Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230079

Classification of Parkinson's Disease from Voice - Analysis of Data Selection Bias

Alexander BRENNER[a,1], Catharina Marie VAN ALEN [a], Lucas PLAGWITZ [a] and Julian VARGHESE [a]
[a] *Institute of Medical Informatics, University of Münster, Münster, Germany*

Abstract. A growing number of studies have been researching biomarkers of Parkinson's disease (PD) using mobile technology. Many have shown high accuracy in PD classification using machine learning (ML) and voice records from the mPower study, a large database of PD patients and healthy controls. Since the dataset has unbalanced class, gender and age distribution, it is important to consider appropriate sampling when assessing classification scores. We analyse biases, such as identity confounding and implicit learning of non-disease-specific characteristics and present a sampling strategy to highlight and prevent these problems.

Keywords. Machine Learning, Parkinson's Disease, Selection Bias

1. Introduction

The incidence and global burden of Parkinson's disease (PD) is increasing [1]. PD is associated with a variety of symptoms, including tremors, rigidity, changes in speech and gait, and non-motor symptoms. Early treatment can reduce burden, but screening for PD can be time-consuming. The mPower study provides a large dataset of PD patients and healthy controls (HC) including voice records from smartphones [3]. Based on this data various studies reported accuracies of up to 90% in the distinction of PD from HC using machine learning (ML) [1,2]. Still, a common problem are repetitive samples of the same individual. Many analyses lack declaration of subject-wise train/test splits, which may lead to identity confounding. In addition, controls should represent clinical practice. Otherwise, models may differentiate people by age instead of disease-specific patterns. We present an approach to identify bias and to derive fair scores with stratified sampling.

2. Methods

The mPower dataset holds records of participants vocalising the phoneme "aaah" [3]. We arranged the classes similar to Tracy et al. [2], but not considering UPDRS (Unified Parkinson's Disease Rating Scale) scores. We designed the stratified sampling to align datasets to a desired distribution of attributes by iteratively removing samples [4]. Our goal was to balance gender and classes, and match the age-distribution (10-years bins). Trade-off steps kept a minimal fraction of the dataset. To control for accuracy loss due to reduced train set size, we additionally assembled sets from random subsampling.

[1] Corresponding Author: Alexander Brenner, E-mail: alexander.brenner@uni-muenster.de.

We used the following feature sets from Surfboard [5] and OpenSmile [6]: 1) Surfboard PD, 2) eGeMAPS, 3) AVEC, and 4) ComParE. We tested two classifiers comparable to methods from previous studies, the Support-Vector-Machine (SVM) and CatBoost.

3. Results

We report accuracy for each combination of data and feature set using 5-fold cross-validation (Table 1). Results for CatBoost and further metrics are in the supplements [4].

Table 1. Mean classification accuracy (in %) on mPower subsets using the SVM and 5-fold CV (± std).

Setting	Surfboard PD	ComParE	AVEC	eGeMAPS
Complete set, CV	83.17 (0.65)	82.41 (0.18)	83.14 (0.47)	76.50 (0.67)
Complete set, grouped CV	71.29 (1.50)	71.45 (1.75)	70.96 (2.02)	65.43 (3.02)
Stratified sampling (50%)	61.35 (2.82)	63.14 (2.30)	62.04 (1.94)	57.99 (2.23)
Stratified sampling (20%)	56.20 (5.29)	57.12 (5.57)	56.53 (6.38)	52.61 (5.59)
Random sampling (50%)	70.19 (2.04)	70.35 (2.11)	69.99 (2.56)	65.14 (1.86)
Random sampling (20%)	67.79 (4.48)	67.41 (2.47)	67.10 (3.09)	63.64 (2.96)

4. Discussion

We observed high accuracy comparable to previous works. Grouping by individuals clearly reduced the scores, confirming identity confounding. When balancing age and gender distribution, we observed a further drop. While smaller sample size slightly decreased accuracy, random sampling outperformed stratified sampling. Although our results are limited to certain methods and a simple phonation task, we showed that it is crucial to consider comparable control groups when assessing performance.

5. Conclusion

We investigated smartphone-recorded phonation in PD detection using ML and the potential evaluation bias on unbalanced data. Many features show promising results on the mPower dataset, but are limited when balancing classes, gender and age. To prevent pitfalls in future analyses, we propose the use of stratified sampling and grouped CV.

References

[1] Mei J, Desrosiers C, Frasnelli J. Machine learning for the diagnosis of Parkinson's disease: a review of literature. Frontiers in aging neuroscience. 2021 May 6;13:633752.

[2] Tracy JM, Özkanca Y, Atkins DC, Ghomi RH. Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. Journal of Biomedical Informatics. 2020 Apr 1;104:103362.

[3] Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, Doerr M, Pratap A, Wilbanks J, Dorsey E, Friend SH. The mPower study, Parkinson disease mobile data collected using ResearchKit. Scientific data. 2016 Mar 3;3(1):1-9.

[4] Brenner A. Stratified Sampling, (2023). https://github.com/alex-bre/iterative_stratified_sampling

[5] Lenain R, Weston J, Shivkumar A, Fristed E. Surfboard: Audio feature extraction for modern machine learning, *ArXiv Prepr. ArXiv200508848*. (2020).

[6] Eyben F, et al. Opensmile: the munich versatile and fast open-source audio feature extractor. InProceedings of the 18th ACM international conference on Multimedia 2010 Oct 25 (pp. 1459-1462).