this study as comparison points. The total UPDRS has a score range of 0 to 176 while the motor UPDRS has a score range of 0 to 108 and examines the patient's overall abilities.. To measure the frequency and severity of tremors and dyskinesia in Parkinson's disease (PD) patients at a resolution of 1 second, algorithms are required that can analyze data from a large number of 3-D accelerometric (ACC) and surface electromyographic (EMG) sensors.

16 PD patients and 8 normal participants executed the algorithms in a home environment while doing free-form, spontaneous daily chores. The study demonstrated that dynamic pattern monitoring failure rates could be kept at 10% by utilizing dynamic support vector machines (DSVM), hidden Markov models (HMM), and dynamic neural networks (DNN). The effectiveness of these machine learning algorithms was confirmed by comparing them to independent clinical evidence of sickness prevalence and severity. In this study, a machine learning approach is developed to recognizee Parkinson's disease as well as a deep learning neural network to gauge the severity of the condition.

II. LITERATURE REVIEW

- *Selecting a Template*

The research on Parkinson's disease is extensive. even if the severity of Parkinsonism has received less attention. In these tests, several machine learning models were employed. Comparing neural networks to decision trees and machine learning algorithms, however, it was shown that they are the most systematic classifier and regression tool. In a research by Das et al..

1. Many classification algorithms were employed to make the PD diagnosis. Several studies have been conducted in an effort to develop a classifier that can forecast Parkinson's disease. The majority of research have employed variables gleaned from speech signals to predict the severity of PD.
2. Boosted accuracy while determining the severity of Parkinson's disease by using bagged decision trees. from patient audio recordings. Based on their UPDRS scores, Malekt et al.'s
3. There are four categories for PD patients: early, moderate, and advanced —used a dataset of 40 features. Using LLBFS, they identified the top nine characteristics for each class. (Local Learning Based Feature Selection).Using an audio dataset, Seeja K.R. et al.
4. Created a neural network for two-class classification.There were many different inputs and outputs used to construct the neural network.

DOI: rp-9788770040723.202

Regression isn't the neural network's primary objective; classification is. The outcome of the categorization has a 79% accuracy rate. The research paper for this study received the most mentions. The goal of this study was to use data from a keyboard typing test to identify whether or not the victim had Parkinson's disease. If so, we assessed how much the patient was utilising that knowledge to prevent them from communicating. For our detection investigation, we built upon a paper titled "High-accuracy identification of early Parkinson's disease utilising several features of finger movement during typing." In this study, multiple machine learning algorithms have been developed and evaluated in order to recognise Parkinson's disease and determine the optimal model for it.

III. Methodology

The approach suggested in Figure 1 arranges the procedures in a sequential manner, beginning with the collection of data and concluding with the selection of the model.

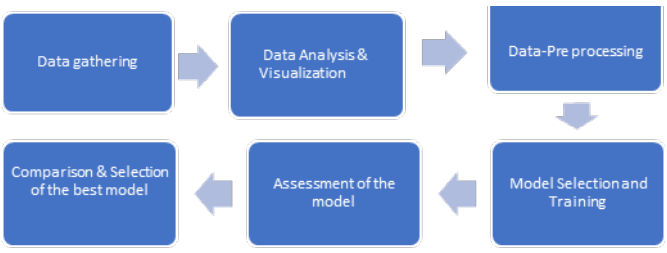

Fig 1. Recommended approach

- *Data Gathering*

Using two separate datasets, Parkinson's disease is identified and the degree of the problem is predicted. The dataset[7] contains the participants' typing data for detection. Australia, Canada, the United Kingdom, and the United States selected to participate in after reviewing the project website, the research. Under protocol number H17013, the Charles Sturt University Human Research Ethics Council in Australia approved the experiment. The duration of the typing activity when participants used their various Windows programmes is included in almost all data files that have been collected. (such as email, word processing, web searches, etc.). The timing accuracy of the keystrokes was achieved by the keystroke capture programme (called "Tappy"), which only delivered timestamps after a brief period of time.

There are two subgroups in the data files that are provided: The participants' personal information is contained in Group *1:* User Archives. (gender, the participant's diagnosis year, whether they have tremors, etc.) Tappy Data in Group 2 offers keystroke counts from certain individuals. (hold time, current hand, previous hand, etc.) The merged dataset consists of the columns "BirthYear," "Gender," "Parkinson," "Tremors," "DiagnosisYear," "Sided," "UPDRS," "Impact," "Levadopa," "DA," "MAOB," and "Other."

The experimental method predicted the UPDRS score for Parkinson's disease symptoms using a number of linear and nonlinear regression models.

In the Parkinson's speech dataset, there are the following columns: Each subject's age is indicated by an integer called subject#. The subject's age, sex, and gender are noted. The number "M" for a male and "F" for a female denotes the gender of the topic. The days since recruitment are shown in the integer component. Motor total interpolated motor UPDRS score for the clinician Linear interpolation of the clinical professional's overall UPDR S score is known as UPDRS. There are several techniques to monitor changes in fundamental frequency, and they are all referred to as "jitter.". The ratio of the voice's noise to tonal components may be measured using the terms NHR and HNR. There are several amplitude fluctuation measurements that make up shimmer. RPDE is a sign of nonlinear dynamic complexity. Known as the DFA, the basic frequency fluctuation can be assessed nonlinearly using the signal fractal scaling exponent.

- *Data Analysis & Visualization*

The selection of the algorithm is greatly influenced by the type and quality of the data gathered. Knowledge of data is therefore essential. The author can visualize the data to better understand it using Python programs like Matplotlib, Seaborn, and others. Finding relevant relationships between classes or variables is aided by visualization.

- *Data Processing*

The information we gathered might not be suitable for analysis since it may be chaotic, have many missing values, be duplicated, noisy, and have extreme values. The author handles data, deals with missing values, replaces missing numbers with "NaN," removes duplicates, fixes outliers if present, and does data pre-processing and tuning. Actually, this stage entails removing irrelevant and insufficient data. This study further reduces the dimension of the data by reducing the amount of features in the dataset in order to prevent overfitting.

- *Selection & Trianing the Model*

As a beginning point, this activity is taken. Since a class is the result of the Parkinson's disease prediction, it is a classification problem. Since we utilize the UPDRS result to estimate the disease's severity during the detection phase, there is a regression problem. The classifier and regression methodology used in the study are as follows:

  - XGBoost
  - Regression neural network

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

The three subsets of the dataset that are utilised for machine learning are the training set, testing set, and validation set. Using a "training dataset," the classifiers are trained, and the parameters are adjusted using a "validation

DOI: rp-9788770040723.202

dataset," according to the author. Utilizing an original "test dataset," the classifier's performance is assessed.

The train and test datasets are frequently split in a 7:3 ratio.

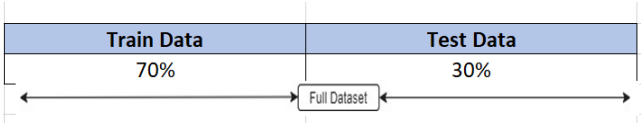

Fig 2. Train Test Split

- *Evaluation of Model*

The effectiveness of machine learning algorithms may be determined using a variety of techniques. Using the confusion matrix, the author evaluates the model's accuracy. A table-based summary of predictions called a confusion matrix is used to assess the model's efficacy. The table displays a combination of anticipated and actual numbers. Below is a matrix of misunderstandings:

IV. RESULT

- Only a portion of the 217 individuals in the dataset for the diagnosis of Parkinson's disease were employed in the study that followed, which includes
- Documents with a minimum of 2000 keystrokes.
- Only a record of "moderate" severity among those who have Parkinson's disease (since the research concentrated on the early detection of the condition).
- People who don't take levodopa (Sinemet® and related medications) to offset any negative effects on typing tests from the medicine.

		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)

There were 53 people in the group as a result. (comprised of PD and non-PD). The dataset also includes columns for the patients' birthdays, diagnosis years, genders, and medications.

In order to evaluate the severity of the illness, the 45 patients with Parkinson's disease in its early stages who volunteered to wear telemonitoring equipment for six months to follow the development of their symptoms are included in this dataset.

Subject number, subject age, subject gender, time since the first recruitment date, motor UPDRS, total UPDRS, and 16 biomedical voice measurements are the columns in the table. One of the 4,824 audio recordings that the chosen participants created is present on each CD. The data's main objective is to predict the 16 voice parameter's motor and overall UPDRS scores. (referred to as "motor UPDRS" and "total UPDRS" respectively).

CSV data were created using the information. (ASCII). A single speech or audio recording is represented by one occurrence per row in the CSV file. Each patient has 200 records total, with their subject number appearing in the first column.

The previous graphic displays the age distribution of both Parkinson's disease sufferers and healthy people. (Fig. 3). The strategy set forth. Fig. 3 depicts an age group with Parkinson's disease and vigorous individuals.

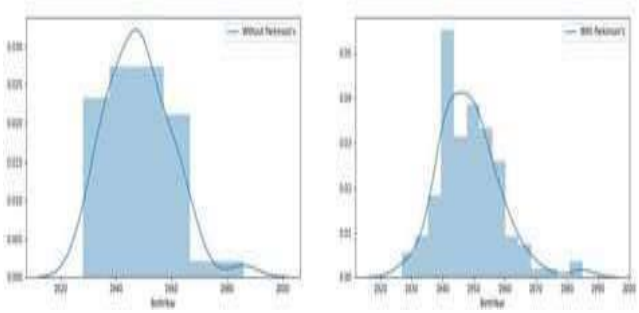

Fig 3. Age distribution of Parkinson's disease patients versus healthy individuals.

Parkinson's sufferers tend to be between the ages of 50 and 60. Therefore, there is a very low likelihood that a young individual would get Parkinson's. The model's development was significantly influenced by these facts. Fig.4 shows the gender-specific number of Parkinson's patients.

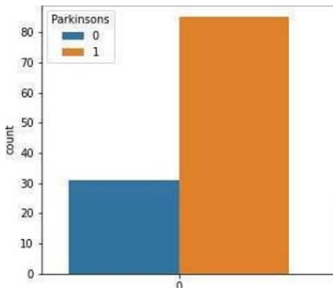

Fig 4. Gender Distribution

We were able to get to this conclusion since the gender distribution bar plot visual (Fig. 4) showed that more women than males had Parkinson's disease. Because it was made clear throughout the data collection stage that this information was gathered for various study with participants chosen at random, we may presume that this distribution shows a tendency that will be useful to our analysis.

Boxplots in Fig. 5 demonstrates how patients with and without Parkinson's disease have different distributions of various time data. (hold time, latency time, and flight time).

DOI: rp-9788770040723.202

The specifics of a particular sort of typing switch are covered in each subplot. In the upper left sub-plot, for instance (denoted as LL above the sub-plot), Typing data is displayed as participants switch from one left-handed key to another.

In contrast, typing data is displayed in the top right sub-plot when participants switch from a left-hand key to a space. (LS). These switches have the labels LL, LR, LS, RR, SL, SR, and SS. Fig. 5 shows the hold time, latency time, and flight time between each key when typing.

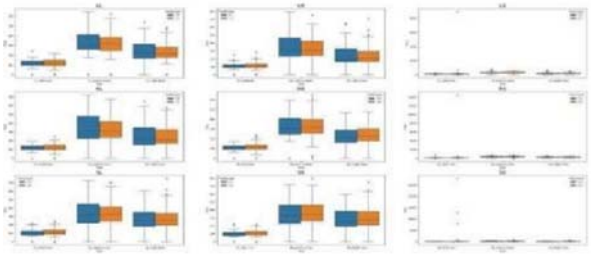

Fig 5. When typing, there are three things to consider: hold time, latency time, and flight time.

These numerous time periods were charted since they are important indicators when comparing a Parkinson's patient with a healthy person.

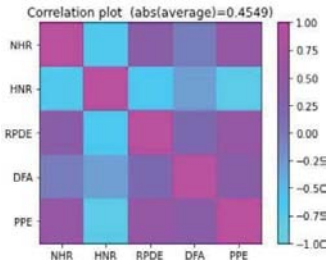

Fig 6 depicts a correlation map highlighting crucial elements required to anticipate the severity of Parkinson's disease

The feature importance approach was used to evaluate the most valuable characteristics and determine their significance. To illustrate the connections between a number of variables and the severity of Parkinson's disease, a correlation graph was made.

Figure 6 depicts a correlation map highlighting crucial elements required to anticipate the degree of Parkinson's disease severity.

The crucial traits for the feature selection stage have been identified using the correlation map displayed in above Fig. 6. Using correlation graphs in addition to EDA techniques like PCA, it was possible to identify important variables that have a significant impact on our forecast. We came to the conclusion that the Signal to Noise Ratio, among other features, is extremely significant and non-negligible for a suitable outcome for our investigation. This includes all voice measures, including Jitter (frequency fluctuation between sound wave cycles), Shimmer (amplitude variation between sound wave cycles), NHR and HNR (Noise to Harmonic Ratio) (Harmonic Noise Ratio). In this study, two models were utilized, one to assess whether a patient had Parkinson's disease or not, and the other to assess the degree of the condition. For detection purpose, Enhanced XGBoost was used.

Algorithm	Accuracy
XGBoost	0.97
Artificial Neural Network	0.84

V. Conclusion

Detection of Parkinson's Disease

- Parkinson's disease is an important subject for research since an early diagnosis might improve patients' health
- With tolerance ranges of 92 to 100%, specificity ranges of 94 to 100%, and an AUC in the range of 0.96 to 100%, this technique was able to differentiate between those with early-stage Parkinson disease and controls with an AUC (Area Under Curve) between 0.97 and 1.00.
- It has been observed that those over 65 had a higher likelihood of being diagnosed with Parkinson's disease.
- The study discovered that women had a higher risk of developing Parkinson's than males do.

Figure 7 compares the suggested model to other widely used models utilized in the research mentioned.

Table4. Comparative Analysis of various models for Parkinson's disease detection

Data Models	Proposed by	Accuracy (%)
SVM (RBF)	Little et al [15]	89.2
KNN+	Richa Mathur et al[23]	89.23
Adaboost.M1	Richa Mathur et al[23]	81.34
Linear SVM	Ipsita et al[15[	76.89
Linear SVM	B.E Sakar er al[14[	90.234
Linear SVM	Achraf Benha et al[17]	89.32
ANN	A.Yasae et al[34]	92.33
SVM (RBF)	C.O Sakar et al[21]	80.13
XGBoost	Proposed for this work	94.49

Fig 4. Machine Learning Model Comparison

Forecasting Parkinson's Disease Severity

- The neural network is the most competent method in this investigation.
- Each auditory component is an essential part of the forecast.
- The severity of the issue is determined by the motor and total UPDRS levels.

It follows that the Parkinson's disease prognosis is exceedingly complex and depends on a wide range of variable factors that are always changing. If the qualities are properly picked, we might be able to create a model that is

DOI: rp-9788770040723.202
both ideal and efficient and that can reliably predict the extent to which a patient's sickness has spread.

VI. Future Scope

The investigation applied a single model to assess Parkinson's disease presence and severity. The study may be further developed by using a number of models and comparing the results in order to identify the most efficient and optimised models for disease detection and to assess the severity of the condition in the patient.

Acknowledgement

We are appreciative of the expertise and insight that our faculty members contributed to the research by sharing their experiences.

References

[1] R. Das, “Comparison of multiple classification methods for diagnosis of Parkinson disease”. Expert Systems With Applications, vol. 37, pp. 1568–1572, 2010.

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

[3] Rajesh, M., &Sitharthan, R. (2022). Introduction to the special section on cyber-physical system for autonomous process control in industry 5.0.Computers and Electrical Engineering, 104, 108481.

[4] K.R. Seeja, Srishti Grover, Saloni Bhartia, Akshama, and Abhilasha Ya- dav, “Predicting Severity Of Parkinson’s Disease Using Deep Learning”, Procedia Computer Science, vol. 132, pp. 1788–1794.

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

[6] Sitharthan, R., Vimal, S., Verma, A., Karthikeyan, M., Dhanabalan, S. S., Prabaharan, N., ...&Eswaran, T. (2023). Smart microgrid with the internet of things for adequate energy management and analysis.Computers and Electrical Engineering, 106, 108556.

[7] Tappy keystroke dataset.

[8] Parkinson’s speech dataset from Github

[9] Google colaboratory.

[10] Keras api reference.

[11] Python machine learning course.

[12] Deep learning course.

[13] Multiple output neural networks.