Parkinson's Disease Classification Employing a Questionnaire of Non-motor Symptoms and Machine Learning Methods

Rafael Alves Padilha[1], **Juliana Paula Felix**[1], **Rogerio Salvini**[1]

[1]Instituto de Informática, Universidade Federal de Goiás (UFG)

Abstract. *This work presents a machine learning approach to aid in the classification of Parkinson's disease (PD). Answers to a 30-question non-motor symptoms questionnaire are used as input for two classifiers that focus on differentiating Parkinson's subjects (PD) from healthy subjects and PD vs. patients with differential diagnoses. The method was evaluated using a cross-validation technique, and the results surpass those in the literature.*

1. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder affecting millions of people worldwide, primarily those of older age, by targeting the central nervous system. It is characterized by the loss of dopamine-producing neurons in the substantia nigra region of the brain, which leads to a range of motor and non-motor symptoms (Jankovic 2008). Machine learning can be a promising tool for aiding in the diagnosis of the disease, which currently has no definite diagnostic protocol. Many studies have explored this by applying methods such as vocal feature and gait pattern analysis in tandem with machine learning techniques (da Silva et al. 2024; Félix et al. 2019). This study proposes a diagnostic method by applying classical machine learning algorithms to questionnaire data addressing the disease's most prominent non-motor symptoms. The study's methodology, results, and conclusions are detailed in Sections 2, 3, and 4.

2. Material and Methods

The study utilized the *Parkinson's Disease Smartwatch Dataset (PADS)*(Varghese et al. 2024b), which comprises a stratified sample of 234 participants. Among these, 82 individuals have Parkinson's disease, 82 present with various differential diagnoses, and 70 are healthy controls. The dataset holds the answer to a series of 30 true or false questions pertaining to non-motor symptoms that can manifest in cases of Parkinson's Disease. The classification experiments were conducted in two binary classification scenarios: The first involving the classes of Healthy Controls (HC) and Parkinson's Disease patients (PD), and the second contrasting the Differential Diagnosis (DD) and Parkinson's Disease (PD) groups. The employed classical machine learning algorithms are Support Vector Machine (SVM), using a linear kernel, and Random Forest, both of which were trained using the technique of 5-fold Cross-Validation (CV). The performance of the algorithms was evaluated by calculating the mean accuracy, balanced accuracy, specificity, sensitivity and F1-score, obtained from all the validation folds.

3. Results

Table 1 report the metrics obtained for the two classification scenarios (HC *vs.* PD and DD *vs.* PD). The Random Forest Classifier, on average, performed slightly better than the SVM in most metrics. Both models achieved better results in the first scenario, with balanced accuracy averaging past 80%, in comparison to the results in the second scenario, in which both algorithms averaged 75%. To this date, the only other machine learning classification attempt making use of the PADS database is (Varghese et al. 2024a), where the authors make use of a classification pipeline featuring the CatBoost classifier algorithm and describe, in the second scenario, balanced accuracy metrics averaging around 68%, a value that was surpassed by the two classifiers in the present study.

Table 1. Average Model Performance - HC *vs*. PD and DD *vs*. PD.

Model	Accuracy	Balanced Accuracy	Specificity	Sensitivity	F1-Score
HC *vs*. PD					
SVM	0.82	0.82	0.84	0.80	0.81
Random Forest	0.86	0.87	0.93	0.80	0.86
DD *vs*. PD					
SVM	0.72	0.75	0.84	0.65	0.74
Random Forest	0.75	0.75	0.81	0.70	0.75

4. Conclusion

The present work deals with a dataset that consists of questionnaire data pertaining to non-motor symptoms of Parkinson's Disease, answered by 234 subjects, including healthy controls (HC), Parkinson's Disease patients (PD) and people with differential diagnoses (DD). The data was split in two binary classification scenarios, one involving the HC and PD groups, and the other the DD and PD groups, then fed to classifier models implementing the Support Vector Machine (SVM) and Random Forest algorithms. Both models performed well in the first scenario, which saw the Random Forest classifier performing slightly ahead, and the results achieved in the second scenario surpass the results reported by another study (Varghese et al. 2024a) using the same database.

Acknowledgments

The authors thank CAPES, through Edital Nº 30/2022 - PDPG- Solidariedade Acadêmica.

References

[da Silva et al. 2024] da Silva, M. I., Felix, J. P., de Stecca Prado, T., de Bastos Chagas, A. L., Bucci, G. d. F. F. B., da Fonseca, A. U., and Soares, F. (2024). Sobre a análise de sinais de voz para o diagnóstico da doença de parkinson. *Journal of Health Informatics*, 16(Especial).

[Félix et al. 2019] Félix, J. P., Vieira, F. H., Cardoso, Á. A., Ferreira, M. V., Franco, R. A., Ribeiro, M. A., Araújo, S. G., Corrêa, H. P., and Carneiro, M. L. (2019). A Parkinson's disease classification method: an approach using gait dynamics and detrended fluctuation analysis. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pages 1–4. IEEE.

[Jankovic 2008] Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(4):368–376.

[Varghese et al. 2024a] Varghese, J., Brenner, A., Fujarski, M., van Alen, C. M., Plagwitz, L., and Warnecke, T. (2024a). Machine Learning in the Parkinson's disease smartwatch (PADS) dataset. *npj Parkinson's Disease*, 10(1):9.

[Varghese et al. 2024b] Varghese, J., Brenner, A., Plagwitz, L., van Alen, C., Fujarski, M., and Warnecke, T. (2024b). PADS-Parkinsons Disease Smartwatch dataset. *physionet.org*.