EasyChair Preprint

№ 11814

NeuroDx: a Novel Machine Learning Paradigm for Unveiling Parkinson's Disease Patterns

Sharayu Garad, Pranoti Naiknaware, Anisha Shinde,
Ashutosh Garad and Meenakshi Pawar

January 20, 2024

NeuroDx: A Novel Machine Learning Paradigm for Unveiling Parkinson's Disease Patterns

Sharayu Garad || Pranoti Naiknavare || Anisha Shinde

Ashutosh Garad || Meenakshi Pawar

Shri Vitthal Education & Research Institute's

College of Engineering Pandharpur

Abstract

This paper explores an innovative slant for detecting Parkinson's disease (PD) by analyzing voice data from patients. To extract meaningful features from the MDVP voice input, a machine learning technique, including a Support Vector Machine (SVM) is employed. The study emphasizes the importance of data collection, preprocessing, and feature engineering to improve model accuracy. Robustness is ensured by cross-validation and testing across diverse patient datasets. Integrating voice-based PD detection in clinical practice holds potential for early diagnosis and personalized care. This research highlights the efficacy of voice-based machine learning in enhancing PD detection, offering a non-invasive and patient-centric approach.

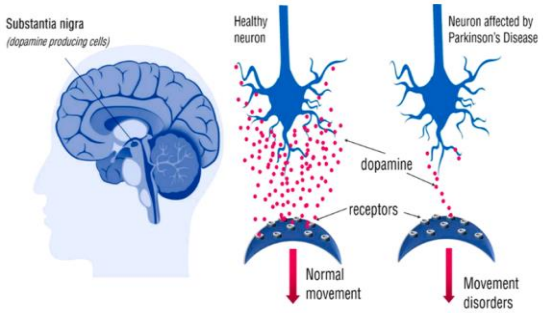
Keywords- Parkinson's Disease (PD), Machine Learning(ML), Support Vector Machine (SVM).

1.Introduction

Parkinson's disease (PD) is a chronic and escalating brain degenerative disorder that primarily affects movement control. It is characterized by the gradual deterioration and loss of dopamine-producing neurons in a region of the brain called the substantia nigra. Dopamine is a neurotransmitter that plays a pivotal role in facilitating smooth, coordinated muscle movements. It is a prevalent neurodegenerative disorder, that poses significant challenges to patients' well-being. Dysphonia often presents as an initial sign of Parkinson's disease.[1] Other symptoms also include Tremors, Bradykinesia, Muscle rigidity, Postural instability, etc.[2]

Parkinson's Disease is typically characterized into 5 stages. Stage 1 includes early motor symptoms, often mild and affecting one side of the body. Stage 2 includes moderate motor symptoms, and bilateral involvement (both sides of the body). Stage 3 is Mid-stage with significant balance and coordination problems. Stage 4 is the advanced stage, requiring assistance with daily activities. Stage 5 includes severe motor symptoms, often confined to a wheelchair or bed.

The development of deep convolutional neural networks (CNN) for automated PD detection using voice signals, achieves an accuracy of 89.75%. The results suggest that integrating this model into smart electronic devices could offer substitute pre-diagnosis methods and aid physicians during in-clinic assessments, potentially improving patients' quality of life and reducing healthcare costs.[3] The use of Machine Learning algorithms like Random Forest, Support Vector Machine, Logistic Regression, and K-Nearest Neighbour (KNN) can also be done to detect PD. [2]



1.1 Literature survey

Previous research into Parkinson's disease (PD) detection primarily focused on modalities like MRI scans, gait analysis, and genetic data.[4] Audio-based methods for early PD detection were relatively unexplored. For instance, Bilal et al. achieved an SVM model accuracy of 0.889 using genetic data, while this study improves upon that with an SVM model boasting an accuracy of 0.9183, underscoring the efficacy of using audio data for PD classification.[5]

In another study, Raundale, Thosar, and Rane used keyentry data to prognosticate the extremity of PD in older patients, while Cordella et al. utilized audio data but relied heavily on MATLAB. In contrast, this research adopts open-source Python models, which are nimble and more memory-conserving.[6]

The plurality of previous PD research emphasized deep learning approaches. For instance, Ali et al. employed deep learning models in collaboration with vocalization data but lacked feature selection, a gap filled by this study, which employs Principal Component Analysis (PCA) to select seven major voice features for enhanced PD detection.[7]

In a similar vein, prior studies, for example, the research conducted by Huang et al., focused on diminishing the dependency on smart wearables for Parkinson's disease (PD) diagnosis by employing conventional decision trees with a focus on speech features. In another approach, Wodzinski et al.[8] harnessed a ResNet model for analyzing audio images, while Wang et al.[9] implemented a combination of 12 machine learning models to evaluate acoustic biomarkers and Alkhatib et al. focused on characterizing shuffling movements in PD patients. Ricciardi et al. analysed brain MRI scans over time to detect Mild Cognitive Impairment (MCI) involves a detailed assessment of structural and functional brain changes, helping identify early signs of cognitive decline in PD patients.[10] Chakraborty, and Mukherjee Guruler [11] used Sugeno–Takagi Fuzzy Inference System based on Fuzzy C-Means clustering. Chen et al.[12] used Fuzzy k-nearest neighbor and Peker, Sen, and Delen[13] used minimum surplus maximum pertinence attribute selection, and complex-valued artificial neural network with high accuracy.

Building on this extensive literature review, the current research has implemented a PD classification model using audio data with the aim of advancing early PD detection through telemedicine. Considering past biomarker data research, this study explores Support Vector Machine for classifying audio data from Parkinson's patients. The K-nearest neighbor model outperforms others, achieving an impressive accuracy of 93.83%. This study adds to the expanding collection of research in the area of Parkinson's disease (PD) detection and telemedicine.

2.Methodology

The suggested approach involves gathering a diverse dataset of voice recordings, encompassing individuals both afflicted by Parkinson's disease and those in good health while ensuring a wide range of disease severity levels. Preprocessing of the voice data by segmenting the recordings, removing background noise, and converting them into suitable formats for analysis is done. Extracting relevant features from the voice data, such as pitch, jitter, shimmer, Mel-frequency cepstral coefficients (MFCCs), and fundamental frequency are the steps of data collection and preprocessing. The next step is Feature Selection and Engineering i.e. conducting exploratory data analysis (EDA) to understand the distribution and relationships of the extracted features. It also includes applying feature selection techniques (e.g., correlation analysis, mutual information) to identify the most relevant features for distinguishing between healthy and Parkinson's disease cases as well as engineer new features if necessary, such as statistical measures or ratios based on the selected features. It is followed by Dataset Splitting and Standardization which consists of Splitting the pre-processed dataset into training and testing sets. It also involves the use of stratified sampling to ensure a balanced representation of both classes. Applying standardization to scale the features to zero mean and unit variance, enhances the model's convergence and performance. Model Selection and Training is the experiment with machine learning algorithm Support Vector Machines (SVM). It utilizes Python's sci-kit-learn library along with numpy and pandas for model implementation, training, and hyperparameter tuning. Training of the model is done using the training dataset, monitoring its performance using cross-validation techniques. Classification reports and confusion matrices can be used to assess the models' performance on both training and testing datasets. Interpretation of feature importance or coefficients to understand the contributions of different features to the classification decisions is useful.

The trained model is assessed using various metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The final model is validated on an independent validation dataset, preferably collected from a different source or population.

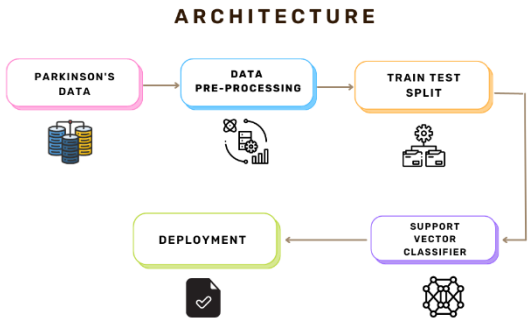


Fig.2) Proposed methodology

Dataset

Table 1: Dataset attributes

Attribute	Purpose
Name	Data is stored in ASCII CSV format where patient name and recording number is stored
MDVP: Fo (Hz)	Fundamental frequency of pitch period
MDVP: Fhi (Hz)	Upper limit of fundamental frequency or maximum threshold of voice modulation
MDVP: Flo (Hz)	Lower limit or minimal vocal fundamental frequency
MDVP: Jitter, Abs, RAP, PPQ, DDP	These are various Kay Pentax's multi-dimensional voice program (MDVP) measures. MDVP is a traditional measure of frequency of vibrations in vocal folds at pitch period to vibrations at start of next cycle called pitch mark [25]
Jitter and Shimmer	Measures of absolute difference between frequencies of each cycle, after normalizing the average
NHR and HNR	Signal to noise and tonal ratio measures, that indicate robustness of environment to noise
Status	0 indicates healthy person while 1 indicates PWP.
D2	Correlation dimension is used to identify dysphonia in speech using fractal objects. It is a nonlinear, dynamic attribute.
RPDE	Recurrence Period Density Entropy quantifies the extent to which signal is periodic
DFA	Detrended Fluctuation Analysis or DFA measures the extent of stochastic self-similarity of noise in speech signals.
PPE	Pitch Period entropy is used to assess abnormal variations in speech on a logarithmic scale
Spread1, spread2	Analysis of extent or range of variations in speech with respect to MDVP: Fo(Hz)

3.Model Implementation

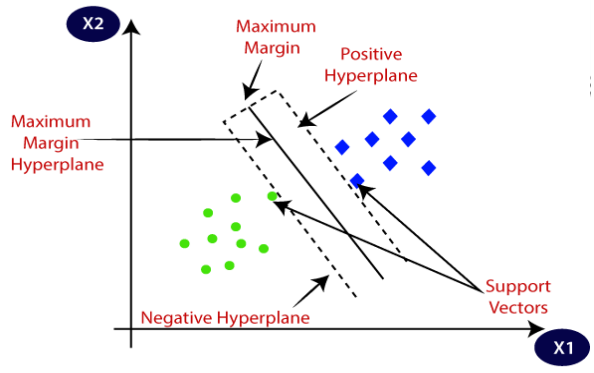


Fig.3) Support Vector Machine

The Support Vector Machine (SVM) stands out as a highly favoured tool in supervised learning, finding applications in both classification and regression tasks. However, its primary use revolves around classification challenges within the realm of machine learning. The fundamental objective of the SVM algorithm is to determine the optimal line or boundary that effectively separates data in multi-dimensional space into different categories. This boundary is aptly referred to as a hyperplane. Once established, this hyperplane enables straightforward categorization of new data points, ensuring they end up in the correct class.

SVM operates by identifying the most extreme data points, which play a crucial role in defining the hyperplane. These exceptional data points are known as "support vectors," hence the name "Support Vector Machine." SVMs prove highly effective when dealing with data that exists in high-dimensional spaces and are especially valuable when the goal is to have a clear and distinct separation between different categories or classes. They can also handle data that doesn't follow a linear pattern by using kernel functions, which transform the data into a higher-dimensional space, thereby allowing for the discovery of linear boundaries within this transformed space. Since PD voice data doesn't exhibit a linear separation pattern, we apply an SVM kernel to convert the data into a higher-dimensional space. SVM excels with PD data due to its efficient memory usage and the utilization of support vectors derived from a subset of the training data points.

4. Result

This study focused on Parkinson's disease detection, employed Support Vector Machine as a classification model using voice parameters. The findings revealed a high level of accuracy, with an impressive 93.86% classification accuracy. Promising results were achieved improving the model's performance. The SVM model exhibited robustness when dealing with outliers and ensured a high level of precision in disease classification. Notably, there were no false positive predictions, highlighting the reliability of this approach for detecting Parkinson's disease.

This research suggests that SVM, in conjunction with voice parameter data, can serve as a valuable tool for accurate and non-invasive Parkinson's disease detection. These results hold promise for the development of practical applications that can assist in the early diagnosis and monitoring of the disease.

5. Analysis of Result

ID	Fo(Hz)	Fhi(Hz)	…	…	Status
S01	119.992	157.302			1
S02	197.076	206.896			0
S03	116.682	131.111			1

In this table, Patient ID uniquely identifies each individual. Fo(Hz) represents “Speech Feature 1”, Fhi(Hz) represents "Speech Feature 2," and so on represents the extracted features from the voice data. These are the various acoustic features, such as pitch, jitter, shimmer, formants, and other characteristics related to speech patterns that are known to change with Parkinson's disease. “Status” is the target variable, where 1 indicates the presence of PD, and 0 indicates the absence.

ML models are trained on such data to learn the relationships between the input features (speech characteristics) and the target variable (Parkinson's disease status). After training, these models can make predictions for new, unseen data to assess whether an individual may have Parkinson's disease based on their voice characteristics.

6. Future Scope

The future of Parkinson's disease (PD) detection using the SVM machine learning algorithm and patient voice data holds great potential. It promises more accurate and early diagnosis by leveraging advanced SVM models and larger, diverse datasets. This approach could also integrate with wearable devices for continuous monitoring and personalized care. As these innovations unfold, it is imperative to consider ethical and privacy concerns, ensuring that the use of voice data is carried out responsibly and with utmost respect for individuals' privacy.

7. Conclusion

This paper delved into a novel method for identifying Parkinson's disease (PD) through the examination of voice data obtained from individuals with the condition. The use of a Support Vector Machine (SVM), a Machine Learning (ML) algorithm for Parkinson's Disease (PD) diagnosis through analysis of human voice offers a promising avenue for early and non-invasive detection. Machine learning (ML) algorithms have showcased their capability in precisely detecting vocal cues linked with Parkinson's disease (PD), potentially paving the way for more easily attainable and punctual diagnosis, ultimately resulting in a substantial enhancement in the well-being of individuals vulnerable to this ailment. This innovative approach holds the potential to revolutionize PD screening and monitoring, offering a cost-effective and efficient solution for early intervention and treatment. However, the achieved accuracy is limited to 93.83%.

8.References

[1] Tao Zhang, Liqin Lin, Zaifa Xue, “A voice feature extraction method based on fractional attribute topology for Parkinson’s disease detection”, Expert Systems With Applications 219 (2023) 119650

[2] Aditi Govindu, Sushila Palwe, “Early detection of Parkinson’s disease using machine learning”, 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0) Peer-review under the responsibility of the scientific committee of the International Conference on Machine Learning and Data Engineering

[3] Onur Karaman, Hakan Cakin, Adi Alhudhaif, Kemal Polat, “Robust automated Parkinson disease detection based on voice signals with transfer learning”. Expert Systems with Applications 178 (2021) 115013

[4] Alatas Bilal, Moradi Shadi, Tapak Leili, Afshar Saeid (2022), “Identification of Novel Noninvasive Diagnostics Biomarkers in the Parkinson’s Diseases and Improving the Disease Classification Using Support Vector Machine”, BioMed Research International, Hindawi

[5] P. Raundale, C. Thosar and S. Rane (2021), "Prediction of Parkinson’s disease and severity of the disease using Machine Learn ing and Deep Learning algorithm," 2021 2nd International Conference for Emerging Technology (INCET), pp. 1-5, doi: 10.1109/INCET51464.2021.9456292.

[6] F. Cordella, A. Paffi and A. Pallotti (2021) "Classification-based screening of Parkinson’s disease patients through voice signal," 2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 1-6, doi: 10.1109/MeMeA52024.2021.9478683.

[7] F. Huang, H. Xu, T. Shen and L. Jin (2021), "Recognition of Parkinson's Disease Based on Residual Neural Network and Voice Diagnosis," 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control

Conference (ITNEC), pp. 381-386, doi: 10.1109/ITNEC52019.2021.9586915.

[8] M. Wodzinski, A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave and E. Nöth, (2019) "Deep Learning Approach to Parkinson's Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 717-720, doi: 10.1109/EMBC.2019.8856972.

[9] W. Wang, J. Lee, F. Harrou and Y. Sun, "Early Detection of Parkinson's Disease Using Deep Learning and Machin e Learning," in IEEE Access, vol. 8, pp. 147635-147646, 2020, doi: 10.1109/ACCESS.2020.3016062.

[10] X. Yang, Q. Ye, G. Cai, Y. Wang and G. Cai, (2022), "PD-ResNet for Classification of Parkinson's Disease from Gait," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 10, pp. 1-11, 2022, Art no. 2200111, doi: 10.1109/JTEHM.2022.3180933.

[11] Atanu Chakraborty, Aruna Chakraborty and Bhaskar Mukharjee, "Detection of Parkinson's Disease Using Fuzzy Inference System", https://link.springer.com/chapter/10.1007/978-3-319-23036-8_7.

[12] Sajad Mohamadzadeh, Sadegh Pasban, Javad Zeraatkar-Moghadam & Amir Keivan Shafiei. "Parkinson's Disease Detection by Using Feature Selection and Sparse Representation"

[13] Musa Peker, Baha Sen, and Dursun Delen, "Computer-Aided Diagnosis of Parkinson's Disease Using Complex-Valued Neural Networks and mRMR Feature Selection Algorithm", Journal of Healthcare Engineering.