Thompson Rivers University

DETECTION OF EARLY PARKINSON'S DISEASE USING GOLD RUSH LIVER CANCER HYBRID ALGORITHM

By
Melvin Biju

PROJECT PROPOSAL REPORT FOR THE PARTIAL FULFILMENT OF THE DEGREE OF

Master Of Science in Data Science

KAMLOOPS, BRITISH COLUMBIA

April 2024

SUPERVISOR

Dr. Mohammed Tawhid

1. Abstract

Millions of people are diagnosed with Parkinson’s every year. People with this condition exhibit abnormality in speech that listeners may overlook, but the recorded speech signals can be analyzed for early detection. Conventional methods of diagnosis are highly subjective and, hence, can be prone to error. Therefore, there is a need for an accurate method of early identification and classification of Parkinson’s. This study aims to develop a machine learning-based model for the early detection of Parkinson's Disease (PD), focusing on the identification of predictive biomarkers and characteristics from 15 comprehensive medical datasets, including voice recordings and clinical data. It will employ the wrapper feature selection and extraction technique to enhance model accuracy and will explore various machine learning classifiers (KNN, Random Forest and Decision tree) to construct an effective classification system. The project's objectives include validating the model's effectiveness against conventional diagnostic techniques, ultimately improving early PD diagnosis and patient care outcomes.

Key words: Parkinson’s disease, Machine learning, medical data, Wrapper, Feature Selection, Prediction

2. Introduction

The brain is the most crucial organ in the human body and is responsible for the effective functioning of all the other organs. One of the diseases that impairs its function is Parkinson's. It is an incurable neurological disease that affects people mostly above the age of 50 [7]. Statistics show that only 4% affected with PD are below 50 years of age [8]. The symptoms of this disease can be quite subtle. The main motor symptoms are slowness of movement, tremor, stiffness, and poor balance [9]. Non-motor symptoms include hypotension, mood disorders, pain, sensory disfunction and loss of weight [9]

Studies show that PD patients commonly have speech disorders [9]. They often portray anomalies like quiet and hurried speech [10]. The analysis of this speech data is considered as an important non-invasive method for PD identification. This makes it important for doctors in diagnosing this disease. The detection of changes in speech pattern is extremely useful to identify PD in its early stages. As stated before, this disease is incurable; however, there are several treatments that reduce the symptoms in early stages.

Evolutionary computation has recently received much attention as a solution to optimization problems. Evolutionary algorithms are bio-inspired generic population-based optimization algorithms that mimic biological processes such as selection, mutation, and reproduction to solve optimization issues. Another metaheuristic technique is the crowd search algorithm (CSA), which was motivated by the clever ways that crows conceal and pilfer food [11]. Crows are believed to be some of the world's most intelligent creatures. Several algorithms that draw inspiration from nature have been proposed in recent studies. For instance, the foundation of genetic algorithms (GA) is natural selection; bat algorithms draw inspiration from microbat echolocation techniques [12], and particle swarm optimization leverages the clever collective behavior of fish schools and bird flocks [14].

Voice frequency analysis is a relatively accurate and non-invasive process. Consequently, voice frequency can be utilized to monitor the development of this irrational illness [16]. Numerous speech tests have been carried out to monitor the disease's course. ML techniques are regularly applied in the medical (healthcare) industry. A range of data modalities, such as handwriting patterns and acoustic voice recordings, are being combined with machine learning algorithms to diagnose Parkinson's disease. We can identify relevant characteristics that aren't typically used in the medical diagnosis of Parkinson's disease (PD) with the aid of machine learning (ML) tools, and we can rely on these alternative indications to detect PD in its preclinical stages. The wrapper method is used to eliminate or reduce the noise in the dataset.

Finally, classification techniques are executed during the final phase. The performance of the classification method has a significant impact on the feature extraction approach. For this reason, selecting the best classification method is a crucial factor that must be considered for this disease. This review examines how ML models trained on sensory data help doctors diagnose patients with Parkinson's disease (PD) at every stage of the treatment process. The aim is to give neurologists early detection insights that could enhance the diagnosis and treatment of Parkinson's disease (PD) by highlighting additional unique techniques and offering new solutions that have not been adequately addressed in the reviews that have already been published. The contribution of this paper is as follows:

- We provide an overview of Parkinson disease, outlining its primary traits as well as its most common motor and non-motor symptoms.
- We test several optimization algorithms and ML models, analyzed the accuracy of ML models for the diagnosis of Parkinson disease on benchmarked medical datasets.
- Finally, the study shows the challenges and discusses the potential future work that can be implemented.

3. Methodology

3.1 Optimization algorithms:

In machine learning, these algorithms adjust the parameters of models to reduce errors on training data, essentially finding the best possible settings for making accurate predictions. When applied to medical data, optimization algorithms play a crucial role in developing predictive models that can identify disease patterns, optimize treatment plans, and improve patient outcomes. In this study, we evaluate the performance of 10 optimization algorithms to choose an algorithm for implementation. We use the CEC benchmark problems with similar parameters for a consistent comparison.

3.2 Feature Selection:

Feature selection is a fundamental process in machine learning that significantly enhances model performance and efficiency. By identifying and retaining only the most relevant features, it helps in reducing overfitting, decreasing training time, and improving the interpretability of models. This process is crucial for simplifying complex data, facilitating easier visualization and analysis, and ensuring models are both accurate and understandable. For this study, we use the Wrapper feature selection model. It has been empirically proven that wrappers obtain subsets with better performance than filters because the subsets are evaluated using a real modelling algorithm [1].

3.1.2 Wrapper method:

The wrapper feature selection method is a search technique for selecting a subset of relevant features for model construction [6]. It relies on the performance of a given machine learning algorithm to evaluate the combination of features rather than relying on the intrinsic properties of the data. In essence, wrapper methods use the predictive model to score feature subsets based on their predictive power, typically through cross-validation.

For medical datasets, they can help identify combinations of features that are most predictive of a patient's condition, considering the complex interactions and dependencies between variables characteristic of medical data. For example, in genomics, certain gene expressions may be strong indicators of disease only in combination with others. Wrapper methods are adept at uncovering these complex patterns, which might be missed by more straightforward feature selection methods that must consider the modeling process. However, it is worth noting that wrapper methods can be computationally intensive, especially with very large datasets, because they require training a new model for each subset of features considered. Despite this, their tailored approach to feature selection often results in improved model performance, making them a powerful tool for predicting medical diagnoses and outcomes.

3.3 Machine Learning models:

Machine learning models offer transformative potential for diagnosing and managing Parkinson's Disease (PD). By processing and learning from vast and complex medical datasets, these models can uncover subtle patterns associated with the early stages of PD that might elude traditional diagnostic methods. Features such as voice patterns, motor movements, and even genetic markers can be fed into algorithms to predict the onset and progression of the disease with high accuracy[17]. The adaptive nature of ML models means they can continuously improve as they are exposed to more data, which is particularly valuable in tracking the progression of PD over time or in response to treatment. Moreover, machine learning can personalize patient care by predicting individual responses to therapies, enhancing the quality of life for those affected.

3.3.1 K-Nearest Neighbors (KNN):

KNN is a non-parametric, instance-based learning algorithm that classifies new cases based on a similarity measure (usually distance functions) [2]. KNN is particularly useful for medical data because of its simplicity and effectiveness in cases where the relationship between the data points is not linear. It can be very effective for diagnostic problems where the class of its neighbors determines the classification of a sample in the feature space. Its reliance on local information means it can adapt to changes in the data distribution, making it robust to noise and outliers often found in medical datasets.

3.3.2 Random Forest (RF)

Random Forest is an ensemble learning method that operates by constructing many decision trees [3] at training time and outputting the class, which is the mode of the classes (classification) [4] of the individual trees. RF is excellent for handling the high dimensionality and complexity often associated with medical data without overfitting due to its ensemble nature. It can manage thousands of input variables without variable deletion, assess the importance of variables, and model complex relationships between features and the target. These properties make it highly suitable for medical datasets where numerous variables and their interactions can affect the outcome.

3.3.3 Decision Tree (DT)

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. [5] DTs are transparent and easy to interpret, which is critical in medical settings for understanding decision-making. They are helpful for medical data as they can handle numerical and categorical data and can model complex relationships by splitting the dataset into subsets based on different conditions. This hierarchical structure mimics clinical decision-making processes, aligning closely with medical diagnostic reasoning.

3.4 Benchmark Data Review:

This study uses 15 benchmark medical datasets of various kinds for benchmark testing. The datasets are collected from the UCI machine learning repository. Table 1 describes the datasets. The purpose of choosing these 15 datasets is to examine whether the suggested algorithm can be assessed on various datasets. Here, some values in a few of the datasets need to be added. The values missing when each feature's mode is determined have been replaced with the most frequent value in the feature. Next, all the missing values in the corresponding features are replaced using the mode.

Datasets	Samples	Features	Classes	Missing values
Breast Cancer Wisconsin (Original)	699	10	2	Yes
Pima Diabetes	768	8	2	Yes
Hepatitis	155	19	2	Yes
Cancer Data	569	30	2	No
Lymphography	148	18	4	No
Stat log (Heart)	270	13	2	No
Colon	62	2000	2	No
Single-photon emission computed tomography (SPECT)	267	22	2	No
Parkinson	756	754	2	No
Indian liver patient dataset	583	10	2	No
Thoracic surgery dataset	470	17	2	No
Lung cancer	32	56	3	Yes
Mice Protein Expression (MPED)	1080	82	8	Yes
Zoo	101	18	2	No
Cardiotocography	2126	23	3	No

Table 1. Dataset description

4. Timeline

Phase	Task	Tentative completion
I	Gather 10 algorithm codes and run for CEC problems	5/10/2023
II	Create feature selection code for the algorithms	12/12/2024
III	Modify the code to run 14 Benchmark Datasets	15/2/2024
IV	Implement Hybrid approach	25/3/2024
V	First Draft of Report and Slides for presentation	10/4/2024
VI	Presentation	23/4/2024

5. References

1. A. Jović, K. Brkić and N. Bogunović, "A review of feature selection methods with applications," 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2015, pp. 1200-1205, doi: 10.1109/MIPRO.2015.7160458. keywords: {Search problems;Filtering algorithms;Information filters;Accuracy;Classification algorithms;Clustering algorithms}
2. Madhumathi, V., & Rajan, T. (2014). Review on: Nearest Neighbor Search with Keywords. International Journal of Science, Engineering and Computer Technology, 4(10), 263-265.
3. Kouzani, A. Z., Nahavandi, S., & Khoshmanesh, K. (2007). Face classification by a random forest. https://doi.org/10.1109/tencon.2007.4428937
4. Degradation of Urban Green Spaces in Lagos, Nigeria: Evidence from Satellite and Demographic Data. https://www.scirp.org/journal/paperinformation.aspx?paperid=99251
5. Decision Trees in Data Science – Pompeo Pontone. https://www.pompeopontone.com/notes/decision-trees-in-data-science/
6. (2011). An Information Theoretic Approach For Feature Selection And Segmentation In Posterior Fossa Tumors. https://core.ac.uk/download/480754437.pdf
7. A. Ranjan and A. Swetapadma, "An Intelligent Computing Based Approach for Parkinson Disease Detection," 2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC), Bangalore, India, 2018, pp. 1-3, doi: 10.1109/ICAECC.2018.8479490. keywords: {Support vector machines;Artificial neural networks;Machine learning;Parkinson's disease;Biological neural networks;Machine Learning;ANN;SVM;k-NN;Parkinson’s Disease},
8. Stephen K. Van Den Eeden, Caroline M. Tanner, Allan L. Bernstein, Robin D. Fross, Amethyst Leimpeter, Daniel A. Bloch, Lorene M. Nelson, Incidence of Parkinson’s Disease: Variation by Age, Gender, and Race/Ethnicity, *American Journal of Epidemiology*, Volume 157, Issue 11, 1 June 2003, Pages 1015–1022, https://doi.org/10.1093/aje/kwg068
9. C. Quan, K. Ren and Z. Luo, "A Deep Learning Based Method for Parkinson’s Disease Detection Using Dynamic Features of Speech," in *IEEE Access*, vol. 9, pp. 10239-10252, 2021, doi: 10.1109/ACCESS.2021.3051432. keywords: {Feature extraction;Speech processing;Task analysis;Perturbation methods;Long short term memory;Handheld

computers;Computational modeling;Parkinson's disease;speech signal processing;deep learning;dynamic features;bidirectional long short term memory},

10. S. Perez-Lloret, L. Nègre-Pagès, A. Ojero-Senard, P. Damier, A. Destée, F. Tison, et al., "Oro-buccal symptoms (dysphagia dysarthria and sialorrhea) in patients with Parkinson's disease: Preliminary analysis from the French COPARK cohort", *Eur. J. Neurol.*, vol. 19, no. 1, pp. 28-37, 2012.
11. Fan Y, Yang H, Wang Y, Xu Z, Lu D. A Variable Step Crow Search Algorithm and Its Application in Function Problems. Biomimetics (Basel). 2023 Aug 28;8(5):395. doi: 10.3390/biomimetics8050395. PMID: 37754146; PMCID: PMC10526407.
12. *Yang XS. A new metaheuristic bat-inspired algorithm. Trumpington Street, Cambridge CB2 1PZ, UK: Department of Engineering, University of Cambridge; 2010.*
13. *Kennedy J, Eberhart RC. Particle swarm optimization. Proc of IEEE international conference on neural networks. 1995. p. 1942–8.*
14. *https://en.wikipedia.org/wiki/Corvus_%28genus%29.*
15. *P. Rincon, Science/nature|crows and jays top bird IQ scale*
16. Rahn, D.A.; Chou, M.; Jiang, J.J.; Zhang, Y. Phonatory impairment in Parkinson's disease: Evidence from nonlinear dynamic analysis and perturbation analysis. *J. Voice* **2007**, *21*, 64–71
17. Prediction of the rate of progression of primary open-angle glaucoma depending on gender and polymorphism of the endothelial NO-synthase (NOS3) gene | Medicni perspektivi. https://journals.uran.ua/index.php/2307-0404/article/view/271215