Project

Project Proposal

Parkinson’s Disease (PD) is a neurodegenerative disorder of the central nervous system that affects over eight million individuals worldwide. It is primarily caused by the loss and decay of dopamine-producing neurons, and thus a decreased level of dopamine. Symptoms include but are not limited to unintended tremors, slowed movement, muscular stiffness, impaired balance, speech changes, and heightened risk of dementia. Symptoms are progressive, often appearing slowly and vaguely, then gradually worsening over time. While PD cannot be cured as of 2024, medicine and treatment can greatly improve symptoms, especially if the disease is caught early in it’s development.

There is no specific laboratory test to diagnose PD. Diagnosis is difficult, often involving repeated clinical visits where the patient’s history is assessed, motor skills are observed, and general symptoms are monitored. For this reason, an effective screening process that can help identify symptoms, especially prior to clinician visits, would be extremely valuable. A machine learning algorithm that could diagnose the disease before symptoms have progressed to dangerous and/or life-altering degrees could be applied for this goal.

The Parkinson's Disease Detection dataset on Kaggle (https://www.kaggle.com/datasets/jainaru/parkinson-disease-detection) provides biomedical measurements of 23 PD patients’ voices, and those of 8 healthy control individuals. The dataset, and this project therein, utilizes symptoms of PD that affect patient’s voices and speech. In the dataset, individual’s health status is stored in binary format (0 for healthy, 1 for diagnosed with PD), and each column stores different vocal measurements from 195 different recordings.

The overall goal is to use predictive models to discern whether or not an individual has Parkinson’s Disease, based purely on data measuring one symptom of PD: speech impairment.

This is a classification project, particularly a binary classification within the healthcare field. The two classes of interest are that of healthy or Parkinson’s-free, and unhealthy or Parkinson’s-diagnosed, individuals.

During the Exploratory Data Analysis phase, the primary goal will be to grasp the nature of the data and how it’s stored. Finding a general understanding of it’s format and what it represents is the necessary first step. From there, I will further explore the data to get a grasp of the differences between healthy and unhealthy individuals to ensure I can assess the models based on my own understanding of the symptoms and their expressions. This phase may also reveal if dimensionality reduction should be considered, though I do not anticipate it will be.

There are a decent number of columns in this dataset, which explore various facets of individual’s voices from recordings. During model building, I may find it interesting to experiment with feature selection, using only some of the recordings and comparing the different models. Off immediate intuition, however, it seems unnecessary to exclude some symptoms in this type of analysis, but model performance evaluation will be the ultimate judge.

I believe building multiple models and validating them will be the best approach. Many machine learning models excel in binary classification. Starting simple, I will utilize decision trees and, naturally, random forests. Especially since there is only one kind of symptom in question here, I believe decision trees may excel in this classification. K-Nearest Neighbors algorithm and Bayes Classifiers will also be used, and the former will likely encompass the bulk of the analysis.

Models will of course be assessed on all metrics, especially Matthew’s Correlation Coefficient and the AUC-ROC curve. This assessment and validation of the various models will require splitting the data into training and testing/validation sets, which raises the immediate concern of the dataset’s size.

I hope I can get the models to perform at acceptable levels, able to accurately decipher when an individual is healthy or has PD. On top of this, it would be an interesting and worth-while endeavor to determine/extract what patterns in the voice recording data tend to indicate PD.

Citations

'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

“Parkinson’s Disease.” *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 5 Apr. 2024, www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055.

“Parkinson’s Disease.” *National Institute of Neurological Disorders and Stroke*, U.S. Department of Health and Human Services, www.ninds.nih.gov/health-information/disorders/parkinsons-disease. Accessed 8 June 2024.