# **Abstract:**

**Parkinson's disease is a progressive disorder that affects the nervous system and the parts of the body controlled by the nerves. Initial symptoms involve tremors in just one hand, stiffness, or slowing of movement. In the early stages of Parkinson's disease, a patient may show little or no expression. Their arms may not swing during a walk. Speech may become soft or slurred. Parkinson's disease symptoms worsen as the condition progresses over time. The symptoms are different for everyone and are unnoticed. Blood tests and Scans do not provide strong evidence of the disease. However, speech gets changed in the early stage which can be used for the prediction of Parkinson’s disease. In this paper, a total of 9 supervised machine learning models are used combined with 6 feature selection techniques. These models have been trained and tested not just for their accuracy but also for the f1 score. Performance analysis of all the models has been presented in this paper and a comparative study has been done. Support Vector Machine is found to give a 97% f1 score and 93% AUC score with a total of 8 features.**

**Introduction:**

**What is Parkinson’s disease, causes and symptoms?**

**Parkinson's disease affects the central nervous system, particularly the portion of the brain that governs movement, and is a chronic, progressive neurodegenerative condition. The substantia nigra area of the brain loses dopamine-producing cells, which causes a variety of motor symptoms include tremors, stiffness, sluggish movement, and issues with balance and coordination. Reduced facial expression, small crowded handwriting (micrographia), Soft slurred or sluggish voice (hypophonia), overall weakness or discomfort and a "masked" face are further signs and symptoms of Parkinson's disease. Parkinson's disease can result in a variety of non-motor symptoms, including sadness, anxiety, sleep issues, and cognitive impairment, in addition to its motor symptoms.**

**Why is there need of machine learning in prediction of Parkinson’s disease?**

**Parkinson's disease is believed to be brought on by a combination of hereditary and environmental factors, while its exact causes are yet unclear. Although, there is currently no cure for Parkinson's disease, there are medicines that can help patients live better lives by managing their symptoms.** Since, Machine learning can analyze huge volumes of data and identify patterns that people would not see, it is being used to forecast Parkinson's illness. This can be especially helpful in forecasting Parkinson's disease or identifying people who are most at risk for getting the condition.

In general, machine learning holds the potential to enhance the efficiency and accuracy of Parkinson's disease prediction , which may result in an earlier diagnosis and more successful treatment.

Machine learning can predict Parkinson's disease using a variety of methods, such as:

**Analysis of medical imaging**: Machine learning algorithms can examine brain scans and spot tiny abnormalities in the structure or operation of the brain that could be signs of Parkinson's disease in its early stages.

**Voice analysis**: Parkinson's disease can affect a person's voice, causing it to become softer, hoarser, or more monotone. Machine learning algorithms can analyze speech patterns to identify early signs of Parkinson's disease.

**Movement analysis**: Parkinson's disease affects movement, causing tremors, stiffness, and slowness. Machine learning algorithms can analyze movement patterns to identify subtle changes that may indicate the early stages of the disease.

**Genetic analysis**: Parkinson's disease has a hereditary component in certain cases. Genetic data can be analyzed by machine learning algorithms to pinpoint those who have a high chance of developing the illness.

Overall, machine learning has the potential to improve the accuracy and efficiency of Parkinson's disease prediction, which can lead to earlier diagnosis and more effective treatment.

**Why voice data is more suitable for the prediction of Parkinson’s disease ?**

The voice is influenced by Parkinson's disease in several ways, making it a helpful tool for early illness identification. Parkinson's disease can alter an individual's voice, resulting in alterations including hoarseness, monotony, breathiness, and decreased volume. These changes are caused by the degeneration of the nerves that control the muscles used in speaking, swallowing, and breathing.

Voice analysis is a non-invasive (not harmful) and cost-effective method for detecting Parkinson's disease. It involves analyzing recordings of a person's speech to detect changes in voice quality that may indicate the presence of the disease. Voice analysis can be performed using a variety of methods, including acoustic analysis and machine learning algorithms.

Acoustic analysis involves measuring various aspects of a person's speech, such as pitch, loudness, and duration. Machine learning algorithms can then be used to then identify patterns in the data that are indicative of Parkinson's disease.

Voice analysis is also convenient and accessible, as it can be performed remotely using a smartphone or other recording device. This makes it a valuable tool for screening large populations, such as elderly individuals who may be at higher risk for Parkinson's disease.

Overall, voice data is a promising tool for the early detection of Parkinson's disease, which can lead to earlier diagnosis and treatment, and ultimately improve patient outcomes.

**Objective and aims:**

Aim: this paper aims to classify the patients having Parkinson and not having Parkinson by maximizing the score(f1 score)/accuracy while compromising the training time.

Objective: Feature selection methods like variance thresholding, Fischer score, Information Gain, particle swarm optimization, genetic algorithm is combined with various machine learning algorithms like Logistic Regression, Decision Tree, Random Forest, Support Vector Machines, Gaussian Naïve Bayes, XGBoost, AdaBoost, Gradient Boosting,

**Previous Work:**

1. Yunfeng Wu et.al. 2017 [1] purposed a generalized logistic regression analysis (GLRA), Support Vector Machine (SVM), and Bagging ensemble algorithm input with the ICPR (Interclass probability risk) features. They found SVM is much better at distinguishing normal vocal patterns with a specificity of 0.8542. Among the three classification methods, the Bagging ensemble algorithm with ICPR features can identify 90.77% vocal patterns, with the highest sensitivity of 0.9796 and largest area value of 0.9558 under the receiver operating characteristic curve.

Wu Y, Chen P, Yao Y, et al. Dysphonic Voice Pattern Analysis of Patients in Parkinson's Disease Using Minimum Interclass Probability Risk Feature Selection and Bagging Ensemble Learning Methods. *Comput Math Methods Med*. 2017;2017:4201984. doi:10.1155/2017/4201984

1. Marziye Keshavarz Shahsavari et al. [2] established the extreme learning machine and hybrid particle swarm optimization to detect Parkinson’s disease. They optimized at most 12 features from 22 using Hybrid Particle Swarm Optimization and Extreme Learning Machine and achieved 88.72% accuracy using their proposed method.

Dhinakaran, J. & Thambusamy, Velmurugan. (2022). Accuracy based Performance Analysis of Classification Algorithms using Parkinson Disease Dataset. 927-933. 10.1109/ICAAIC53929.2022.9792981.

1. Akshaya Dinesh et al. [3] proposed an ML approach to detect Parkinson’s disease from sound recordings. They used various learning methods and filter based instance selection schemes. They selected the top 10 features from 22 and achieved 91.21% accuracy by using Boosted Decision Tree.

Dinesh, Akshaya & He, Jennifer. (2017). Using machine learning to diagnose Parkinson's disease from voice recordings. 1-4. 10.1109/URTC.2017.8284216.

1. Geeta Yadav, Yugal Kumar and G. Sahoo (2012) [4] used sequential minimization optimization, logistic regression and decision stump for the prediction of PD and obtained the best score from the support vector machine model with 76% accuracy.

G. Yadav, Y. Kumar and G. Sahoo, "Predication of Parkinson's disease using data mining methods: A comparative analysis of tree, statistical and support vector machine classifiers," 2012 NATIONAL CONFERENCE ON COMPUTING AND COMMUNICATION SYSTEMS, Durgapur, India, 2012, pp. 1-8, doi: 10.1109/NCCCS.2012.6413034.

1. Sachin Shetty et al. [5] used SVM Based ML approach to recognize Parkinson’s Disease. At first, they picked up feature vectors using a host of statistical tools and selected the best seven features. Finally, they applied SVM to measure performance. They obtain good overall accuracy of 83.33% with seven selected instances and SVM.

Shetty, Sachin & Rao, Yerramreddy. (2016). SVM based machine learning approach to identify Parkinson's disease using gait analysis. 1-5. 10.1109/INVENTIVE.2016.7824836.

**Dataset Description**

A dataset has been selected from the University of California, Irvine’s (UCI) public Machine Learning repository. This dataset is composed of a range of biomedical voice measurements from 31 people. Out of 31 subjects, the data related to 8 healthy and 23 Parkinson’s disease affected patients has been used for training and testing different predictive models. The dataset contains 4290 samples (195 rows x 22 columns) of biomedical voice measurements [6]. It has 195 vowel voice records taken against 31 subjects, where 147 recordings are taken from patients suffering from Parkinson’s disease and 48 recordings from healthy people. Each column of the data file corresponds to an individual voice recording and contains certain attributes and features which are used as inputs in the machine learning algorithm. The voice features are classified into 6 categories i.e., Amplitude, Pulse, Frequency, Voicing pattern, Pitch and Harmonicity.

*B. The Input Variables*

Six Amplitude parameters representing the local shimmer,

The local shimmer in dB, 3-point amplitude perturbation

quotient, 5-point amplitude perturbation quotient, 11-

point amplitude perturbation quotient, average absolute

difference between the amplitude of consecutive periods.

• Eleven Frequency parameters representing the average

vocal fundamental frequency, maximum vocal

fundamental frequency, minimum vocal fundamental

frequency, jitter in percentage, Absolute jitter, relative

amplitude perturbation, period perturbation quotient,

jitter cycles, two nonlinear measures of fundamental

frequency variation and pitch period entropy.

• Two harmonicity parameters representing the Harmonic-to-

Noise ratio and Noise-to-Harmonic ratio.

• Two Complexity parameters representing the recurrence

period density entropy measure and the Correlation

dimension.

• One Signal scaling parameter representing the signal

fractal scaling exponent of detrended fluctuation analysis.

*C. The Output Variables*

The 'Status' variable distinguishes healthy individuals from patients with PD. The single binary output variable given is:

1- Parkinson’s disease(Positive)

0- Parkinson’s disease(Negative)

Table 1, indicates the health status of the subject.**Citation:** '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)