Exploring Biases in Machine Learning Models for Neurodegenerative Diseases Diagnosis Through Gait and Voice Analysis

Ana Luísa de Bastos Chagas[1], **Giordana de Farias F. B. Bucci**[1], **Juliana Paula Félix**[1,2], **Rogerio Salvini**[1], **Hugo Nascimento**[1], **Fabrizzio Soares**[1]

[1]Instituto de Informática, Universidade Federal de Goiás
[2]Escola Politécnica e de Artes, Pontifícia Universidade Católica de Goiás

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

Abstract. *This work examines potential biases in machine learning models for diagnosing neurodegenerative diseases (NDDs) through gait and voice analysis. It investigates how common techniques, such as gait signal windowing and using multiple voice samples per individual indiscriminately, can lead to inflated performance estimates when samples from the same individual are treated independently. Using two public databases, it compares scenarios where augmented samples are treated independently versus grouped by individual. Results show that independent treatment leads to artificially higher performance metrics. The findings highlight the need for proper intra-individual variability handling to ensure reliable clinical decision support for NDD diagnosis.*

1. Introduction

Neurodegenerative Diseases (NDDs), such as Parkinson's (PD), Huntington's (HD), and Amyotrophic Lateral Sclerosis (ALS), are progressive, incurable, and life-threatening, causing neuronal deterioration. Despite unique features of each disease, patients often experience memory loss, involuntary movements, mobility problems, speech issues and unstable gait [Berman and Bayati 2018]. Since NDDs like PD and ALS lack definitive diagnostic tests, diagnosis usually relies on clinical symptom observation, often resulting in late detection and limited treatment options [Mayeux 2003]. Given the debilitating nature of these diseases, early detection can significantly impact patient outcomes by enabling timely interventions, improving quality of life, and optimizing healthcare resources. Developing more accurate, alternative diagnostic methods is therefore highly desirable.

Several studies have demonstrated the potential of machine learning models to analyze complex patterns in gait [Modaresnia et al. 2024, Fraiwan and Hassanin 2021] and voice data [Little et al. 2009, Ouhmida et al. 2021], achieving high accuracy in differentiating between healthy individuals and those with NDDs. These results highlight the growing role of machine learning (ML) as a valuable tool to support clinicians in the diagnostic process. However, a closer examination of these previous studies reveals potential methodological limitations that may inflate performance metrics.

The scarcity of data for rare diseases hinders the training of robust machine learning models. To mitigate this, data augmentation techniques are commonly used to expand training samples. In gait analysis, walking sequences are segmented into smaller windows (data windowing), as done in [Fraiwan and Hassanin 2021, Modaresnia et al. 2024]

while vocal assessments collect multiple speech samples per individual when composing the database [Little et al. 2009, Ouhmida et al. 2021]. While these techniques help increase the amount of training data, failing to properly separate samples by their individual source can introduce bias, as models may learn to recognize personal characteristics rather than generalizable disease patterns if samples originated from the same individual end up in both training and testing phases.

Previous research [Felix et al. 2022] has suggested that this issue can lead to overly optimistic performance estimates, not reflecting the real capacity of the model to diagnose. In this sense, this undergraduate project focused on investigating this said bias in gait and vocal samples domain, and assess its implications for the real-world efficiency of machine learning-based diagnostic models for NDDs. By analyzing both domains, and taking into consideration that these are distinct areas of study, this study highlights that the bias may persist even across different modalities. First, three classification tasks are conducted, investigating this bias across three different diseases (PD, HD, ALS) on gait domain. Second, this study proceeds on analyzing PD vs. CO (control) using voice data.

The remainder of this work is organized as follows: Section 2 describes the proposed method. Results and a following discussion are presented in Sections 3 and 4. Our conclusions are drawn in Section 5.

2. Materials and Methods

Python 3.10.12 was used to develop this work, with Google Colaboratory as the environment. Libraries utilized include *tsfresh* for feature extraction and *scikit-learn* for classification tasks. The following two subsections present separate experiments conducted on gait and voice data, each designed to assess classification performance under the two different evaluation scenarios that will be presented below.

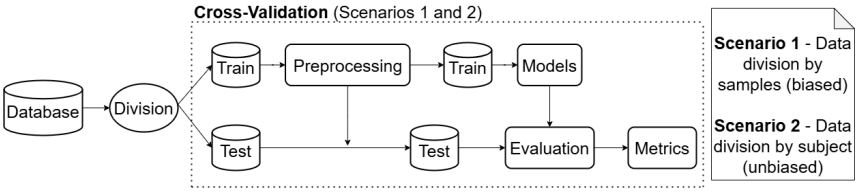

Figure 1. Flowchart of the methodology applied in this work.

2.1. Gait data

For gait-related experiments, the publicly accessible Gait in Neurodegenerative Diseases Database[1] was used, containing walking data from 15 individuals with PD, 20 individuals with HD, 13 individuals diagnosed with ALS, and 16 healthy controls. The gait data were collected using force sensors, placed on both the left and right feet.

The stance phase of gait, which refers to the period during which the foot is in contact with the ground, supporting the body's weight, is already provided as a pre-processed time series in the database. From this stance interval time series derived from the right foot, a total duration of one minute of walking is considered in the experiments. As recommended by the dataset authors, the first 20 seconds are discarded to mitigate potential initialization effects in walking, resulting in 40 seconds of effective gait data per patient.

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

These remaining 40 seconds are divided into four non-overlapping 10-second windows. In this way, 64 subjects generated a total of 252 samples. From each window, four features are extracted: mean, standard deviation, entropy, and average signal power. The resulting feature set is then used as input to train the following machine learning models: Support Vector Machine (SVM) with a linear kernel, k-Nearest Neighbors (KNN) with k=5, Naïve Bayes (NB), Linear Discriminant Analysis (LDA), and Decision Tree (DT). These models are applied to perform the following classification tasks, including PD vs. Control (CO), HD vs. CO, ALS vs. CO, and NDD vs. Control (CO), where NDD represents a combined class including all three neurodegenerative disease groups.

The experiment is conducted under two distinct scenarios. In Scenario 1, each individual sample (10-second window) is treated independently. Thus, in each classification task, Leave-One-Out Cross-Validation (LOOCV) is applied over n samples, where n corresponds to the total number of windows generated by the data classes involved. LOOCV is particularly suitable for this assessment as it maximizes data usage and provides a reliable estimate of model performance, crucial given the limited class sizes. In Scenario 2, samples follow the Leave-One-Subject-Out rule, where the process is based on the total number of participants in each classification task. In this method, all samples from a single participant are used exclusively for testing, while the samples from the remaining participants constitute the training set. Consequently, each participant selected for the training/testing phase contributes to four data samples.

2.2. Voice data

On the voice domain, the Parkinson Dataset With Replicated Acoustic Features[2] was used for the classification of Parkinson's Disease vs. healthy controls. This dataset includes 80 participants, with 40 diagnosed with PD and 40 healthy controls. Three voice samples per subject were recorded, resulting in 210 total recording samples. Each sample consists of 27 different features directly available in the database, such as pitch variation, amplitude perturbations, Harmonic-to-Noise Ratios (HNR), among other signal-related characteristics. All 27 features are used as input to feed the same five machine learning models used for the gait experiment (SVM, KNN, LDA, NB, DT). The evaluation follows a 5-fold cross-validation approach, considering the two main scenarios previously described. The 5-fold validation is now used instead of LOOCV because this dataset contains a larger number of patients per class. Thus, it is no longer necessary to handle the computational cost of LOOCV (considerably higher than the current approach) to obtain a reliable result.

For this voice-related experiment, Scenario 1 evaluates samples independently, allowing multiple samples from the same individual in both training and testing sets, while Scenario 2 assigns all samples from a person to either set, exclusively.

3. Results

Table 1 shows the results for the four classification tasks and both assessment scenarios for the gait-related experiment. In Scenario 1, where data samples were treated independently, DT outperformed all models, achieving up to 88.79% accuracy (ALS vs. CO), while SVM performed the worst, even misclassifying all control samples in some cases

[2]https://archive.ics.uci.edu/dataset/489/parkinson+dataset+with+replicated+acoustic+features

Table 1. Results obtained for different classifications (gait data).

	PD vs CO			HD vs CO			ALS vs CO			NDD vs CO		
Scenario 1 – leaving one sample out												
	Acc. (%)	Sens. (%)	Spec. (%)	Acc. (%)	Sens. (%)	Spec. (%)	Acc. (%)	Sens. (%)	Spec. (%)	Acc. (%)	Sens. (%)	Spec. (%)
SVM	66.94	53.33	79.60	43.57	43.42	43.75	77.59	57.69	93.75	74.60	100.00	0.00
KNN	74.19	65.00	82.81	67.86	59.21	78.12	84.48	76.92	90.62	76.19	81.91	59.38
NB	72.58	53.33	90.62	77.86	67.11	90.62	77.59	59.62	92.19	66.67	58.51	90.62
LDA	79.03	73.33	84.38	79.29	67.11	93.75	81.03	65.38	93.75	74.60	93.62	18.75
DT	81.45	80.00	82.81	80.71	82.89	78.12	88.79	86.54	90.62	84.52	90.43	67.19
Scenario 2 – leaving one subject out												
SVM	62.90	45.00	79.69	22.14	40.79	0.00	77.59	57.69	93.75	74.60	100.00	0.00
KNN	71.77	61.67	81.25	61.43	51.32	73.44	81.90	71.15	90.62	73.02	79.79	53.12
NB	72.58	53.33	90.62	77.86	67.11	90.62	78.45	59.62	93.75	66.67	58.51	90.62
LDA	79.03	75.00	82.81	75.00	63.16	89.06	80.17	63.46	93.75	72.62	91.49	17.19
DT	70.16	61.67	78.12	73.57	76.32	70.31	85.34	84.62	85.94	81.75	87.77	64.06

(NDD vs. CO). However, Scenario 2, leaving one subject out, showed overall lower accuracy, sensitivity, and specificity, suggesting that considering gait samples originated from the same individual as separate entities results in overestimated performance metrics, biasing results. Notably, LDA performed best for PD vs. CO (79.03% acc.), while NB stood out for HD vs. CO (77.86% acc.). These findings highlight the importance of careful training-test separation to ensure robust classification for neurodegenerative disease diagnosis.

Results for PD vs. CO under both assessment scenarios for the voice-related experiment are presented in Table 2. Accuracy varied between 66.67% and 83.33%, with NB and KNN showing the strongest performance in both cases. Scenario 1, where replicated voice data were handled as independent samples, achieved higher results across all evaluation measures. In Scenario 2, where it is ensured that the source individual plays a role during the training and testing distribution phase, performance dropped considerably. This also implies that treating data from the same subject as independent may introduce bias, potentially compromising the model's validity and reliability.

Table 2. Results obtained for PD *vs.* CO (voice data).

	Scenario 1 (*5 K-fold* per samples)			Scenario 2 (*5 K-fold* per subject)		
Algorithm	Acc.(%)	Sens.(%)	Spec.(%)	Acc.(%)	Sens.(%)	Spec.(%)
SVM	74.58	75.82	73.18	70.00	69.60	71.48
KNN	82.50	81.69	83.39	72.08	70.62	74.05
NB	83.33	81.65	84.96	82.92	81.42	84.44
LDA	76.25	75.71	76.66	66.67	65.91	68.24
DT	71.67	69.91	73.35	66.67	66.22	67.33

4. Discussion

The results suggest that, no matter the data domain (gait or voice), the chosen validation strategy can significantly influence reported accuracy rates for NDD diagnosis. In particular, the difference between the tested scenarios suggests that common practices in the

literature (as done in previous works with high accuracy reported rates, such as 99.91% and 99.17% in [Modaresnia et al. 2024, Fraiwan and Hassanin 2021] for gait data, and 91.40% and 93.10% in [Little et al. 2009, Ouhmida et al. 2021] for voice data) may overestimate the actual performance of models if intra-subject variability is not properly considered. This observation raises questions about model generalization, in special, for NDD diagnosis, and highlights the need for more rigorous validation approaches aiming diseases diagnosis. Thus, when comparing our findings with previous studies that report high accuracy rates, it is possible that, if the same methodological constraints applied here were used, those results values would likely be lower.

Thus, given that the unbiased scenario, in both data domains, generally presents inferior results, it is inferred that the typical technique of considering samples independently for training and testing – when they originate from the same individual – may induce bias in machine learning classifiers, artificially inflating their results. This raises significant concerns about the reliability of existing algorithms intended to support NDD diagnosis, as even seemingly minor methodological choices in data handling can lead to overly optimistic performance estimates. Our experiments demonstrate that extra precaution must be taken when developing and validating artificial intelligence systems for these diseases, particularly given the serious implications of misdiagnosis in clinical settings.

5. Conclusion

This work investigated potential biases in machine learning approaches for diagnosing neurodegenerative diseases, focusing specifically on how data augmentation techniques and the use of replicated samples affect model reliability. Through experiments with gait signals and voice recordings from two public databases, we evaluated how treating windowed gait samples and multiple voice recordings from the same individual as independent instances impacts classification results and the reliability of health-related decision support systems. Five algorithms (SVM, KNN, Naive Bayes, LDA, and Decision Trees) were tested under two validation scenarios to understand whether treating samples individually might lead models to learn individual characteristics rather than disease patterns.

By addressing this issue, this study contributes to the development of more accurate and trustworthy ML models, which are crucial in assisting healthcare professionals with early and precise NDD diagnoses. The insights from this research reinforce the importance of rigorous validation strategies to ensure that AI-driven diagnostic tools provide meaningful benefits in real-world healthcare settings. Findings highlight how standard ML practices may need to be adapted when dealing with medical data to ensure reliable clinical decision support and effective implementation in medical workflows.

In addition to the scientific contributions listed above, this research has led to a journal article [da Silva et al. 2024], full papers presented in national and international conferences [Chagas et al. 2024b, Felix et al. 2025], and a poster awarded as the best at a local event [Chagas et al. 2024a], all with the active participation of the undergraduate student authors. The project is still ongoing, and further investigations are planned to explore additional data sources and refine methodologies to mitigate biases in ML models for NDD diagnosis.

Acknowledgments

The authors thank CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Financing Code #001; CAPES/PDPG n. 30/2022 – Programa Emergencial de Solidariedade Acadêmica; and CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico.

References

Berman, T. and Bayati, A. (2018). What are neurodegenerative diseases and how do they affect the brain? *Frontiers for Young Minds*, 6.

Chagas, A., Lobo, P. S., Felix, J., do Nascimento, H., and Salvini, R. (2024a). Analyzing the impact of voice data replication on machine learning models for parkinson's disease diagnosis. In *Anais da XII Escola Regional de Informática de Goiás*, pages 263–264, Porto Alegre, RS, Brasil. SBC.

Chagas, A. L., Bucci, G., Felix, J., Fonseca, A., Nascimento, H., and Soares, F. (2024b). Avaliando a sobreamostragem de dados temporais de marcha no diagnóstico automático de doenças neurodegenerativas. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS), Goiânia, GO, Brazil, June 25–28, 2024*, pages 1–12. SBC.

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).

Felix, J., da Silva, M. I., Chagas, A. L., Salvini, R., Nascimento, H., and Soares, F. (2025). Analyzing the effect of replicated voice samples in Parkinson's disease classification. In *2025 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–5, Vancouver, Canada. IEEE. To appear.

Felix, J., Fonseca, A. U., Nascimento, H., and Guimarães, N. (2022). Rede neural multicamadas para classificação de doenças neurodegenerativas a partir de sinais de marcha. In *Anais do XXIV Congresso Brasileiro de Automática*, pages 1354–1361. SBA.

Fraiwan, L. and Hassanin, O. (2021). Computer-aided identification of degenerative neuromuscular diseases based on gait dynamics and ensemble decision tree classifiers. *Plos one*, 16(6):e0252380.

Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., and Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4):1015–1022.

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

Modaresnia, Y., Torghabeh, F. A., and Hosseini, S. A. (2024). A deep time-frequency approach in automated diagnosis of neurodegenerative diseases using gait signals. *Basic and Clinical Neuroscience*, 15(6):759–774. [Online].

Ouhmida, A., Fattah, J., Khaireddin, Y., and Maaroufi, M. (2021). Voice-based deep learning medical diagnosis system for parkinson's disease prediction. In *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*. IEEE.