

**INT 246 Project: Parkinson’s Diseases classification**

Name: Ritik Kumar

Regd. No: 11808486

Section: KM109

Submitted to : Dr. Sukhvir Kaur

**Abstract**

In this project, we are experimenting with a real world dataset, and to explore how machine learning algorithms can be used to predict Parkinson’s Disease(PD). Parkinson’s Disease is a progressive neurodegenerative disease with multiple motor and non-motor characteristics. PD patients commonly face vocal impairments during the early stages of the disease.

We were expected to gain experience using data-mining and machine learning libraries, skLearn , nltk , pandas etc, and we’re expected to submit a report about the dataset and the algorithms used.

**Introduction**

Health informatics systems have been widely used in the detection and monitoring of important diseases in recent years. Information systems bases on artificial learning are utilized in the monitoring of Parkinson’s Disease (PD), which is frequently seen in people over 60 years of age. PD is a progressive neurodegenerative Disease with multiple motor and non-motor characteristics. Due to the prolonged life of the patients with an early diagnosis, high accuracy and reliable health informatics systems are needed for the detection of the PD patients. These systems also aim to reduce the workload of clinicians.

PD detection systems are focused on recognizing the severity of symptoms using several types of instruments. One of the most common symptoms is the vocal problem, and most patients faces vocal defections in the early stages of the disease. Therefore, health systems based on vocal disorders have leading position on recent PD detection studies. In these studies, several speech signal processing techniques were used to obtain clinically relevant features, and extracted features were fed into various artificial learning methods to obtain reliable decisions in PD classification. While Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are the common algorithms in PD classification, Random Forest (RF) [10], and K-Nearest Neighbors (KNN) are also properly used due to their simplicity and also ease of understanding. The success of mentioned algorithms is directly related to the quality of the features selected from the data.

**Related Work**

In this section, we summarize some recent studies on PD classification that use machine learning algorithms and we also cover the recent deep learning methods in PD classification.

**A. Machine Learning for PD classification**

The success of the PD classification studies is directly related to the selection of relevant feature extraction and artificial learning methods. In literature, many studies have used the same publicly available dataset consisting 31 instances (23 PD patients and 8 healthy individuals) with 195 sound recordings. Another PD dataset has 40 examples of 20 PD patients and 20 healthy individuals with multiple speech recordings. Both datasets have commonly extracted features such as vocal fundamental frequency, measures of variation in fundamental frequency, measures of variation in amplitude etc. Since most of the PD detection studies are conducted experiments with these datasets, obtained features from both datasets generally are known as baseline features. Apart from the baseline features, other features that are based on signal processing techniques were also employed in PD detection. Signal-to-noise ratio (SNR), Mel-frequency cepstral coefficients (MFCC) and Tunable Q-factor Wavelet Transform (TQWT) are important tools for extracting relevant features in PD classification . Rather than using separate feature types in model training, most studies use the combination of individual feature types to perform classification task. Extended feature space in these studies can be reduced via feature selection methods . Although, there are lots of symptoms among the people subjecting to the PD including slowed movement, posture and balance deficiencies, dysphonia which is defined as the changes in speech and articulation, is the most meaningful forerunner of PD. This is the reason why many studies are focused on speech based PD classification.

PD patients mainly face vocal defections which directly influence the vocal loudness, instability and frequency abnormality. Voice breaks and impaired vocal quality are also the other impairments that can be seen in PD patients. Speech processing techniques is commonly used to detect anomalies in speaking and it is often preferred in automated extraction of PD-related vocal features. During the last decade, several machine learning based studies have been performed in the detection of PD using vocal features.

Least Absolute Shrinkage and Selection Operator (LASSO), Minimum Redundancy Maximum Relevance (mRmR), Relief and Local Learning-Base Feature Selection (LLBFS) were the methods used for feature selection and the performance of the selected features were evaluated with Random Forest (RF) and Support Vector Machines (SVM) classifiers. These classifiers resulted the performances up to 98.6% of precision rate using features from the shimmer, HNR and vocal fold excitation. Their study was also found out that the feature set with the lowest classification error was obtained from the Relief selection.

**B. Deep Learning for PD Classification**

Besides common machine learning algorithms, a subdivision of machine learning called deep learning also has been successfully implemented in the PD studies. For instance, study by used a well-designed smart pen to capture handwritten dynamics from healthy individuals and PD patients. In this study, the handwritten dynamics were modeled as a time series data, and used as inputs to the proposed CNN. Suggested CNNs were built on already-trained deep learning architectures as LeNet, Cifar10 and ImageNet. In order to compare the performances of the proposed CNN, Open Path Forest (OPF) classifier was trained with the raw time series data. Over all experiments, CNN results in better performances than the OPF with the help of its ability to learn important features to distinguish PD patients from healthy individuals

**DATASET**

The data used in this study were taken from UCI Machine Learning repository and it has been recently used in study. The dataset was gathered at the Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University and it contained 188 PD patients (107 men and 81 women) and 64 healthy individuals (23 men and 41 women). The age of PD patients varied between 33 and 87 years, while the age of healthy subjects ranged from 41 to 82 years. During the data collection, frequency response of the microphone was set to 44.1 KHz, and after the doctor review, repeated repetition of the vowel/a/letter in ea

ch person was collected with three replicates.