Introduction:

Parkinson’s disease (PD) is a disabling brain disorder that affects movements, cognition, sleep, and other normal functions. This disease is said to be the second most common neurodegenerative disorder after Alzheimer’s. It is caused due to the decrease of dopamine levels in the brain. The symptoms usually include tremor in hands, arms, legs, jaw, or head, muscle stiffness, slowness of movement, lack of balance and coordination. It may also result in depression, emotional damages, difficulty in swallowing and talking, constipation, skin problems etc. Unfortunately, there is no current cure and the disease worsens over time. Early prediction of its progression can significantly help in providing required treatments. Traditional diagnosis depends on clinical assessment, which can be challenging, especially in the early stages. Since PD patients exhibit characteristic vocal features, voice recordings can be used for detection. This study aims to apply various machine learning models for the prediction of PD and compare their performances to identify the most optimal algorithm for prediction. Then the algorithm is applied on unseen data to predict the status of the patient.

Methodology:

Dataset description:

The dataset is downloaded from UCI Machine Learning Repository. The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals in 2008-06-26. The original study published the feature extraction methods for general voice disorders [1]. This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of the 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient, the name of the patient is identified in the first column.

The dataset contains the following attributes:

* name - ASCII subject name and recording number
* MDVP:Fo(Hz) - Average vocal fundamental frequency
* MDVP:Fhi(Hz) - Maximum vocal fundamental frequency
* MDVP:Flo(Hz) - Minimum vocal fundamental frequency
* MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP - Several measures of variation in fundamental frequency
* MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA - Several measures of variation in amplitude
* NHR – Noise to Harmonics ratio
* HNR – Harmonics to Noise ratio
* status - Health status of the subject i.e if 1 then Parkinson's, if 0 then healthy
* RPDE, D2 - Two nonlinear dynamical complexity measures
* DFA - Signal fractal scaling exponent
* spread1, spread2, PPE - Three nonlinear measures of fundamental frequency variation

The data was analysed and checked for missing values. It was found that the data did not contain any missing values. The dataset contained 147 recordings that were diagnosed with PD and 48 healthy recordings.

Data visualisation: /\*some graphs and explanation\*/

Conclusion and Future scope:

In conclusion, the study demonstrates the potential of machine learning models in predicting PD progression using vocal features. Among the models evaluated, SVM stands out for its balanced performance in accuracy, sensitivity, specificity, and precision.

Literature Survey

The study [2] uses voice dataset from UCI Machine Learning repository, Orange v2.0b, and Weka v3.4.10 for analysis. SVM achieved 88.9% accuracy, Random Forest achieved 90.26% accuracy, and Naïve Bayes obtained an accuracy of 69.23%.

Speech signals from 252 subjects for PD diagnosis were used where significant differences were found in voice recordings of PD patients compared to healthy individuals . Deep neural networks were employed for voice data analysis. Ensemble of Classifiers achieved an accuracy of 95%, F1-score of 0.97 and Mathews Correlation Coefficient (MCC) of 0.87 [3].

The study [4] highlights the importance of feature selection and algorithm choice in Parkinson's disease diagnosis. Performance evaluation of 20 ML algorithms for Parkinson's disease using acoustic datasets was compared. Different cross validation techniques like LOOCV and Stratified K-Fold cross validation and hyperparameter tuning techniques like Grid Search were used for evaluation.

The **dataset f**rom UC Irvine ML Repository, consisting of 195 voice recordings and a **dataset c**ollected at Istanbul University, including 188 records from Parkinson's patients and 64 records from healthy individuals were used in [5] and an accuracy of 96.41% and 97.35% were obtained using ensemble voting classifiers.

The Study used EEG, MRI, handwriting, and speech data for classification and concluded that deep learning showed promising results. Future directions include integration of different data modalities for improved accuracy, longitudinal data analysis for disease progression forecasting, emphasis on explainable AI and federated learning for collaborative model training [6].

Reference

[1] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering.

[2] PD1 -  Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms

[3] PD2 - Parkinson disease prediction using machine learning-based features from speech signal

[4] PD3 –

[5] PD4 –

[6] PD5 –