University of Essex

School of Computer Science and Electronic Engineering

CE888 Reassessment

Student No: 1807381

Name: Peter Osewingie

Supervisors: Dr Ana Matran – Fernandez,

Dr Raza Haider

Date: 29 April 2021

Table of Contents

Keywords: Parkinson's disease, Voice Analysis, UCI Machine learning.

Abstract

This study anchored on predicting Parkinson's disease while utilizing a voice Database that assists with treating the individuals in early phases. Parkinson's disease is a neurological issue that prompts shaking and difficulty in walking, balance, and coordination. In most pessimistic scenarios, some patients have a big challenge walking or standing to an extent they cannot live independently and require a wheelchair to move around. In contrast, help is needed for all everyday exercises. The first part of the paper expounds on the data processing procedure, normalization procedure, modelling, and analysis. The following section presents the methodology used in detecting Parkinson's disease. The last part reveals the results gathered while applying various methods. Parkinson's disease (PD) has many implications and effects on affected individuals.

The individual hears or experiences things that are not genuine, such as hallucinations and delusions. Parkinson's disease patients commonly have a low-volume voice with a droning quality. The speech pattern of Parkinson's patients frequently delivered in short overflows with unseemly hushes among words and long stops before starting discourse. The voice dataset has the variables like MDVP: FO (Hz) - Average vocal fundamental frequency, MDVP: Fhi (Hz) -Maximum vocal fundamental frequency, jitter, shimmer. The data set split into train and test, where the training dataset utilized to prepare the model. The test dataset used to test the XGB model, which delivered a higher precision of 100 per cent. The machine learning approach proved to be reliable for detecting, monitoring, and managing Parkinson's Disease. Therefore, it is helpful for patients and clinicians to monitor and manage the disease at the best accuracy level.

1. Introduction

Parkinson's disease (PD) is a disorder that attacks the brain, and it makes the patient shake, difficulties in walking, balancing and coordinating and speech disorder with an estimate of people who suffer from it to be about 10 million worldwide [1, 2]. Moreover, PD influences the cerebrum's nerve cells that produce dopamine to give indications, including muscle immovability, tremors, sync changes and speech.

Parkinson's disease impacts a human's voice, affecting them to murmur or have shuddered in talking. There are various neurodegenerative sicknesses; therefore, Alzheimer's disease is the most known complication, followed by PD. It is relied upon to increment in the coming years; subsequently, creating an identification framework for down-to-earth examination is vital for ideal treatment.

As PD's indications happen all around requested and generally, the older, noticing the disorder utilizing dysphonia assessments, has an indispensable part in the examination [3]. The characterization calculations from Artificial intelligence (AI) and machine learning utilized to predict and explore Parkinson's disease [4] [5]. The excellent highlights from the data set passed to contribute to the models to get the expected results. The assumption execution can be endorsed from the accuracy procured of the classification algorithm.

The assurance of Parkinson's ailment has logically improved the exactness boundary through further analyses. Therefore, this paper will explore the effectiveness of using supervised classification algorithms, such as the XGBoost classifier, to accurately diagnose individuals with the disease [1, 6].

The current statistical estimates of the disease attract the attention of the government and the medical sector to the increasing need for addressing the problem. Therefore, determining the prevalence and incidence helps the community facilitate the provision of necessary resources to support research in this field.

2. Background

This section extrapolates the main issue in society, for instance, Parkinson's disease that poses a significant risk to the public wellbeing.

The clinical diagnosis of Parkinson's diseases (PD) can be affirmed based on histopathologic standards [7].

Clinical indicative grouping of PD should be possible on a far-reaching survey of the writing information and determination based on the affectability and particularity of the trademark clinical features.

A report by the Parkinson's foundation highlights that 60,000 Americans are diagnosed with the disease each year while over 10 million people worldwide live with PD. Accordingly, it increases with age, and men are 1.5 times more likely to have it than women. The medications and therapeutic surgeries for PD are costly. Also, the estimated costs for treating the disease in the US is $3.3 billion yearly [8]. However, between 1994 and 2013 in the United Kingdom, the health care cost for PD patient and control was about £2471 (US$3716) per patient in the essential year postdiagnosis ($P < 0.001$), growing to £4004 (US$6021) per patient ($P < 0.001$) 10 years following decision because of more raised degrees of use across all classes of clinical consideration use.

Costs in patients with markers of forefront PD, (i.e., presence of levodopa-equivalent consistently dose > 1100 mg, dyskinesias, falls, dementia, psychosis, facility certification fundamentally in light of PD, or nursing home game plan) were on ordinary higher by £1069 (US$1608) per patient than those with PD without these markers [9].

The current literature expounds more on this research area. The area under research is Detecting Parkinson's disease using the Data-Driven Classification Model. Unique and specific brain oscillatory behaviour characterizes the resting stage as awake and sleeps in a preserved way.

The literature also reveals that the thalamocortical dysrhythmia (TCD) model provides a common underlying mechanism present in specified neurological disorders [10]. Imminent with centre pathologic examinations in the agent populace of patients demonstrating PD is expected to explore the clinical, pathologic, and nosologic studies dependent on recurrence of the event, attributes, and danger factors in patients [11, 12]. Networks, Regression, and Decision Trees recently utilized for ascertaining the presentation score of the classifiers' dependable analysis of PD [5, 13]. However, a proposed framework for detecting the early stages of Parkinson's diseases.

The data classification completed by utilizing the KNN method.

As a supervised learning algorithm, KNN requires labelled data to learn a function that provides the required output when unlabeled data entered. This algorithm is easy to use and has low calculation time,

not to mention high prediction power, therefore, commonly used in genetics, data compression, and forecasting. However, its accuracy depends on the quality of data and requires high memory to store training data. Nevertheless, the least complex technique in gathering the comparability is KNN. Among classification strategies, KNN utilized when current realities for information dissemination are insufficient [14].

This technique has two sections: a) decide k close neighbours, b) deciding class type utilizing these nearby neighbours.
The demonstration indicated a 93.7 per cent of precision for every four upgraded highlights, an exactness of 94.8 per cent per 7 streamlined highlights, and 98.2 per cent precision for nine advanced highlights is accomplished, which is a meaningful outcome contrasted with different investigations. In this study, data from the UCI archive [15].

The information incorporates 192 voice test accounts from males and females. Each subject has had six voice signal accounts. Twenty-three individuals experience the ill effects of PD, and the rest are sound. Individuals were around 46 to 85 years. The primary inconvenience of the KNN calculation is that it is a passive student; for instance, the order is depleted by using preparing information and from the training data [16].
A comparative analysis to detect PD infection utilizing different classifier. Support vector machine (SVM), feed-forward back-spread based artificial neural organization (fbann) and arbitrary tree (rt) classifiers utilized and examination between them is made to separate among PD and healthy patients. The study has utilized the UCI machine learning repository [8, 9].

The dataset comprises 195 voice tests from 31 people involving the two males and females.
From the taken subjects, twenty-three were resolved with PD and eight were healthy. In order to improve the grouping precision with a nominal error rate, a 10-fold cross-approval repeated multiple times (100) has been executed for all three classifiers.

The knn classifier has accomplished a 97.37 per cent acknowledgement exactness, consequently outflanking the other two classifiers. Moreover, [17, 18] suggested a framework to order PD and Non-PD patients by the following classifiers; binary logistic regression, linear discriminant analysis (lda), partial least square regression (pls), random tree (rnd tree) and support vector machine (svm) [18].

The Parkinson's disease dataset is retrieved from the UCI Repository. This data is extracted from patients and comprises 197 unique samples and 22 features.

Fisher separating feature choice calculation was discovered to be a viable element positioning framework. The random tree calculation accomplished 100 per cent arrangement precision while the lda, c4.5, cs-mc4 and knn yielded exactness results more noteworthy than 90 per cent. Among all, the c-pls calculation accomplished minimal precision of 69.74 per cent—the multilayer perceptron (mlp) with a back-engendering learning algorithm.

Nevertheless, there are existing gaps; for example, there are no other models and frameworks that support the data-driven classifications of a model in detecting Parkinson's disease. The frameworks from other studies must support the existing model of detecting Parkinson's disease. Furthermore, another gap is the lack of a detailed explanation regarding the techniques of identifying PD. The available approach explaining this concept is shallow; thus, it does not involve scientific methodologies.

The steps that led to identifying these gaps are the lack of enough information from the existing literature in this study area. Therefore, these gaps need to be reduced by conducting much research to help solve these gaps amicably.

3. Methodology

It is crucial to diagnose Parkinson's disease accurately and timely to control its spread among the patients. The current methods include dopaminergic imaging, which relies on Single Photon Emission Computed Tomography (SPECT) combined with I-Ioflupane for detecting Parkinson's disease in the early stages. Other studies applied grey matter, white matter, cerebrospinal fluid only. Others rely on detecting non-linear relation linking different biomarkers using deep learning and logistic regression.

Support vector machine (SVM) classifier has been reported to have an accuracy of 94%. In this case, studies applied the classification of objects by extracting the voxels from the striatum's and using partial least squares techniques.

Some studies even used SPECT images, achieving 95 - 96.2% accuracy using convolution neural networks (CNNs) [19]. Nevertheless, the studies relied on limited studies, such as only ten patients or on healthy participants.

However, this study used several tools to assemble the proficient model to detect

Parkinson's illness. Dataset used to build the model for this study retrieved from the UCI website. The Parkinson's Disease Detection Dataset has 195 unique values and 24 columns. The matrix column entries/attributes for the dataset indicate the name and the ASCII subject name and recording number.

Name - ASCII subject name and recording number.

MDVP: Fo(Hz) - Average vocal fundamental frequency

MDVP: Fhi(Hz) - Maximum vocal fundamental frequency

MDVP: Flo(Hz) - Minimum vocal fundamental frequency

MDVP: Jitter (%), MDVP: Jitter(Abs), MDVP: RAP, MDVP: PPQ, Jitter: DDP - Several measures of variation in fundamental frequency.

MDVP: Shimmer, MDVP: Shimmer(dB), Shimmer: APQ3, Shimmer: APQ5, MDVP: APQ, **Shimmer: DDA** - Several measures of variation in amplitude.

NHR, HNR - Two measures of the ratio of noise to tonal components in the voice.

Status - The health status of the subject (one) - Parkinson's, (zero) – healthy.

RPDE, D2 - Two non-linear dynamical complexity measures.

DFA - Signal fractal scaling exponent

Spread1, Spread2, PPE - Three non-linear measures of fundamental frequency variation.

The process involved data pre-processing, normalization, and finally, modelling and analysis.

Each of the tools in this study had its role driving the aim of the research. Each step is discussed under the following subsections [6].

Data Preprocessing

Data pre-processing is a crucial step in this paper as the data requires transformation into a valuable and efficient format.

The dataset does not follow a Gaussian distribution; therefore, the analysis would quickly adopt K-Nearest Neighbors or Neural Networks since this does not rely on assumptions regarding the data distribution. Otherwise, the analysis would require the standardization of data.

Therefore, this section involved handling missing data and noise from the dataset. This section entails two cycles which is normalization and adjusting the dataset, and is given in finer detail below:

Normalization

Normalization is a procedure applied as a stage of preparing a dataset for a machine learning model. The need for normalization is to adjust the numeric sections'

estimations in the dataset to a typical scale without changing contrasts in the scopes of values.

Xnew= (X- Xmin)/ (Xmax- Xmin)

Where Xnew specific component spoke to by a segment in the dataset, x is a value of this column. The column's minimum value represented as Xmin, and the maximum value of the column is Xmax. However, this study used the MinMaxScaler function from sklearn to normalize the features [6].

Modelling and Analysis

xgboost: XGBoost is a gradient boosting library that helped this research implemented machine learning algorithms under the Gradient Boosting framework. XGBoost a parallel tree boosting that solves many Machine Learning problems quickly [1, 6, 20].

The Jupyter notebook code attached demonstrates how the XGB classifier has solved this machine learning problem.

4. Results and Discussion

The research paper aimed to develop a model that detects Parkinson's disease using Data-Driven Classification.

This machine learning project analysis has utilized various factors/variables to detect Parkinson's disease. The XGB Classifier was used for the classification and used the sklearn library to prepare the dataset.

The algorithm commonly used in applied machine learning for structured data, the most crucial components of the algorithm include Boosting gradient algorithm, Stochastic Gradient Boosting, and Regularized Gradient Boosting. The algorithm has the best execution speed compared to other gradient boosting techniques. Furthermore, it provides various tuning parameters for cross-validation, regularization, missing values, and APIs.

The XGB Classifier model produced an accuracy of 100 per cent, which is excellent considering the number of lines of code and the size of the dataset in this python project. The results indicate that the machine learning approach is reliable for detecting, monitoring, and managing Parkinson’s disease.

Therefore, the model provides a way to help patients and clinicians to monitor and manage the disease accordingly.

5. Conclusion

In conclusion, this study has leveraged the XGBoost classifier's, which gives an efficient Parkinson's disease prediction model with high accuracy of 100 per cent when detecting and predicting Parkinson's

disease before getting it to most exceeding results.

Analysis of voice data is significant in the current decade to comprehend and indicative techniques for human infections.

The current technique gives the finding of PD utilizing voice dataset through machine learning algorithms. Early recognition of Parkinson's disease is valuable as it will assist with keeping the patients from the most noticeably terrible stage.

The early detection of PD is essential to help facilitate the early initiation of therapeutic interventions and strategies to manage the problem. The current study summarizes available methods for early detection, including regression, support vector machine (SVM), and other classification algorithms. At the same time, this study validates the machine learning method to have 100 per cent accuracy, the need for further analysis regarding the area under the ROC curve to classify early PD from normal and healthy individuals.

Future studies need to conduct other tests to validate this method. For instance, logistic regression on the data could be helpful with testing the statistical significance of the data to infer it as a helpful model when predicting Parkinson's disease.

6. References

[1] G. Abdurrahman and M. Sintawati, "Implementation of xgboost for classification of parkinson's disease," in *Journal of Physics: Conference Series*, 2020, vol. 1538, no. 1: IOP Publishing, p. 012024.

[2] N. Ball, W.-P. Teo, S. Chandra, and J. Chapman, "Parkinson's disease and the environment," *Frontiers in neurology*, vol. 10, p. 218, 2019.

[3] P. Gronek *et al.*, "The mechanism of physical activity-induced amelioration of Parkinson's disease: A narrative review," *Aging and disease*, vol. 12, no. 1, p. 192, 2021.

[4] Z. K. Senturk, "Early diagnosis of Parkinson's disease using machine learning algorithms," *Medical hypotheses*, vol. 138, p. 109603, 2020.

[5] T. V. Sriram, M. V. Rao, G. S. Narayana, D. Kaladhar, and T. P. R. Vital, "Intelligent Parkinson disease prediction using machine learning algorithms," *Int. J. Eng. Innov. Technol*, vol. 3, pp. 212-215, 2013.

[6] M. Al-Sarem, F. Saeed, W. Boulila, A. H. Emara, M. Al-Mohaimeed, and M. Errais, "Feature Selection and Classification Using CatBoost Method for Improving the Performance of Predicting Parkinson's Disease," in *Advances on Smart and Soft Computing*: Springer, 2021, pp. 189-199.

[7] D. J. Gelb, E. Oliver, and S. Gilman, "Diagnostic criteria for Parkinson disease," *Arch Neurol*, vol. 56, no. 1, pp. 33-9, Jan 1999, doi: 10.1001/archneur.56.1.33.

[8] L. J. Findley, "The economic impact of Parkinson's disease," *Parkinsonism & related disorders*, vol. 13, pp. S8-S12, 2007.

[9] S. Weir *et al.*, "Short-and long-term cost and utilization of health care resources in Parkinson's disease in the UK," *Movement Disorders*, vol. 33, no. 6, pp. 974-981, 2018.

[10] S. Vanneste, J.-J. Song, and D. De Ridder, "Thalamocortical dysrhythmia detected by machine learning," *Nature communications*, vol. 9, no. 1, pp. 1-13, 2018.

[11] D. Aarsland, K. Andersen, J. P. Larsen, A. Lolk, and P. Kragh-Sorensen, "Prevalence and characteristics of dementia in Parkinson disease: an 8-year prospective study," *Arch Neurol*, vol. 60, no. 3, pp. 387-92, Mar 2003, doi: 10.1001/archneur.60.3.387.

[12] D. Aarsland, K. Andersen, J. P. Larsen, and A. Lolk, "Prevalence and characteristics of dementia in Parkinson disease: an 8-year prospective study," *Archives of neurology*, vol. 60, no. 3, pp. 387-392, 2003.

[13] D. Gupta, " Optimized cuttlefish algorithm 193 for diagnosis of Parkinson's disease,," *Cognitive Systems Research,*, vol. 52, 2018.

[14] R. Mathur, V. Pathak, and D. Bandil, "Parkinson disease prediction using machine learning algorithm," in *Emerging Trends in Expert Applications and Security*: Springer, 2019, pp. 357-363.

[15] A. Benba, A. Jilbab, A. Hammouch, and S. Sandabad, "Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease," *2015 International conference on electrical and information technologies (ICEIT)*, 2015: IEEE, pp. 300-304.

[16] M. Islam, K. A. Mamun, M. Khan, and H. Deng, "Performance assessment of artificial neural network classifier for predicting movement and laterality of deep brain local field potential," in *3rd workshop on Machine learning and*

interpretation in neuroimaging (MLINI 2013) in Neural information processing systems (NIPS 2013), Nevada, USA, 2013.

[17] R. Polikar, A. Topalis, D. Green, J. Kounios, and C. M. Clark, "Comparative multiresolution wavelet analysis of ERP spectral bands using an ensemble of classifiers approach for early diagnosis of Alzheimer's disease," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 542-558, 2007.

[18] R. G. Ramani and G. Sivagami, "Parkinson disease classification using data mining algorithms," *International journal of computer applications*, vol. 32, no. 9, pp. 17-22, 2011.

[19] B. Prasad, T. Nagabhushan, and G. Pahuja, "Early Detection of Parkinson's Disease by Using SPECT Imaging and Biomarkers," 2019.

[20] J. Goyal, P. Khandnor, and T. C. Aseri, "A Comparative Analysis of Machine Learning classifiers for Dysphonia-based classification of Parkinson's Disease," *International Journal of Data Science and Analytics*, vol. 11, no. 1, pp. 69-83, 2021.