**Extreme Gradient Boosting for Parkinson’s Disease Diagnosis from Voice Recordings**

Ibrahim Karabayir, Samuel Goldman, Suguna Pappu, Oguz Akbilgic

**Abstract**

**Background**: Parkinson’s Disease (PD) is a clinically diagnosed neurodegenerative disorder that affects both motor and non-motor neural circuits. Speech deterioration (hypokinetic dysarthria) is a common symptom, which often presents early in the disease course. Machine learning can help movement disorders specialists improve their diagnostic accuracy using non-invasive and inexpensive voice recordings.

**Method**: We used “Parkinson Dataset with Replicated Acoustic Features Data Set” from the UCI-Machine Learning repository. The dataset included 45 features including sex and 44 speech test based acoustic features from 40 patients with Parkinson’s disease and 40 controls. We analyzed the data using various machine learning algorithms including tree-based ensemble approaches such as random forest and extreme gradient boosting. We also implemented a variable importance analysis to identify important variables classifying patients with PD.

**Results**: The cohort included total of 80 subjects; 40 patients with PD (55% men) and 40 controls (67.5% men). PD patients showed at least two of the three symptoms; resting tremor, bradykinesia, or rigidity. All patients were over 50 years old and the mean age for PD subjects and controls were 69.6 (SD 7.8) and 66.4 (SD 8.4), respectively. Our final model provided an AUC of 0.940 with 95% confidence interval 0.935-0.945 in 4-folds cross validation using only six acoustic features including Delta3 (Run2), Delta0 (Run 3), MFCC4 (Run 2), Delta10 (Run 2/Run 3), MFCC10 (Run 2) and Jitter_Rap (Run 1/Run 2).

**Conclusions**: Machine learning can accurately detect Parkinson’s disease using an inexpensive and non-invasive voice recording. Such technologies can be deployed into smartphones for screening of large patient populations for Parkinson’s disease.

**Keywords**: Parkinson’s Disease, Gradient Boosting, Machine Learning, Artificial Intelligence
Introduction

Parkinson’s disease (PD) is a neurodegenerative disorder of largely unknown cause [1]. After Alzheimer’s disease, it is the second most common neurodegenerative disease [2]. In 2010, there were approximately 680,000 people over 45 years old with PD in the US in 2010 and this number is expected to rise to 1,238,000 in 2030 [3]. Diagnosis of PD currently relies on clinical examination, however, by the time PD becomes clinically apparent, there is more than 50% loss of substantia nigra neurons and an 80% decline in striatal dopamine levels [4], [5].

Current gold standard in PD diagnosis is based on motor signs and symptoms (bradykinesia, resting tremor, rigidity, postural reflex impairment) and response to dopaminergic drugs [6], [7]. However, the accuracy of commonly used clinical diagnostic are around 80% [8] [9], implying a large population with PD is undiagnosed or may be misdiagnosed with other types of neurodegenerative disorders [10]. On the other hand, PD is well-recognized as a disorder that affects both motor and non-motor neural circuits [6], [11], [12]. Hence, discovery of novel motor or non-motor markers of PD or improving the accuracy of currently available diagnostic tools have potential to better diagnose patients with PD.

Noninvasive speech tests have been explored as a marker of disease [8], [13], since deterioration of speech is consistently observed in patients with PD [14]–[16]. In their work, Naranjo et al. [17], [18] showed that patients with PD could be detected with moderately high accuracy using acoustic features extracted from a speech test. In this study, we implemented machine learning methods, specifically Extreme Gradient Boosting [19], on Naranjo et al.’s [17] data to significantly improve PD detection accuracy from acoustic features extracted from voice recordings.

Methods

Data: We utilized “Parkinson Dataset with Replicated Acoustic Features Data Set” that was donated to University of California Irvine Machine Learning repository by Naranjo et al. [17] in April 2019. The dataset includes sex and age and 44 acoustic features extracted from voice recordings of 40 patients.
with PD and 40 controls. The voice recordings were repeat three times (three runs) applied as sustained phonation of the vowel /a/ for 5 seconds on one breath. Digital recordings were implemented at 441 KHz sampling rate and 16 bits/sample rate [17].

The 44 acoustic features extracted from voice recordings comprised five categories; pitch and amplitude local perturbation, noise, special envelope, and nonlinear measures. Four pitch local features (jitter relative, jitter absolute, jitter RAP (relative absolute perturbation), jitter PPQ (pitch perturbation quotient), and five amplitude perturbation measures (shimmer local, shimmer dB, APQ3 (3 point Amplitude Perturbation), APQ5 (5 pint Amplitude Perturbation) were extracted using a waveform matching algorithm. As measures of relative level of noise in speech [17], five different variants of harmonic-to-noise ratio (HNR) corresponding to different frequency bandwidths (HNR05 [0-500 Hz], HNR15 [0-1500 Hz], HNR25 [0-2500 Hz], HNR35 [0-3500 Hz], HNR38 [0-3800 Hz]) were extracted using the VoiceSauce Toolbox [20]. Glottal-to-Noise Excitation Ratio (GNE), quantifying the amount of voice excitation, was also calculated. Since PD is known to affect articulation [21], 13 Mel Frequency Cepstral Coefficients (MFCCs) associated with articular position and 13 Delta Coefficients as time dependent derivatives of MFCCs were also extracted. In addition, Recurrence Period Density Entropy (RPDE), Detrended Fluctuation Analysis (DFA), and Pitch Period Density Entropy (PPE) were also extracted as non-linear measures of voice recordings. Further details of the dataset can be found in Naranjo et al. [17].

Features: As described above, each acoustic feature was calculated three times for different runs of the speech test. Thus, in addition to testing the diagnostic accuracy of our analytic approach, we were also able to investigate intra-individual changes in response from different runs of the test. We considered acoustic features calculated for all three runs as individual predictors. Moreover, for a given acoustic feature, we created three artificial variables representing the change from one run to another (Figure 1). Therefore, our feature set included 264 acoustic features and sex for 80 subjects.
**Classification**: We implemented an extreme gradient boosting algorithm to distinguish between subjects with PD and controls. Gradient boosting is an ensemble machine learning consisting of several weak models (shallow decision trees rather than overfitting deep ones) and it can be used for both regression and classification problems [19]. Because it uses weak classifiers, it is more robust against overfitting compared to a random forest, a similar method that allows overfitting of individual tree predictors [19], [22], [23]. In our work, we mainly implement 4-fold cross validation to identify any overfitting by randomly splitting data into four distinct folds. We also repeat this process multiple times and present average results.

**Variable Importance Analysis, Feature Selection, and Re-Classification**: We first built the gradient boosting model using 265 features with four folds cross-validation and repeated this process 100 times. At each run, for each model built within 4-fold cross validation (4x100 models), we implemented a feature importance analysis that calculates the relative contribution of each feature to the corresponding model. A higher value of this metric for a specific feature implies it as a more important feature than another feature that has lower value of this metric [24]. By averaging the feature importance obtained from 400 individual models, we obtained a ranking of the 265 features. Next, we built new classification models with 4-fold cross-validation by incrementally adding the top 15 most important features selected from the previous step into the model with respect to their importance ranking. We repeated each of these steps 100 times to better estimate the effect of each feature on the model performance when they are introduced into model. We then identified the step where the model performance started diminishing or stopped increasing. Finally, using the features introduced up to that step.
specific step, we rebuilt gradient boosting models with 4-fold cross validation and report various performance metrics such as specificity, sensitivity, positive predictive value, accuracy, F1 score, and area under the receiver operators’ curve (AUC).

Comparisons: We compared our results to other commonly used classification models including logistic regression (LR), k-nearest neighborhood (KNN), support vector machines (SVM), and random forest (RF).

Results

Cohort: Our cohort included 40 subjects with PD (55% men) and 40 healthy controls (67.5% men). All subjects were over 50 years old and the mean age and standard deviation (SD) for PD subjects and controls was 69.6 (SD 7.8) and 66.4 (SD 8.4), respectively. PD diagnosis required at least two of resting tremor, bradykinesia, or rigidity [17].

Classification: We initially built the gradient boosting model with 4-fold cross-validation using the entire set of 265 predictors, 50 learners and a learning rate of 0.1. We repeated this process 100 times by randomly splitting the data into four folds and obtained an average F1 score of 0.862 with 95 confidence interval (CI) 0.860-0.864.

Variable Importance Analysis: As described in Methods section, using the total of 400 models, we obtained variable ranking based on their importance in classification. The top 15 variables are placed in the x-axis of Figure 2.

Feature Selection and Re-classification: To obtain a compact model, we repeated our 4-fold classification strategy 15 times by incrementally introducing a new variable into the model based on the order of importance. Figure 2 summarizes the F1-score with associated 95% confidence intervals for each step of this re-classification.
Figure 2 shows that F1 score increases up to 0.860 (95%CI 0.853-0.868) in the first six steps of the feature selection protocol and then slightly declines in following steps. In other words, after introducing the top six variables - Delta3 (Run2), Delta0 (Run3), MFCC4 (Run 2), Delta10 (Run 2/Run 3), MFCC10 (Run 2) and Jitter_Rap (Run 1/Run 2) - into the model, further additional variables did not improve the classification accuracy. We further implemented a grid search by changing the number of learners {20, 50, 60, 70, 80, 100} to identify whether the performance could be improved. However, there was no significant difference in AUC values of models with different parameter settings.

Comparisons: Since the F1 score using six variables was similar to accuracy obtained using all 265 variables, we decided to continue with this compact model of six variables. To compare the performance of our gradient boosting based model, we implemented the same 4-fold classification procedure using other commonly used classification methods (LR: Logistic Regression, SVM: Support Vector Machines, RF: Random Forest, KNN: K-nearest Neighbor) and provide the results in Table 1.
Table 1 Comparison of alternative machine learning methods

| Metrics   | AUC [95 % CI] | Gradient Boosting | LR | SVM | RF | KNN |
|-----------|---------------|-------------------|----|-----|----|-----|
|           |               |                   |    |     |    |     |
| F1        | 0.860         | 0.836             | 0.816 | 0.839 | 0.771 |
|           | [0.853- 0.868] | 0.829-0.844       | 0.808-0.824 | 0.831-0.847 | 0.762-0.780 |
| AUC       | 0.940         | 0.902             | 0.895 | 0.937 | 0.852 |
|           | [0.935-0.946] | 0.895-0.908       | 0.889-0.902 | 0.933-0.942 | 0.845-0.860 |
| Accuracy  | 0.864         | 0.834             | 0.820 | 0.845 | 0.776 |
|           | [0.857-0.871] | 0.8261-0.841      | 0.813-0.827 | 0.838-0.853 | 0.768-0.784 |
| Sensitivity | 0.852       | 0.855             | 0.807 | 0.817 | 0.771 |
|           | [0.840-0.863] | 0.845-0.866       | 0.796-0.819 | 0.806-0.829 | 0.758-0.785 |
| Specificity | 0.877      | 0.813             | 0.833 | 0.873 | 0.782 |
|           | [0.867-0.887] | 0.801-0.824       | 0.822-0.843 | 0.863-0.884 | 0.769-0.794 |
| PPV       | 0.883         | 0.830             | 0.837 | 0.875 | 0.791 |
|           | [0.874-0.892] | 0.821-0.840       | 0.828-0.846 | 0.866-0.885 | 0.781-0.801 |

Table 1 shows that gradient boosting outperforms all other machine learning models for all accuracy metrics considered.

The main reason for implementing 4-fold cross-validation in our work was to make our results comparable to the work of Naranjo et al. [17], [18], which is the original study utilizing this data.

However, we also implemented a 5-fold cross-validation using the six selected variables and obtained an F1 score of 0.869 (95%CI 0.861-0.877) and AUC of 0.944 (95%CI 0.939-0.949). We also repeated all results in Table 1 for 5-fold cross-validation and gradient boosting remained as the best performing method for all accuracy metrics (Supplementary Material).

Discussion
We were able to accurately classify persons with Parkinson’s disease by analysis of voice recordings using machine learning. Acoustic features extracted from speech test recordings offer a potential for computerized non-invasive diagnostic tools. The data we used in this study included 44 acoustic features generated separately for three runs of the same speech test task. In their original studies on the same data, Naranjo et al. [17], [18] proposed a statistical approach by treating the results of these runs as repeated measures and obtain an 4-fold classification accuracy, sensitivity, specificity, precision, and AUC of 0.779, 0.765, 0.792, 0.806, and 0.879, respectively. Our results show that better accuracy metrics (0.864, 0.852, 0.877, 0.940) are obtained via gradient boosting algorithms using only six acoustic features.

As reported above, our results showed that Delta3 (Run 2), Delta0 (Run 3), MFCC4 (Run 2), Delta10 (Run 2/Run 3), MFCC10 (Run 2) and Jitter_Rap (Run 1/Run 2) variables were the most important variables to detect Parkinson’s disease from speech. Figure 3 shows that these variables are indeed very different for controls and patients with Parkinson’s disease.

Figure 3 Distribution of six important variables across controls vs cases
It is worth noting that none of the acoustic variables obtained from the first run of the speech test were selected in the final model as a predictor. There were three variables from second run, one from the third run, one variable representing change between run 1 to run 2, and one variable representing change from run 2 to run 3. When we took a closer look at these two variables representing change, Jitter Rap Run 1 / 2 and Delta10 Run 2/3, in terms of the correlations of three runs (Figure 4), the responses to the same task in Run 1 and 3 were very highly correlated (similar) in patients with PD, while the same correlations were much smaller in controls. Interestingly, similar patterns were reported on typing style heterogeneity in patients with PD [25].

Figure 4 Correlations between three runs of two important variables
One of the important limitations in our study is that age data was not available, therefore, we could not include age as a predictor nor could implement subgroup analysis for age. However, it was reported that all subjects were older that 50 years old and the reported mean and standard deviation of the age did not present any statistically significant (two tail t-test, p>0.05) difference between the average age of PD and controls. Another important limitation to our study is that the mean PD duration was not known.

Conclusions

Gradient boosting can be used to identify patients with Parkinson’s disease from a simple non-invasive speech test. Further studies are required to investigate whether similar results can be obtained from records of normal conversation or phone calls, in order to screen larger patient populations at different stages of Parkinson’s disease. The value of this approach to identify early prodromal PD remains to be determined.

Declarations: This manuscript is based on a publicly available de-identified data from UCI-Machine Learning Data Repository.

All authors have approved this submission and no author have any conflict of interest to report.

There is no funding supported this project.

OA and SG conceptualized the study. IK and OA carried out the analysis and drafted the first version. All authors contributed provided critical feedback and contributed in writing the manuscript.

References

[1] C. Tanner and S. Goldman, “Epidemiology of Parkinson’s Disease,” *Neuroepidemiology*, vol. 14, no. 2, pp. 317–335, 1996.

[2] E. R. Dorsey *et al.*, “Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030,” *Neurology*, 2007.

[3] C. Marras *et al.*, “Prevalence of Parkinson’s disease across North America,” *npj Park. Dis.*, 2018.

[4] J. M. Fearnley and A. J. Lees, “Ageing and parkinson’s disease: Substantia nigra regional
[5] G. W. Ross, R. D. Abbott, H. Petrovitch, C. M. Tanner, and L. R. White, “Pre-motor features of Parkinson’s disease: the Honolulu-Asia Aging Study experience,” *Parkinsonism Relat. Disord.*, vol. 18, pp. S199–S202, 2012.

[6] P. Rizek, N. Kumar, and M. S. Jog, “An update on the diagnosis and treatment of Parkinson disease,” *CMAJ*. 2016.

[7] O. Suchowersky, S. Reich, J. Perlmutter, T. Zesiewicz, G. Gronseth, and W. J. Weiner, “Appendix A: Practice parameter: Diagnosis and prognosis of new onset Parkinson disease (an evidence-based review): Report of the Quality Standards Subcommittee of the American Academy Neurology,” *Contin. Lifelong Learn. Neurol.*, 2007.

[8] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, “Accurate Telemonitoring of Parkinson’s Disease Progression by Noninvasive Speech Tests,” *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, Apr. 2010.

[9] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, and G. Logroscino, “Accuracy of clinical diagnosis of Parkinson disease,” *Neurology*, 2016.

[10] A. Schrag, Y. Ben-Shlomo, and N. Quinn, “How valid is the clinical diagnosis of Parkinson’s disease in the community?,” *J. Neurol. Neurosurg. Psychiatry*, 2002.

[11] A. E. Lang and A. M. Lozano, “Parkinson’s disease: Second of two parts,” *New England Journal of Medicine*. 1998.

[12] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach,” *Biometrics*, vol. 44, no. 3, p. 837, 1988.

[13] B. Harel, M. Cannizzaro, and P. J. Snyder, “Variability in fundamental frequency during speech in prodromal and incipient Parkinson’s disease: A longitudinal case study,” *Brain Cogn.*, 2004.
[14] L. O. Ramig, C. Fox, and S. Sapir, “Speech treatment for Parkinson’s disease,” Expert Review of Neurotherapeutics. 2008.

[15] W. Maetzler, I. Liepelt, and D. Berg, “Progression of Parkinson’s disease in the clinical phase: potential markers,” Lancet Neurol., vol. 8, no. 12, pp. 1158–1171, Dec. 2009.

[16] S. Skodda, “Effect of Deep Brain Stimulation on Speech Performance in Parkinson’s Disease,” Parkinsons. Dis., vol. 2012, pp. 1–10, 2012.

[17] L. Naranjo, C. J. Pérez, Y. Campos-Roca, and J. Martín, “Addressing voice recording replications for Parkinson’s disease detection,” Expert Syst. Appl., vol. 46, pp. 286–292, Mar. 2016.

[18] L. Naranjo, C. J. Pérez, J. Martín, and Y. Campos-Roca, “A two-stage variable selection and classification approach for Parkinson’s disease detection by using voice recording replications,” Comput. Methods Programs Biomed., vol. 142, pp. 147–156, Apr. 2017.

[19] J. H. Friedman, “Stochastic gradient boosting,” Comput. Stat. Data Anal., vol. 38, no. 4, pp. 367–378, Feb. 2002.

[20] Y.-L. Shue, P. Keating, and C. Vicenik, “VOICESAUCE: A program for voice analysis,” J. Acoust. Soc. Am., vol. 126, no. 4, p. 2221, 2009.

[21] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, “Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson’s disease symptom severity,” J. R. Soc. Interface, 2011.

[22] C. Leistner, A. Saffari, P. M. Roth, and H. Bischof, “On robustness of on-line boosting - A competitive study,” in 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009, 2009.

[23] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Ann. Stat., 2001.

[24] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
L. Giancardo et al., “Computer keyboard interaction as an indicator of early Parkinson’s disease,” *Sci. Rep.*, vol. 6, no. 1, p. 34468, Dec. 2016.

### Supplementary Material

#### 4 Fold Results

| SCORES | KNN | XGBOOST | LR | SVM |
|--------|-----|---------|----|-----|
| F1     | 0.860382 | 0.852663 | 0.861101 | 0.836294 |
| AUC    | 0.940337 | 0.893531 | 0.84544 | 0.89717 |
| ACC    | 0.864375 | 0.857321 | 0.874349 | 0.83375 |
| SENS   | 0.8515 | 0.839399 | 0.863661 | 0.855 |
| SPEC   | 0.87725 | 0.867234 | 0.887266 | 0.8125 |
| PPV    | 0.882852 | 0.874209 | 0.891495 | 0.83023 |

#### 5 Fold Results

| SCORES | KNN | XGBOOST | LR | SVM |
|--------|-----|---------|----|-----|
| F1     | 0.869155 | 0.861254 | 0.872056 | 0.836294 |
| AUC    | 0.960407 | 0.910666 | 0.960328 | 0.90135 |
| ACC    | 0.87335 | 0.860283 | 0.880717 | 0.83735 |
| SENS   | 0.861 | 0.849315 | 0.872465 | 0.855 |
| SPEC   | 0.886 | 0.876091 | 0.895909 | 0.81235 |
| PPV    | 0.892768 | 0.884346 | 0.90119 | 0.83023 |

Research Square Preprint | CC BY 4.0