Journal of Biomedical Informatics 104 (2020) 103362

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease

John M. Tracy[a], Yasin Özkanca[b], David C. Atkins[c], Reza Hosseini Ghomi[d,*]

[a] Member of DigiPsych Lab, University of Washington, Seattle, WA, USA
[b] Electrical & Electronics Engineering, Ozyegin University, Istanbul, Turkey
[c] Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA
[d] Department of Neurology, University of Washington, Seattle, WA, USA

ARTICLE INFO

*Keywords:*
Parkinson's disease
Deep phenotype
Feature selection
Voice technology
Audio features
Voice biomarkers

ABSTRACT

Voice technology has grown tremendously in recent years and using voice as a biomarker has also been gaining evidence. We demonstrate the potential of voice in serving as a deep phenotype for Parkinson's Disease (PD), the second most common neurodegenerative disorder worldwide, by presenting methodology for voice signal processing for clinical analysis. Detection of PD symptoms typically requires an exam by a movement disorder specialist and can be hard to access and inconsistent in findings. A vocal digital biomarker could supplement the cumbersome existing manual exam by detecting and quantifying symptoms to guide treatment. Specifically, vocal biomarkers of PD are a potentially effective method of assessing symptoms and severity in daily life, which is the focus of the current research. We analyzed a database of PD patient and non-PD subjects containing voice recordings that were used to extract paralinguistic features, which served as inputs to machine learning models to predict PD severity. The results are presented here and the limitations are discussed given the nature of the recordings. We note that our methodology only advances biomarker research and is not cleared for clinical use. Specifically, we demonstrate that conventional machine learning models applied to voice signals can be used to differentiate participants with PD who exhibit little to no symptoms from healthy controls. This work highlights the potential of voice to be used for early detection of PD and indicates that voice may serve as a deep phenotype for PD, enabling precision medicine by improving the speed, accuracy, accessibility, and cost of PD management.

## 1. Introduction

Voice technology has enjoyed significant advancement in non-medical applications in recent years. Given this progress, our goal is to explore how voice can serve as a deep phenotype for various diseases. Here we present methodology for processing a raw audio signal comprised of a patient's voice in order to extract relevant acoustic features to detect disease, in this case, Parkinson's Disease (PD). PD is a progressive, neurodegenerative disease characterized by tremor, muscle rigidity, bradykinesia, and postural instability [1]. In addition to these hallmark motor symptoms, speech has also been shown to be affected. Aileen Ho et al. demonstrated that in a sample of 200 PD participants, 74% had some level of speech impairment [2]. Out of voice, fluency, and articulation, voice is the most frequently affected and to a greater extent in the earlier stages of speech deterioration in PD.

The specific pathophysiology underlying such changes, however, remains unclear. Recent work has demonstrated the effectiveness of employing Machine Learning algorithms to classify PD from non-PD using extracted voice features [3–5]. While these and similar studies report high classification accuracy, upwards of 98%-99%, they have several limitations. The more well-known problem is small cohort sizes which undermines the generalizability of the results to a larger and more diverse population. The lesser known problem is identity confounding where multiple voice samples are collected from each individual and these samples show up in both the training and testing data [6]. This leads to over-optimistic model performance because the model has learned to identify characteristics of specific individuals and is using that information to predict the label in the test set. In essence, this is a type of data leakage where information is unintentionally shared between the training and test sets, leading to inflated performance metrics. More fundamental is that studies simply classifying between PD and non-PD provide limited application for improving a patient's quality of life. While tools to increase diagnostic accuracy are needed, diagnosis is typically made only after the disease has

* Corresponding author at: Department of Neurology, University of Washington, 1959 NE Pacific Street, Seattle, WA 98195, USA.
*E-mail address:* rezahg@uw.edu (R. Hosseini Ghomi).

https://doi.org/10.1016/j.jbi.2019.103362
Received 1 March 2019; Received in revised form 10 December 2019; Accepted 17 December 2019
Available online 19 December 2019
1532-0464/ Published by Elsevier Inc.

**Table 1**
Participant and task level filtering.

<table>
<tr><th colspan="3"></th><th>PD</th><th>Control</th></tr>
<tr><td colspan="5">Demographics Survey (Participant Screening)</td></tr>
<tr><td>Variable Name</td><td>Survey Question</td><td>Possible Responses</td><td>Accepted Responses</td><td>Accepted Responses</td></tr>
<tr><td>Professional Diagnosis</td><td>Have you been diagnosed by a medical professional with Parkinson disease?</td><td>True | False | No Response</td><td>True</td><td>False</td></tr>
<tr><td>Diagnosis Year</td><td>In what year were you diagnosed with Parkinson disease?</td><td>Valid Year | Invalid Year | No Response</td><td>Valid Year</td><td>No Response</td></tr>
<tr><td>Surgery</td><td>Have you ever had any surgery for Parkinson disease, other than DBS?</td><td>True | False | No Response</td><td>False</td><td>False</td></tr>
<tr><td>Deep Brain Stimulation</td><td>Have you ever had Deep Brain Stimulation?</td><td>True | False | No Response</td><td>False</td><td>False</td></tr>
<tr><td>Are-Caretaker</td><td>Are you a spouse, partner or care-partner of someone who has Parkinson disease?</td><td>True | False | No Response</td><td>False</td><td>True | False | No Response</td></tr>
<tr><td colspan="5">MDS-UPDRS Part II (Participant Screening)</td></tr>
<tr><td colspan="3">Possible Total Score Range</td><td>Score Cut-Off</td><td>Score Cut-Off</td></tr>
<tr><td colspan="3">0 – 40 | Didn't Fill Out</td><td>< =9</td><td>0 | Didn't Fill Out</td></tr>
<tr><td colspan="5">Voice Task (Task Screening)</td></tr>
<tr><td>Variable Name</td><td>Description</td><td>Possible Responses</td><td>Accepted Responses</td><td>Accepted Responses</td></tr>
<tr><td>Medication Timepoint</td><td>Time at which a voice task was completed relative to taking PD medication</td><td>I don't take Parkinson's medications | Immediately before Parkinson's medication | Just after Parkinson's medication (at your best) | Another Time | No Response</td><td>I don't take Parkinson's medications | Immediately before Parkinson's medication</td><td>I don't take Parkinson's medications | No Response</td></tr>
</table>

**Table 2**
Demographics (n represents the number of participants).

| | Control (n = 2023) | PD (n = 246) |
|---|---|---|
| **Gender** | | |
| Male | 1711 | 147 |
| Female | 308 | 99 |
| Prefer not to answer | 2 | 0 |
| Did not answer | 2 | 0 |
| **Race** | | |
| Black or African | 50 | 1 |
| Caribbean | 6 | 0 |
| East Asian | 106 | 1 |
| Middle Eastern | 37 | 1 |
| Mixed | 36 | 2 |
| Native American | 7 | 1 |
| Other | 21 | 0 |
| Pacific Islander | 4 | 0 |
| South Asian | 72 | 2 |
| White or Caucasian | 1389 | 233 |
| Multiple | 92 | 1 |
| No response | 7 | 0 |
| **Age** | 31.9 (12.2) | 61.2 (10.0) |
| **Disease Duration (2015 – diagnosis-year)** | NA | 4.0 (4.0) |

progressed into more debilitating stages, i.e. when symptoms become noticeable. Some researchers have identified this limitation and have taken a different approach – one that attempts to classify the severity of disease using speech processing algorithms [7,8]. These highlight the potential for cost-efficient methods of disease monitoring, but it's unclear whether one component, namely speech, will be able to characterize the full severity of PD over time.

A pressing need still remains for methods capable of detecting the disease earlier in its course, which is where the application of machine learning algorithms have great potential. One example of work being done in this area is the Parkinson's Progression Markers Initiative (PPMI), an observational, international study aimed at characterizing imaging, clinical, genetic, and biospecimen PD progression markers [9]. By collecting information from various PD cohorts at earlier stages of the disease, the study provides a unique window for examining PD progression. Recent work analyzing the study's neuroimages using graph convolutional networks has already shown promising classification results [10,11]. Our work presents a far less invasive method for distinguishing early stage PD from controls, namely voice acquisition. Changes in speech have been implicated as a prodromal symptom in PD. Variability in fundamental frequency has been detected as early as 5 years before the onset of clinically diagnosable symptoms [12]. Early detection of PD cannot be understated as earlier treatment reduces immediate symptoms, improves long-term clinical outcomes, and has been suggested, though not conclusively demonstrated, to slow disease progression [13,14]. However, early detection is extremely difficult to achieve let alone study, because it requires knowing who will develop PD before the onset of symptoms. A fitting place to start then is by looking at those individuals who have been diagnosed but are in an early stage of the disease, i.e. showing few and non-disruptive symptoms. We build on our previous work [15] by going beyond basic disease classification (e.g. detecting the presence or absence of Parkinson's Disease), toward detection of disease early in its course. The hope is to build a voice model of Parkinson's robust enough to detect early symptoms, in advance of average clinical diagnosis to serve as a screening and disease management tool. Using non-invasive, easily-obtained, low-cost, patient-generated voice data provides a significant improvement to disease detection. This paper outlines voice analysis methodology applicable to any disease along with controlling for identity confounding, an issue common in many data sets. We will use a PD database to demonstrate application of this methodology and present performance of disease detection models. The primary contribution of this paper is its relevance for the early detection of PD. While the hypothesis that voice features can differentiate between PD and HC is known, this is not our hypothesis. Our approach extends this general classification to the more specific case between HC and those with PD who are exhibiting little or no symptoms. This distinction is significant because it supports the use of machine learning as a clinical application for early detection of PD, which has the potential to greatly improve quality of life and clinical outcomes. We aren't aware of any other studies that have attempted classifying between mild PD and HC.

## 2. Materials and methods

### 2.1. Data collection

mPower is an ongoing mobile research study that collects sensor and survey data from individuals diagnosed with PD and control subjects without the diagnosis [16]. The study is managed by Sage Bionetworks with data made available on their Synapse research portal. Once enrolled, participants were presented with a dashboard of study tasks. Study tasks were divided into surveys and four activities: walking, tapping, memory, and voice. Individuals with a PD diagnosis were prompted to complete each task 3 times a day at different time points relative to taking their medications (immediately before, after, and at another time), while controls were prompted to complete each task 3 times at any point during the day. The voice task consisted of an initial 5 s recording of ambient noise level followed by a 10 s recording of the participant vocalizing the single phoneme "aaah" into the microphone. Each audio file represented 1 voice task. iPhones (4th generation or newer) and iPods (5th generation or newer) were the only devices used by participants in the collection of voice data. Notably, no linguistic data was collected, and voice data was limited to only acoustic features.

Data was accessed using the Synapse Python Client. The voice data set was collected from March to September of 2015 and consists of 65,022 unique tasks with 5,826 unique individuals. Python was used for data analysis and modeling along with the following imported libraries: SKlearn, NumPy, Pandas, Matplotlib.

We screened participants and filtered individual voice tasks using a combination of survey and PD severity rating criteria. Our final dataset consisted of 2,289 individuals (2023 Controls, 246 PD) and 15,227 voice tasks (9994 Control tasks, 5233 PD tasks). See Table 1 for screening details and Table 2 for a breakdown of demographic information between the two groups.

The demographics survey was used to identify PD participants and controls and to screen for factors that might confound the distinction between the two groups. Participants with a PD diagnosis were identified as those who answered 'True' to having a professional diagnosis of PD and by providing a diagnosis-year. Healthy controls were identified as those who answered 'False' to having a professional PD diagnosis and by not providing a date for diagnosis-year or onset-year. Participants that had previously undergone surgery related to PD or Deep Brain Stimulation were excluded. Participants that answered true to a professional PD diagnosis but also answered true to 'are-caretaker' were additionally excluded due to confusion about whether the participant had the diagnosis or was caring for someone with the diagnosis.

To identify the subgroup of PD participants at a mild stage of disease, a modified form of The Movement Disorder Society's Unified Parkinson's Disease Rating Scale (UPDRS) was used. The UPDRS is the most widely used clinical rating scale for PD [17]. It is a combined rater and self-administered questionnaire consisting of 4 parts. Several questions from part I and all questions from Part II do not require a trained rater and can be completed directly by the participant [17]. In the mPower study a shortened version was provided that included 6 questions from Part I and 10 questions from Part II in order to increase likelihood of completion. Notably, changes to the UPDRS create a potential change in validity of the tool, however, we decided to continue analysis with the data available. Part I pertains to non-motor experiences of daily living (e.g. emotional status) while Part II pertains to motor experiences of daily living (e.g. difficulty getting dressed). In the full UPDRS rating, Part I and Part II each consist of 13 questions. Since more questions from Part II were provided and this project is trying to identify subtle discrepancies in voice characteristics stemming from aberrant functioning in the muscles used to generate speech, Part II was chosen as our severity measure.

Martinez-Martin et al. propose a method for assessing PD patients as mild, moderate, and severe based on the UPDRS [18]. Using triangulation, they identify cutoff scores for each of three severity levels on each of the 4 sub-sections. They found the cutoff scores for Part II to be 12/13 for mild/moderate and 28/29 for moderate/severe. A score of 13 averages to 1 point per question and since we only had access to 10 of 13 questions of the Part II subsection, we adjusted our threshold cutoff for the 10 questions to be 9/10 for mild/moderate stage, subtracting 3 points.

Since participants were able to fill out the UPDRS multiple times, we used their highest score when considering their severity level. The mild PD subgroup consisted of those participants whose highest score on the UPDRS was less than or equal to 9. Comparing the mean disease duration between those considered mild vs moderate/severe using our criteria yielded averages of 4.00 years (mild) and 5.29 years(moderate/severe). We validated the mean difference using a *t*-test and obtained a p-value of 0.005 which supports our approach to severity classification. Controls were also able to fill out the UPDRS. Only controls that either did not fill out the UPDRS or scored a 0 on the part II subsection were included in the analysis. Table 3 shows the mean and standard deviation of each question and total of part II for both groups.

To avoid any confounding between the groups, we also controlled for medication use. Parkinson’s medication has been shown to reduce symptom affects and sometimes completely relieve symptoms temporarily [19]. For PD participants, only those that took their medication after the task were included in analysis. Only controls not taking any medications were included. Some PD participants had not yet begun medication which corresponded with a more recent diagnosis or suggested a less severe progression of symptoms after diagnosis. This also corroborates our methods for identifying those participants with a milder form of the disease.

The voice activity data contained a .m4a audio file for the recording, a .m4a file for countdown before the recording, and a medication time point variable indicating at what time medication was taken in relation to completing the task. Additional information included phone descriptors such as type and app version.

### 2.2. Feature extraction and preprocessing

Prior to feature extraction, background noise was removed from the recordings using activlev, a voice activation detection algorithm from MATLAB's Voicebox toolkit [20]. 2268 features corresponding to the Audio/Visual Emotion and Depression Recognition Challenge (AVEC) 2013 [21], and 62 features corresponding to the Geneva Minimalistic Acoustic Parameter Set (GeMaps) [22] were extracted using the OpenSmile toolkit [23].

The AVEC features consisted of 32 energy and spectral low-level descriptors (LLDs) × 42 functionals, 6 voice-related LLDs × 32 functionals, 32 delta coefficients of the voicing related LLDs × 19 functionals, and 10 voiced/unvoiced durational features. Applied functionals included 23 statistical, 4 regression, 9 local minima/maxima, and 6 other related functionals [21]. These constitute a standard set of commonly used features in audio signal analysis and emotion recognition.

The creation of GeMAPS was an effort to derive a standard, minimalistic acoustic feature set in response to the recent proliferation of acoustic parameters and varied extraction processes conducted by researchers. While some overlap between the sets exists, GeMAPS and AVEC were combined to produce the final feature set prior to modeling.

**Table 3**
MDS-UPDRS Part II Scores for mild PD severity participants (n represents the number of participants. The controls are fewer because not all completed the UPDRS.) *A single PD participant’s scores were averaged if they completed the survey multiple times prior to averaging for each subtask and total).

| | Control (n = 457) | PD (n = 246)* |
|---|---|---|
| **MDS-UPDRS2.1** | 0.00 (0.00) | 0.56 (0.74) |
| **MDS-UPDRS2.4** | 0.00 (0.00) | 0.30 (0.48) |
| **MDS-UPDRS2.5** | 0.00 (0.00) | 0.37 (0.51) |
| **MDS-UPDRS2.6** | 0.00 (0.00) | 0.17 (0.37) |
| **MDS-UPDRS2.7** | 0.00 (0.00) | 0.74 (0.88) |
| **MDS-UPDRS2.8** | 0.00 (0.00) | 0.49 (0.71) |
| **MDS-UPDRS2.9** | 0.00 (0.00) | 0.43 (0.50) |
| **MDS-UPDRS2.10** | 0.00 (0.00) | 1.00 (0.70) |
| **MDS-UPDRS2.12** | 0.00 (0.00) | 0.45 (0.54) |
| **MDS-UPDRS2.13** | 0.00 (0.00) | 0.07 (0.29) |
| **Part II Sum** | 0.00 (0.00) | 4.59 (2.57) |

### 2.3. Machine learning methods

Three popular models for binary classification were chosen: L2-regularized logistic regression, random forest, and gradient boosted decision trees. We chose classic machine learning models as opposed to the newer, deep learning models because of our emphasis on clinical application and the limited training data available. Because the mechanisms of traditional machine learning models are well understood, they currently provide more clinical utility since they preserve interpretability. Additionally, deep learning models require a large amount of training data to outperform conventional models, and there’s competing evidence over whether deep learning models provide a performance advantage over conventional models on small datasets. For example, Pishgar et al. found that a conventional SVM model performed better than an LSTM model across sensitivity and specificity metrics on a small dataset comprised of voice features from the 2018 FEMH Voice Disorder challenge [24]. Fang et al. found only slight performance gains from using a DNN over a SVM model on dataset of 406 voice samples from the MEEI voice disorder database [25]. Therefore, we chose to compare three conventional models on our dataset. Controls and PD were given binary labels of 0 and 1. Prior to modeling, feature values were scaled using standardization. Our implementation of logistic regression used the newton-sg solver, which converges faster after features are standardized. Since tree-based methods are independent of feature scale, this was not expected to affect the other algorithms. Models without scaling were additionally run; the logistic regression model experienced substantial performance loss while the other two models showed no change in performance.

The dataset was skewed more heavily towards controls (n = 2023) than PD (n = 246), so we selected recall, precision, and f1-score as our evaluation metrics to compare each model performance. ROC curves and AUC scores are also shown to compare overall model performance across different decision thresholds.

Cross-validation was used for all prediction metrics, and two different methods for train-test splitting were used to highlight the significance of identity confounding when using data that contains repeated tasks for the same participant. The first method ignored the individual level (e.g. the same participant repeated the task) and instead split at the task level, meaning that repeated measures from the same individual could potentially appear in both training and test data. Data was first split into training and testing sets using a 70/30 train-test split, and subsequently, hyper-parameters were grid-searched with 5-Fold Cross Validation. The second method accounted for individual level by splitting on participants’ unique id for the initial 70/30 train-test split. This ensured that the same participant would not show up in both the training and the testing sets. Likewise, this method was used

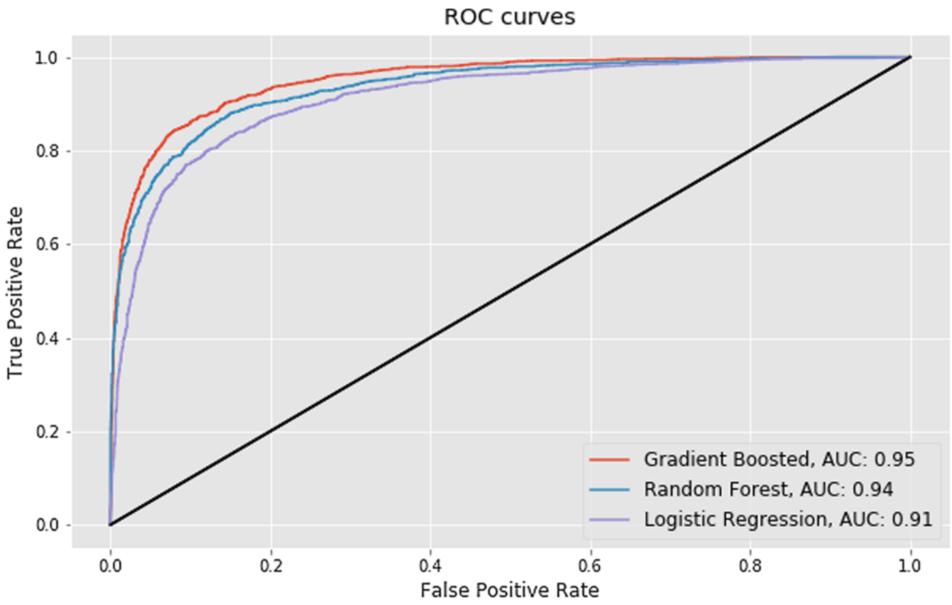

**Fig. 1.** ROC curves and AUC scores when splitting the data at task level for mild PD vs control.

during hyper-parameter tuning to ensure no participant crossover in the training and validation sets on each of the 5 folds. For both methods, the best model was chosen based on the highest mean accuracy across folds. It was then refit to the training data and evaluated on the testing set.

## 3. Results and discussion

In Fig. 1, we see that all models produced high AUC scores with the gradient boosted model performing best with a score of 0.95. Similarly, the gradient boosted model had the highest recall and F1 score as shown in Table 4. All models showed better precision than recall, which is somewhat expected given the majority of the data comes from the control group. Overall, results from the first method are promising as they show high rates of precision (low false-positives) and high rates of recall (low false-negatives) for the PD group.

A concern that arose during exploratory analysis was identity confounding, which results from having the same individuals in both the training and the test set due to replicated tasks. Identity confounding occurs when a model captures information that is unique to the individual as opposed to the target variable, which can lead to overestimation of model performance. Future work to help automate detection of PD signs and symptoms by using a robust model built on a training data set of many patients with PD but without included any data of the patient in question in order to reduce identity confounding. Separately, another model could be used to track an individual patient's symptoms over time to help manage the disease and treatment where the patient's data is used within the model for training and prediction of changes in symptoms.

**Table 4**
Model performance for mild PD vs Control (parameters selected by grid search and evaluation metrics with data split by task).

| Algorithm | Best Parameters | Recall | Precision | F1 |
|---|---|---|---|---|
| Logistic Regression (L2-Regularized) | C = 0.00599<br>Solver = ”Newton-CG”<br>Tolerance = 1e-4 | 0.759 | 0.811 | 0.784 |
| Random Forest | N_Estimators = 1000<br>Max_Features = ”Auto” | 0.693 | 0.902 | 0.783 |
| Gradient Boosted Trees | N_Estimators = 1000<br>Learning_Rate = 0.1 | 0.797 | 0.901 | 0.836 |

Fig. 2 shows that most of the data comes from participants that completed the task multiple times, being more heavily pronounced in the PG group. Participants were grouped based on how many times they completed the task. Adding up the number of tasks for each participant within a group gives you the total number of tasks, displayed on the y-axis (e.g. for participants that completed the task once, the total number of tasks will equal the number of participants, for participants that completed the task twice, it will be double). This illustrates how much the data is dominated by individuals who repeated the task multiple times, which makes potential for identity confounding more likely.

Although there is an a priori concern of identity confounding when a given patient's data can be in both the test and train partitions, Neto et al. [6] have proposed a statistical test for evaluating the potential confounding effect of identity in repeated measures data. Disease labels are shuffled subject-wise for a number, p, equal to the number of possible permutations and model performance metrics are accumulated for each permutation as shown in Fig. 3. This breaks the association between features and labels but preserves feature and subject association. Neto et al. suggest using the maximum number of permutations but this quickly becomes computationally expensive as the number of groups increases. To balance computational cost with sufficient sample size, we chose 5000 permutations. The Random Forest model was chosen for its comparable performance to the gradient boosted model but superiority in training/evaluation speed. AUC scores were selected as the metric of model performance. If there was no identity confounding, then we should see AUC scores close to 0.5, indicating that the model is randomly guessing. If identity confounding was present, then we should see AUC scores closer to 1.0. Fig. 4 shows the distribution of AUC scores for 500 permutations, where the scores range from 0.762 to 0.828 with a mean of 0.799, suggesting identity confounding is present.

Each model was trained with the best set of parameters found when fitting from the first method. Scores closer to 0.50 indicate that the model is randomly guessing, whereas scores closer to 1 indicate identity confounding.

Given the evidence of identity confounding, we re-ran the prediction models using the second train-test splitting method described in the methods section, which guarantees that an individual's data is either in the test or train partition but not both. Results are shown in Fig. 5 and Table 5 below.

With identity confounding removed, model performance drops as expected. Fig. 5 shows AUC scores remain much higher than random chance with the highest AUC score of 0.88 achieved by the gradient

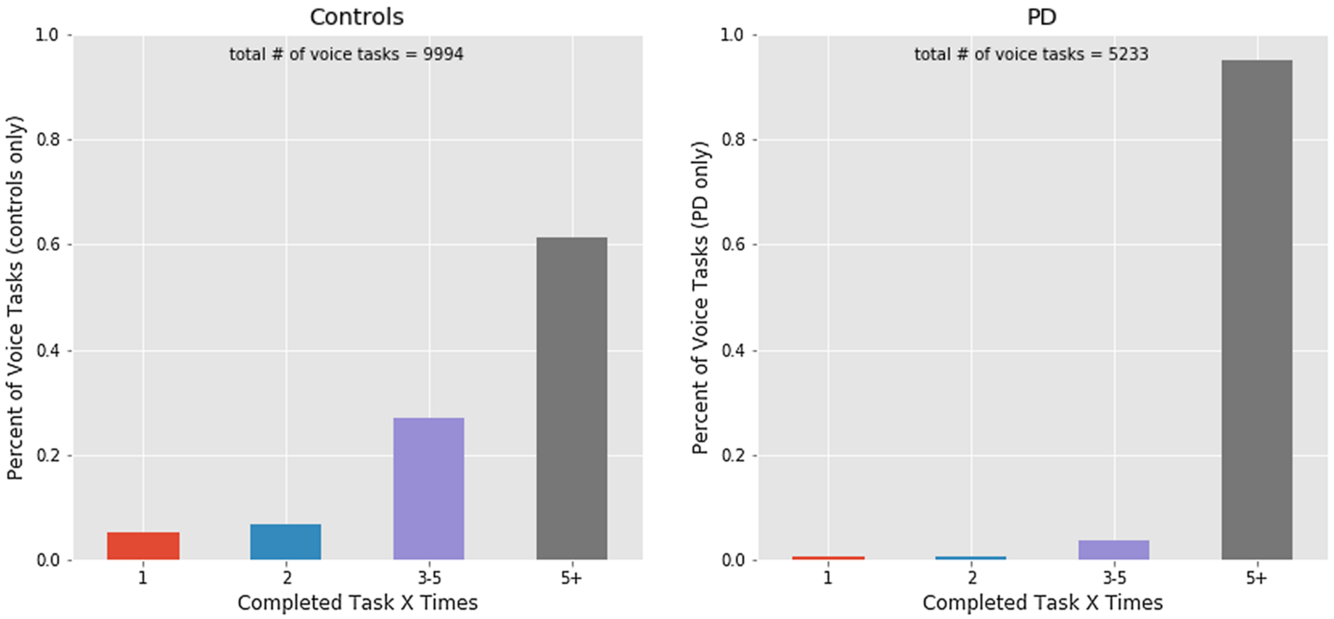

**Fig. 2.** Percentage of voice tasks for a given number of repetitions for Controls and Mild PD. Participants were grouped by how many times they completed the task, and percentages were calculated separately for each group.

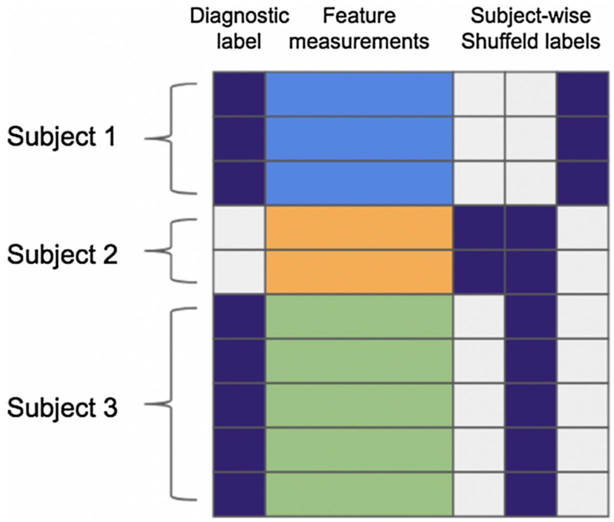

**Fig. 3.** Method for detecting identity confounding. Labels are shuffled subject-wise for an x number of permutations.

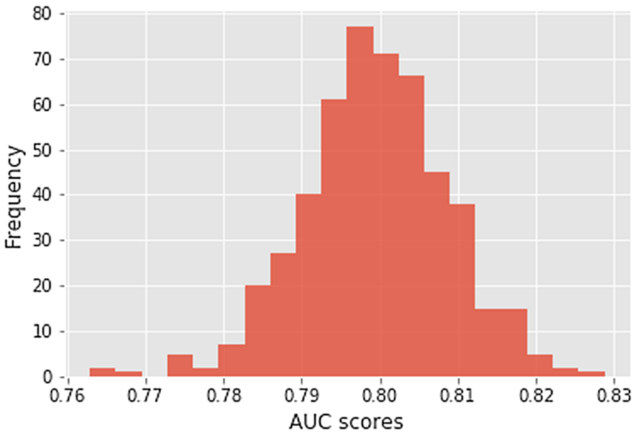

**Fig. 4.** Frequency histogram showing the distribution of AUC scores from 500 random forest models.

boosted model. Like the results from the first method, precision performance is better than recall, with most marked drop seen in the random forest model. Still, recall and precision remain modestly high for the PD group with the best recall (0.646), f1-score (0.688), and the best precision (0.849) achieved by the gradient boosted model. Notably, around 20% of true-positive cases are still identified with low false-positive rates (0.00–0.05) for the logistic regression and gradient boosted model. In light of the absence of any robust early detection techniques to date, these results show that voice features coupled with machine learning approaches can be used as an early detection indicator for Parkinson's.

Fig. 6 shows all 2330 features from the gradient boosted model ranked by importance, which measures a feature's contribution to predicting the target variable. A feature's importance is calculated by taking the total reduction in variance across all the internal nodes for which that feature was chosen as the partitioning variable (weighted by node probability), averaging across all trees that the feature is present in, and normalizing with respect to the other features [26]. Fig. 6 shows a steep drop in importance for the top 100 features followed by a steady decline, indicating that a small proportion of the predictors play a larger role in predicting the target variable. Table 6 lists the top 10 features that were derived from the gradient boosted model. The description includes the low-level descriptor (LLD) and the applied functional. A moving average filter with a window length of 3 was used for all low-level descriptors. Additionally, the source of the feature is provided (AVEC or GeMAPS). While there is no absolute mapping of acoustic features to anatomy and physiology [27], many of these features have been linked to the signs and symptoms of PD. Specifically, 3 of the top 10 features directly measure derivations of fundamental frequency and another 2 features are derived from the second formant. Fundamental frequency and formants are derived directly from the anatomy of the vocal tract and have been shown to be affected by PD [12,28–30]. All of these features capture the ability to phonate and articulate with precision while speaking and in PD. Articulation of vowels has been shown to be impaired even when instructed to speak as clearly as possible, which has been directly linked to the pathology affecting dopaminergic neurotransmission [30]. Additional testing was performed to validate differences in the features between groups. Since users were able to complete the voice task multiple times, their mean value was taken for each feature. A Bonferroni correction was applied

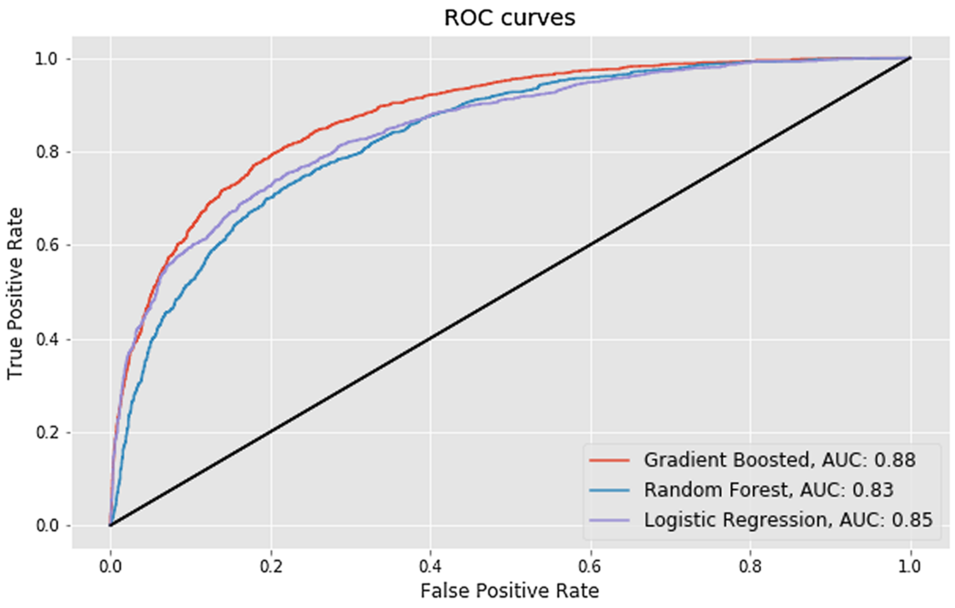

**Fig. 5.** ROC curves when splitting the data at the individual level for mild PD vs control.

**Table 5**
Model performance for mild PD vs control (best parameters selected by grid search, and evaluation metrics when splitting the data at the individual level).

| Algorithm | Best Parameters | Recall | Precision | F1 |
|---|---|---|---|---|
| Logistic Regression (L2-Regularized) | C = 0.00077<br>Solver = ”Newton-CG”<br>Tolerance = 1e-4 | 0.608 | 0.710 | 0.655 |
| Random Forest | N_Estimators = 400<br>Max_Features = ”Auto” | 0.391 | 0.777 | 0.520 |
| Gradient Boosted Trees | N_Estimators = 1000<br>Learning_Rate = 0.1 | 0.646 | 0.849 | 0.688 |

to account for multiple testing. All but 2 features were statistically significant at $p < 0.01$, and 6 of the features were statistically significant at $p < 0.005$ (see Table 7).

The empirical results reported herein should be considered in light of study limitations. The biggest limitations are that the data collection for the demographics survey was done through self-reporting, and there is no validation on external datasets. One longitudinal study found, using neurological examination and post-mortem evaluation, that only 35–40% of individuals reporting a prevalence of PD had typical PD [31]. This draws into question the validity of responses, most notably the professional-diagnosis itself. Participants may have different understandings of what a professional diagnosis constitutes, and we know diagnosis accuracy does differ depending on the provider’s training. Measures were taken by study originators to encourage accuracy of reporting such as making the study completely voluntary with no monetary incentive or potential for personal benefit. The authors additionally filtered for factors that suggest uncertainty of a true diagnosis (see Table 1) to increase the likelihood of accurate responses. While these measures were taken, they cannot compensate for the lack of verification by a trained professional, and the authors strongly suggest readers take this into consideration when interpreting results.

Additionally, our study relies on the ability to accurately distinguish PD participants with mild symptoms from those with moderate/severe

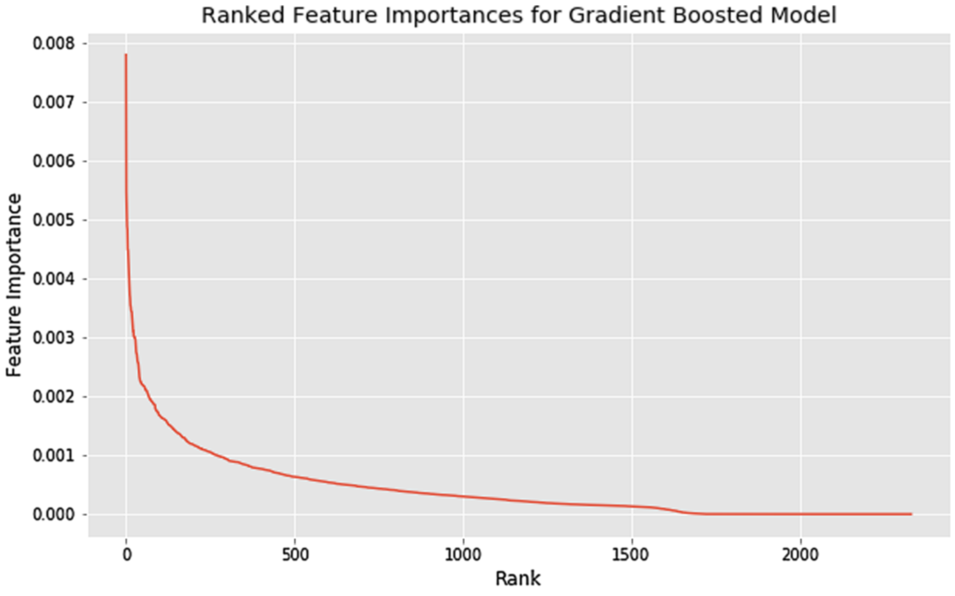

**Fig. 6.** Gradient boosted model feature importances ranked in descending order.

**Table 6**
Top 10 feature importances from the Gradient Boosted Model (a moving average filter with a window length = 3 was used for smoothing all low-level descriptors).

| Rank | Feature | Description | Importance | Source |
|---|---|---|---|---|
| 1 | F0final_sma_linregc1 | LLD: fundamental frequency<br>Functional: slope of a linear approximation of the contour | 0.007802 | AVEC |
| 2 | slopeV500-1500_sma3nz_amean numeric | LLD: spectral slope between the two given bands<br>Functional: arithmetic mean | 0.005474 | GeMAPS |
| 3 | F0final_sma_de_quartile1 | LLD: 1st order differential of the fundamental frequency<br>Functional: 1st quartile | 0.005182 | AVEC |
| 4 | F2frequency_sma3nz_stddevNorm numeric | LLD: 2nd formant<br>Functional: Normalized standard deviation | 0.004893 | GeMAPS |
| 5 | pcm_fftMag_mfcc_sma [7]_percentile99.0 | LLD: Mel-Frequency Cepstral Coefficient 7 (using pulse-code modulation and calculating the magnitude of the Fast Fourier Transform)<br>Functional: 99th percentile | 0.004833 | AVEC |
| 6 | voicingFinalUnclipped_sma_percentile99.0 | LLD: voicing probability of the fundamental frequency<br>Functional: 99th percentile | 0.004491 | AVEC |
| 7 | pcm_fftMag_fband250-650_sma_linregc1 | LLD: Frequency Band 250–650<br>Functional: slope of a linear approximation of the contour | 0.004486 | AVEC |
| 8 | F0final_sma_de_iqr2-3 | LLD: 1st order differential of the fundamental frequency<br>Functional: inter-quartile range (quartile3 – quartile 2) | 0.004324 | AVEC |
| 9 | pcm_fftMag_spectralFlatness_sma_percentile1 | LLD: Spectral Flatness (using pulse-code modulation and calculating the magnitude of the Fast Fourier Transform)<br>Functional: 1st percentile | 0.004240 | AVEC |
| 10 | F2frequency_sma3nz_amean numeric | LLD: 2nd formant<br>Functional: arithmetic mean | 0.003990 | GeMAPS |

**Table 7**
Differences in top 10 features between HC and PD groups. *** significant at $p < 0.005$; ** significant at $p < 0.01$, * significant at $p < 0.05$. A Bonferroni correction of n = 10 was applied to account for multiple tests.

| Rank | Feature | HC (n = 2023) | PD (n = 246) | T statistic | P value |
|---|---|---|---|---|---|
| 1 | F0final_sma_linregc1 | M = 7.30e−03<br>SD = 0.145 | M = −2.34e−02<br>SD = 0.117 | 3.786 | 1.80e−04*** |
| 2 | slopeV500-1500_sma3nz_amean numeric | M = −2.46e−02<br>SD = 1.27e−02 | M = −2.38e−02<br>SD = 1.55e−02 | −0.745 | 0.457 |
| 3 | F0final_sma_de_quartile1 | M = −2.57e−04<br>SD = −5.83e−04 | M = 5.03e−04<br>SD = 1.47e−03 | 2.601 | 9.85e−03** |
| 4 | F2frequency_sma3nz_stddevNorm numeric | M = 8.72e−02<br>SD = 5.92e−02 | M = 0.110<br>SD = 5.80e−02 | −5.728 | 2.40e−08*** |
| 5 | pcm_fftMag_mfcc_sma[7]_percentile99.0 | M = 3.363<br>SD = 15.01 | M = 1.488<br>SD = 14.86 | 1.867 | 6.29e−02 |
| 6 | voicingFinalUnclipped_sma_percentile99.0 | M = 0.867<br>SD = 2.99e−02 | M = 0.874<br>SD = 2.82e−02 | −4.068 | 6.00e−05*** |
| 7 | pcm_fftMag_fband250-650_sma_linregc1 | M = −0.132<br>SD = 0.130 | M = −0.166<br>SD = 0.104 | 4.747 | 3.04e−06*** |
| 8 | F0final_sma_de_iqr2-3 | M = 2.45e−04<br>SD = 6.20e−04 | M = 4.30e−04<br>SD = 8.62e−04 | −3.270 | 1.21e−03*** |
| 9 | pcm_fftMag_spectralFlatness_sma_percentile1 | M = 4.15e−04<br>SD = 1.30e−03 | M = 6.03e−04<br>SD = 1.01e−03 | −2.659 | 8.21e−03** |
| 10 | F2frequency_sma3nz_amean numeric | M = 1330.64<br>SD = 157.19 | M = 1370.67<br>SD = 151.34 | −3.900 | 1.18e−04*** |

symptoms. The most commonly used instrument for assessing PD severity is the MDS-UPDRS rating scale which consists of 4 subscales ranging from 6 to 18 questions each. Only a subset of questions from the first 2 subscales were included in the study, which constrained our ability to adequately measure severity. Modifying of the scale itself also calls into question the validity of using the subset of questions to determine severity.

The data also demonstrates several limitations of this study design including significant disparities in age, gender, race, and similar movement disorders. In clinical practice, early diagnosis of PD is limited by differential diagnoses similar to PD. Therefore, further research that includes training examples from populations with different movement disorders is necessary. For example, Varghese et al. are conducting a 2 year observational study to identify new phenotypical biomarkers of PD and Essential Tremor (ET) given the high rates of misdiagnosis between the two disorders [32]. More studies incorporating similar movement disorders are paramount to the utility of machine learning as a clinical application.

## 4. Conclusion

Here, we demonstrated the use of voice as a biomarker in the early detection of PD as a test case for general voice analysis methodology. Additionally, this project corroborates previous findings regarding the need to check for identity confounding in studies that use repeated tasks. We were able to build a feature extraction and analysis architecture generalizable to any disease to build voice biomarkers while addressing a source of bias in digital biomarker data collection. Identity confounding is an important source of bias and leads to erroneous accuracies if not adjusted for. This demonstrates that although machine learning is a powerful tool in health care, special attention to the limitations of our data sets is needed in applying these methods. After accounting for identity confounding, performance across models dropped, but AUC scores remain high with the best score of 0.85. With the Logistic Regression model, approximately 20% of true-positive cases were identified low false-positive rates close to 0. These results highlight using voice features and machine learning as a promising avenue for early detection methods as well as the significance of accounting for identity confounding in open-source projects that contain repeatable tasks.

Although reasonable methods were used to identify severity levels in PD, no clinical validation of severity was provided in the study. Future studies that provide clinically validated severity levels are needed to corroborate the methods and findings of this study. Additionally, there was a large skew in the number of times a task was repeated between participants. New methods aimed at balancing participation for repeatable tasks, such as limiting number of submissions per individual, will help to curb the effects of identity confounding in future studies.

Ultimately, our initial findings hold promise to develop future applications which can use a simple, brief task to screen for PD. A mobile device, or any device with a microphone, can be used to collect voice data and provide a screening result with a recommendation to see their physician – both for earlier disease detection and for disease management to help guide treatment.

## 5. Statement of ethics

Ethical oversight of the study was provided by the Western Institutional Review Board (WIRB #20141369). Subjects who participated were required to complete an online consent process.

## CRediT authorship contribution statement

**John M. Tracy:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration. **Yasin Özkanca:** Investigation, Resources, Data curation. **David C.Atkins:** Validation, Writing - review & editing. **Reza Hosseini Ghomi:** Conceptualization, Validation, Resources, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

## Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Dr. Hosseini Ghomi is a stockholder of NeuroLex Laboratories, a voice technology company.

## Acknowledgement

The authors thank Dong Si of the Division of Computing and Software Systems, University of Washington for editorial comments. Data was contributed by users of the Parkinson mPower mobile application as part of the mPower study developed by Sage Bionetworks and described in Synapse [30] doi: https://doi.org/10.7303/syn4993293.

## Disclosure statement

Dr. Hosseini Ghomi is a stockholder of NeuroLex Laboratories.

## Funding

Dr. Hosseini Ghomi's work was supported by NIH R25 MH104159 during the completion of this work. This funding provided protected time in Dr. Hosseini Ghomi's schedule. Dr. Hosseini Ghomi's work in preparing this manuscript was supported during his VA Advanced Fellowship Program in Parkinson's Disease which provided protected research time.

## References

[1] R.J. Uitti, Y. Baba, Z.K. Wszolek, D.J. Putzke, Defining the Parkinson's disease phenotype: initial symptoms and baseline characteristics in a clinical cohort, Parkinsonism Relat. Disord. 11 (2005) 139–145.

[2] A.K. Ho, R. Iansek, C. Marigliani, J.L. Bradshaw, S. Gates, Speech impairment in a large sample of patients with Parkinson's disease, Behav. Neurol. 11 (1999) 131–137.

[3] R.A. Shirvan, E. Tahami, Voice analysis for detecting Parkinson's disease using genetic algorithm and KNN classification method, 2011 18th Iranian Conference of Biomedical Engineering (ICBME), 2011, pp. 278–283, , https://doi.org/10.1109/ICBME.2011.6168572.

[4] A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, L.O. Ramig, Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease, IEEE Trans. Biomed. Eng. 59 (2012) 1264–1271.

[5] A.U. Haq, et al., Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson's disease using voice recordings, IEEE Access 7 (2019) 37718–37734.

[6] E.C. Neto, et al., Detecting confounding due to subject identification in clinical machine learning diagnostic applications: a permutation test approach, ArXiv171203120 Stat, 2017.

[7] A. Bayestehtashk, M. Asgari, I. Shafran, J. McNames, Fully automated assessment of the severity of Parkinson's disease from speech, Comput. Speech Lang. 29 (2015) 172–185.

[8] A. Zhan, et al., Using smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score, JAMA Neurol. 75 (2018) 876–880.

[9] The Parkinson Progression Marker Initiative (PPMI) - ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0301008211001651.

[10] Multi-View Graph Convolutional Network and Its Applications on Neuroimage Analysis for Parkinson's Disease. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371363/.

[11] X.S. Zhang, J. Chou, F. Wang, Integrative Analysis of Patient Health Records and Neuroimages via Memory-based Graph Convolutional Network, ArXiv180906018 Cs Stat, 2019.

[12] L.K. Bowen, G.L. Hands, S. Pradhan, C.E. Stepp, Effects of Parkinson's disease on fundamental frequency variability in running speech, J. Med. Speech-Lang. Pathol. 21 (2013) 235–244.

[13] R.A. Hauser, et al., Long-term outcome of early versus delayed rasagiline treatment in early Parkinson's disease, Mov. Disord. 24 (2009) 564–573.

[14] D.L. Murman, Early treatment of Parkinson's disease: opportunities for managed care, Am. J. Manag. Care 18 (2012) S183–S188.

[15] T.J. Wroge, et al., Parkinson's disease diagnosis using machine learning and voice, The 2018 IEEE Signal Processing in Medicine and Biology Symposium (IEEE), (2018).

[16] B.M. Bot, et al., The mPower study, Parkinson disease mobile data collected using ResearchKit, Sci. Data 3 (2016).

[17] C.G. Goetz, et al., Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results, Mov. Disord. 23 (2008) 2129–2170.

[18] P. Martínez-Martín, et al., Parkinson's disease severity levels and MDS-Unified Parkinson's Disease Rating Scale, Parkinsonism Relat. Disord. 21 (2015) 50–54.

[19] G.M. Schulz, M.K. Grant, Effects of speech therapy and pharmacologic and surgical treatments on voice and speech in parkinson's disease: a review of the literature, J. Commun. Disord. 33 (2000) 59–88.

[20] M. Brookes, Voicebox: Speech Processing Toolbox for Matlab, 1997.

[21] M. Valstar, et al., AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. in 3–10 (ACM Press, 2013). https://doi.org/10.1145/

2512530.2512533.

[22] F. Eyben, et al., The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, IEEE Trans. Affect. Comput. 7 (2016) 190–202.

[23] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: The Munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM International Conference on Multimedia, ACM, 2010, pp. 1459–1462. https://doi.org/10.1145/1873951.1874246.

[24] M. Pishgar, F. Karim, S. Majumdar, H. Darabi, Pathological voice classification using mel-cepstrum vectors and support vector machine, ArXiv181207729 Cs Eess Stat (2018).

[25] S.-H. Fang, et al., Detection of pathological voice using cepstrum vectors: a deep learning approach, J. Voice 33 (2019) 634–641.

[26] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, Chapman and Hall/CRC, 1984.

[27] J. Kreiman, B.R. Gerratt, M. Garellek, R. Samlan, Z. Zhang, Toward a unified theory of voice production and perception, Loquens 1 (2014).

[28] B. Harel, M. Cannizzaro, P.J. Snyder, Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: a longitudinal case study, Brain Cogn. 56 (2004) 24–29.

[29] P. Gómez-Vilda, et al., Parkinson disease detection from speech articulation neuromechanics, Front. Neuroinform. 11 (2017).

[30] J.A. Whitfield, D.D. Mehta, Examination of clear speech in Parkinson disease using measures of working vowel space, J. Speech Lang. Hear. Res. 62 (2019) 2082–2098.

[31] T. Foltynie, F.E. Matthews, L. Ishihara, C. Brayne, MRC CFAS, The frequency and validity of self-reported diagnosis of Parkinson's Disease in the UK elderly: MRC CFAS cohort, BMC Neurol. 6 (2006) 29.

[32] J. Varghese, et al., A smart device system to identify new phenotypical characteristics in movement disorders, Front. Neurol. 10 (2019).