Robustness and Cross-Dataset Performance of Machine Learning Models in Parkinson's Disease Diagnosis

Ana Luísa B. Chagas[1], Giordana de Farias F. B. Bucci[1] Pedro L. S. Lobo[1], Rogerio Salvini[1], Fabrizzio Soares[1], Juliana Felix[1,2]

[1]Instituto de Informática, Universidade Federal de Goiás (UFG)
[2]Escola Politécnica e de Artes, Pontifícia Universidade Católica de Goiás (PUC-Goiás)

{analuisa23, giordanabucci, pedro_lemes}@discente.ufg.br,
{rogeriosalvini, fabrizzio, julianafelix}@ufg.br

Abstract. *Parkinson's disease (PD) is a progressive neurodegenerative disorder whose diagnosis remains largely based on subjective clinical evaluation. In this study, we examine the generalization capacity of machine learning models for PD diagnosis from gait data. Descriptive features extracted from force signals were used to train multiple classifiers, evaluated through intra-dataset and cross-dataset experiments. Results showed consistent performance, particularly for ensemble methods such as ExtraTrees, demonstrating the robustness of the proposed approach. These findings highlight the importance of cross-dataset validation for assessing the real-world applicability of diagnostic models.*

1. Introduction

Parkinson's Disease (PD) is a progressive and incurable neurodegenerative disorder characterized by neuronal death and impaired synaptic communication [Poewe et al., 2017]. Among its most relevant symptoms are motor alterations, including tremors, rigidity, bradykinesia, and postural instability, which often affect gait [Braak and Braak, 2000]. Due to the absence of conclusive diagnostic tests, diagnosis is predominantly based on clinical observation and motor evaluations, which can lead to late detection and limit therapeutic interventions [Mayeux, 2003].

With the advancement of computational techniques, Machine Learning (ML) has shown great promise in diagnosing PD through gait data, enabling the identification of complex patterns and distinguishing patients from healthy individuals [Beyrami and Ghaderyan, 2020; Balaji et al., 2020]. However, the generalization and applicability of these models remain challenging, as differences between data (variations in collection protocols, sensors used, and sample characteristics) used to train models and data collected from real-world environments, such as hospitals, clinics, or rehabilitation centers, can significantly affect their performance. Cross-database experiments, in which a model is trained on one dataset and tested on another, are one way to assess this generalization, but there are very few works exploring this in literature, such as Joshi et al. [2017]. This evaluation is particularly important because, in real clinical settings, data collection conditions rarely replicate the exact protocols used in the original dataset. Thus, testing models on different datasets or under conditions that simulate real-world variations allows researchers to identify limitations, enhance robustness, and ensure that machine learning methods can be reliably applied beyond the original experimental context.

In this study, we introduce a novel feature set extracted from gait signals and evaluate model generalization by analyzing the performance of multiple machine learning

algorithms both within a single dataset and across distinct datasets, which differ in sensor types, sampling frequencies, and data collection protocols. This approach allows us to assess not only intra-dataset performance but also cross-database robustness, providing a more realistic indication of how these models may perform in clinical scenarios where data acquisition conditions vary and rarely match those of the original datasets.

The remainder of this work is organized as follows: the materials and methods used in the study are described in Section 2, the results are presented in Section 3, and the conclusions are discussed in Section 4.

2. Materials and Methods

This study was developed in Python 3.10.12 using Google Colaboratory. Data analysis employed the Tsfresh library for feature extraction and Scikit-learn for implementing and evaluating classification models.

2.1. Databases

The Parkinson's Disease Gait Database (GaitPDB)[1] [Hausdorff, 2008], publicly available on PhysioNet [Goldberger et al., 2000], was used as the main dataset to evaluate the proposed strategy. It includes gait recordings from individuals with PD (93 participants) and healthy control subjects (50 participants). The data consist of time series of vertical ground reaction forces measured by eight sensors under each foot, as participants walked at their usual pace for about two minutes, sampled at 100 Hz. In this study, we used the total force signal from both feet, obtained by summing the eight sensor outputs per foot, as provided in the database.

To assess cross-dataset generalization, we used data available from the Gait in Neurodegenerative Diseases Database (GaitNDD) [Hausdorff, 2000; Hausdorff et al., 2000][2], publicly available on PhysioNet. This smaller dataset includes gait recordings from 15 participants with PD and 16 healthy control subjects. Recordings were obtained using instrumented insoles with force-sensitive resistors under each foot, providing one time series for the right and one for the left foot, with amplitudes roughly proportional to the applied force. The data were sampled at 300 Hz and are available for up to five minutes per participant. For this experiment, no downsampling was performed, and only the first two minutes were analyzed to ensure temporal consistency with the GaitPDB dataset.

2.2. Data windowing, Feature Extraction and Balancing

For feature extraction, the first 20 seconds of each recording from both databases were excluded to avoid initialization noise at the start of walking. The remaining time series corresponding to the gait signals from each foot were segmented into smaller windows of 5 seconds to increase the number of data instances. From each window, a set of descriptive features was extracted, namely: the fundamental frequency, harmonic amplitudes (first, second, and third), harmonic distortions (second and third), total harmonic distortion (THD), mean, standard deviation, kurtosis, skewness, and power.

Since these features were computed separately for each foot, each data window was represented by a feature vector of 24 columns. After extraction, the Synthetic Mi-

[1] https://physionet.org/content/gaitpdb/1.0.0/
[2] https://physionet.org/content/gaitndd/1.0.0/

nority Over-sampling Technique (SMOTE) was applied to balance the classes, ensuring proper representation of both groups for model training.

2.3. Classification and Evaluation

In the classification stage, different supervised learning algorithms were evaluated, such as Linear SVM (linear kernel), KNN (5 neighbors), Logistic Regression, HistGradient Boosting, Random Forest, LightGBM, and ExtraTrees. Model training and validation were performed using five-fold cross-validation to ensure greater statistical robustness. The performance of each model was assessed through accuracy, sensitivity, specificity, and F1-score metrics, providing a comprehensive evaluation of the models true capability. In this context, two experiments were conducted. In the first one, the GaitPDB dataset was used both for training and evaluation of the classifiers through five-fold cross-validation. In the second experiment, aimed at assessing the robustness of the classifiers, each model was trained using data from GaitPDB and tested using all data from GaitNDD.

3. Results and Discussion

The results obtained after applying the new feature set are in Table 1. Overall, solid performance was achieved, with the ensemble-based classifiers, particularly ExtraTrees and HistGradient Boosting, showing the highest metrics. ExtraTrees reached an accuracy of 80.96%, sensitivity of 86.69%, specificity of 72.97%, and F1-score of 84.14% when trained and tested on the same dataset (GaitPDB). KNN also performed notably well, with an accuracy of 74.25% and F1-score of 76.78%, suggesting that even relatively simple models benefit from the new descriptive features extracted from the gait signals.

The consistency observed between intra-dataset evaluation and cross-dataset testing underscores the potential robustness of these models in practical scenarios. When trained on GaitPDB and tested on GaitNDD, the performance metrics remained relatively close, with ExtraTrees achieving 80.69% accuracy and 80.50% F1-score. Despite the difference in sampling frequency, the stability of these results is noteworthy because cross-database evaluation is rarely explored in literature, yet it closely reflects real clinical conditions, where data collection protocols, sensor types, and patient behavior cannot be guaranteed to match the original dataset. Consequently, these experiments provide a meaningful assessment of model generalization, suggesting that the proposed feature set and algorithms can maintain reliable performance even under varying conditions.

Table 1. Results obtained from intra-dataset and cross-dataset experiments.

	Train and Test in GaitPDB				**Train in GaitPDB / Test in GaitNDD**			
Algorithm	Acc.(%)	Sens.(%)	Spec.(%)	F1(%)	Acc.(%)	Sens.(%)	Spec.(%)	F1(%)
SVM Linear	67.01	63.77	71.53	69.21	65.54	65.54	67.92	65.74
KNN	74.25	73.21	75.70	76.78	75.53	75.53	75.78	75.65
Logistic Regression	66.78	64.46	70.01	69.27	65.14	65.14	67.04	65.33
HistGradient Boosting	80.02	83.12	75.70	82.87	79.45	79.45	74.26	79.41
Random Forest	79.48	83.93	73.30	82.65	79.35	79.35	72.17	79.23
LightGBM	78.61	81.91	74.01	81.68	78.88	78.88	73.54	78.82
ExtraTrees	80.96	86.69	72.97	84.14	80.69	80.69	72.01	80.50

4. Conclusion

This study presented an investigation into the performance of machine learning models for Parkinson's disease diagnosis using a novel feature set extracted from gait data. The

results obtained across the two datasets were found to be closely aligned, suggesting a level of robustness and consistency of these models when applied to different data sources. This experiment emphasizes the relevance of testing models in conditions that mimic real clinical environments, where data collection may vary and not always occur under the same controlled settings as in the original databases. Such evaluations are essential to understand how these models might perform in practical, everyday clinical use.

Acknowledgments

The authors would like to acknowledge the support of CNPQ – Conselho Nacional de Desenvolvimento Científico, CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Financing Code #001, and CAPES/PDPG n. 30/2022 – Programa Emergencial de Solidariedade Acadêmica.

References

Balaji, E., Brindha, D., and Balakrishnan, R. (2020). Supervised machine learning based gait classification system for early detection and stage classification of Parkinson's disease. *Applied Soft Computing*, 94:106494.

Beyrami, S. M. G. and Ghaderyan, P. (2020). A robust, cost-effective and non-invasive computer-aided method for diagnosis three types of neurodegenerative diseases with gait signal analysis. *Measurement*, 156:107579.

Braak, H. and Braak, E. (2000). Pathoanatomy of Parkinson's disease. *Journal of Neurology*, 247:II3–II10.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Hausdorff, J. (2000). Gait in neurodegenerative disease database. https://physionet.org/content/gaitndd/1.0.0/. Last acessed in October 5th 2025.

Hausdorff, J. (2008). Gait in parkinson's disease database. https://physionet.org/content/gaitpdb/1.0.0/. Last accessed in October 5th 2025.

Hausdorff, J. M., Lertratanakul, A., Cudkowicz, M. E., Peterson, A. L., Kaliton, D., and Goldberger, A. L. (2000). Dynamic Markers of Altered Gait Rhythm in Amyotrophic Lateral Sclerosis. *Journal of Applied Physiology*, 88(6):2045–2053.

Joshi, D., Khajuria, A., and Joshi, P. (2017). An automatic non-invasive method for parkinson's disease classification. *Computer Methods and Programs in Biomedicine*, 145:135–145.

Mayeux, R. (2003). Epidemiology of neurodegeneration. *Annual Review of Neuroscience*, 26(1):81–104.

Poewe, W., Seppi, K., Tanner, C. M., Halliday, G. M., Brundin, P., Volkmann, J., Schrag, A.-E., and Lang, A. E. (2017). Parkinson disease. *Nature Reviews Disease Primers*, 3(1):17013.