Analyzing the Impact of Voice Data Replication on Machine Learning Models for Parkinson's Disease Diagnosis

Ana Luísa B. Chagas[1], **Pedro L. S. Lobo**[1], **Juliana P. Felix**[1], **Hugo A. D. do Nascimento**[1], **Rogerio Salvini**[1]

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

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

Abstract. *This study examines the effect of voice data replication on machine learning models for Parkinson's Disease diagnosis. Using a dataset of 80 individuals, we compare two evaluation scenarios: treating voice samples as independent and considering the source individual when composing training and test sets. Results show that treating replicated samples as independent leads to inflated performance metrics, highlighting the importance of properly handling intra-individual variability in PD diagnosis models.*

1. Introduction

Parkinson's Disease (PD) is a progressive and incurable neurodegenerative disorder caused by neuronal death, leading to impaired synaptic communication. Among the symptoms of PD, speech disorders stand out as a significant symptom [Braak and Braak 2000]. Machine learning has emerged as a promising avenue for research, particularly in using voice signal analysis to aid in diagnosis, as demonstrated by previous studies [Idrisoglu et al. 2023]. However, an important aspect was previously highlighted by [Naranjo et al. 2016], where intra-individual variability was statistically analyzed, and they suggested the need to address voice-features from the same individual differently when performing statistical experiments. Thus, this work aims to analyze this bias through a machine learning aspect. We believe that mishandling replicated samples can artificially inflate sample sizes and compromise result generalization. The next section details our materials and the method used for this analysis. Results follow in Section 3, and conclusions are drawn in Section 4.

2. Materials and Methods

In this work, we used the *Parkinson Dataset With Replicated Acoustic Features* [Prez 2016] database, that contains data from 80 subjects, of which 40 were diagnosed with Parkinson's Disease and 40 are healthy control subjects (CO). Three voice samples from each subject were recorded, totaling 210 samples available in the database. Each sample is composed of 27 varied features, including pitch local and amplitude perturbations, Harmonic-to-Noise Ratios (HNR) and other signal features. All 27 features are used as input to five classical machine learning algorithms: Support Vector Machine (SVM – with a linear kernel), K-Nearest Neighbors (K=5), Naive Bayes (NB), Linear Discriminant Analysis (LDA) and Decision Tree (DT). The classifiers are evaluated using the 5-fold cross-validation technique. However, two scenarios of evaluation are proposed in order to analyze the effects of voice-data replication: i) Case 1, where we iterate over the data samples without taking into consideration their original subject, and ii) Case 2, where the iteration considers the source individual, ensuring that data from voice samples from the same person do not appear simultaneously in the training and testing stages. For both scenarios, the mean accuracy, specificity and sensitivity are calculated and compared.

3. Results

Table 1 shows the results achieved for the classification of Parkinson's Disease vs. Control subjects for both scenarios. Accuracy ranging from 66.67% to 83.33% was achieved, with Naive Bayes and K-Nearest Neighbors having the best overall performance in both scenarios. The results confirm that the first scenario (Case 1), in which replicated voice-data were treated as distinct samples, presented better results in all three evaluation metrics. In contrast, when the problem of voice-replicated data is dealt by ensuring the source individual plays a role during the training and testing distribution phase (Case 2), the classification results are significantly lower. This suggests that treating distinct samples from the same original subject as independent data has the potential to introduce bias into classification models, affecting their validity and overall reliability.

Table 1. Results obtained for PD *vs.* CO.

	Case 1 (*5 K-fold* per samples)			Case 2 (*5 K-fold* per subject)		
Algorithm	Accuracy(%)	Sens.(%)	Spec.(%)	Accuracy(%)	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. Conclusion

In this work, a dataset comprised by acoustic features derived from three voice recordings collected from 80 individuals was used in an experiment to analyze the effects of voice replication from the same original individual on the outcome trends of machine learning algorithms applied to PD diagnosis. Results align with the trends observed in previous studies, reinforcing the notion that treating distinct samples from the same subject as independent data can introduce bias into classification models, undermining their reliability.

Acknowledgments

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

References

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

Idrisoglu, A., Dallora, A. L., Anderberg, P., and Berglund, J. S. (2023). Applied Machine Learning Techniques to Diagnose Voice-Affecting Conditions and Disorders: Systematic Literature Review. *J Med Internet Res*, 25:e46105.

Naranjo, L., Pérez, C. J., and Martín, J. (2016). Addressing voice recording replications for tracking Parkinson's disease progression. *Medical Biological Engineering Computing*, 55(3):365–373. Epub 2016 May 21.

Prez, C. (2016). Parkinson Dataset with replicated acoustic features. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5701F.