Original Paper

Robust Feature Engineering for Parkinson Disease Diagnosis: New Machine Learning Techniques

Max Wang[1], BAdvComp (R&D) (Hons); Wenbo Ge[1], BE (Hons); Deborah Apthorp[1,2], PhD, BPsych (Hons); Hanna Suominen[1,3,4], MSc, Docent, SFHEA, PhD

[1]Research School of Computer Science, College of Engineering and Computer Science, The Australian National University, Canberra, ACT, Australia
[2]School of Psychology, Faculty of Medicine and Health, University of New England, Armidale, NSW, Australia
[3]Machine Learning Research Group, Data61, The Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, Australia
[4]Department of Future Technologies, Faculty of Science and Engineering, University of Turku, Turku, Finland

Corresponding Author:
Hanna Suominen, MSc, Docent, SFHEA, PhD
Research School of Computer Science
College of Engineering and Computer Science
The Australian National University
ANU Research School of Computer Science, Hanna Neumann Building, Room 2.35
145 Science Road
Canberra, ACT, 2600
Australia
Phone: 61 431 913 826
Email: hanna.suominen@anu.edu.au

Abstract

Background: Parkinson disease (PD) is a common neurodegenerative disorder that affects between 7 and 10 million people worldwide. No objective test for PD currently exists, and studies suggest misdiagnosis rates of up to 34%. Machine learning (ML) presents an opportunity to improve diagnosis; however, the size and nature of data sets make it difficult to generalize the performance of ML models to real-world applications.

Objective: This study aims to consolidate prior work and introduce new techniques in feature engineering and ML for diagnosis based on vowel phonation. Additional features and ML techniques were introduced, showing major performance improvements on the large mPower vocal phonation data set.

Methods: We used 1600 randomly selected /aa/ phonation samples from the entire data set to derive rules for filtering out faulty samples from the data set. The application of these rules, along with a joint age-gender balancing filter, results in a data set of 511 PD patients and 511 controls. We calculated features on a 1.5-second window of audio, beginning at the 1-second mark, for a support vector machine. This was evaluated with 10-fold cross-validation (CV), with stratification for balancing the number of patients and controls for each CV fold.

Results: We showed that the features used in prior literature do not perform well when extrapolated to the much larger mPower data set. Owing to the natural variation in speech, the separation of patients and controls is not as simple as previously believed. We presented significant performance improvements using additional novel features (with 88.6% certainty, derived from a Bayesian correlated t test) in separating patients and controls, with accuracy exceeding 58%.

Conclusions: The results are promising, showing the potential for ML in detecting symptoms imperceptible to a neurologist.

(JMIR Biomed Eng 2020;5(1):e13611) doi: 10.2196/13611

KEYWORDS

machine learning; mobile phone; nonlinear dynamics; Parkinson disease; signal processing, computer-assisted; speech; biomarkers

Introduction

Background

Parkinson disease (PD) affects approximately 1% of the population by the age of 70 years. It is characterized by the deterioration of dopamine-producing neurons in the brain, resulting in symptoms such as abnormal gait, speech, and tremor [1]. Current treatments can provide temporary relief from symptoms and slow its progression [2]; however, these treatments cannot repair damage from the disease. Thus, obtaining an accurate early diagnosis is of high importance.

PD is currently diagnosed with a standardized but subjective test administered by a neurologist or a clinician, the *Unified Parkinson's Disease Rating Scale* (UPDRS) [3]. It is not easy to diagnose, as only a subset of symptoms is present in any one patient [4] and there are many diseases with similar symptoms [5]. PD is especially difficult to diagnose in its early stages, as it is believed that most symptoms only manifest once 20% to 40% of dopamine-producing neurons have deteriorated [6]. Autopsy is one of the only reliable ways to confirm diagnosis, and studies have shown a misdiagnosis rate ranging between 9% and 34% [3,7].

Therefore, the search for a more objective measure for diagnosis is a timely topic in the research community. Discovering more quantifiable biomarkers from sources such as gene expression [8] and bodily fluids [9] is a promising option; however, it is likely that costs will be prohibitive for most early-stage patients uncertain about diagnosis. *Machine learning* (ML) on data from more accessible sources is another viable option, potentially offering an objective and low-cost tool to assist the neurologist in diagnosis through a smartphone-derived version of the UPDRS [10].

Objectives

In this study, we consolidated feature engineering techniques with advances in ML to develop a strong model for PD diagnosis on a large vocal phonation data set. We explored the challenges involved in training ML models on noisy, crowdsourced data and delved into the field of signal processing for audio data. We also found that the experimental setup for classifying healthy and diagnosed individuals does not match the diagnosis process of a neurologist. In addition to merely presenting results as a performance metric, we provided insights into the behavior of these models and showed that it is possible for ML to exceed the performance of clinicians in precise phonation analysis and potentially uncover new biomarkers for PD.

There is a major interest in using ML to assist in PD diagnosis, and current results are positive—often the reported accuracy percentage is in the high 90s using only speech or accelerometer data [11]. However, these results should not be taken at face value, as the experimental setup involves differentiating between (potentially incorrectly) diagnosed PD patients and *healthy controls* (HCs). This oversimplifies the complexities involved in a neurologist's diagnosis in a clinical situation, where the clinician must exclude a number of other causes for the symptoms and handle early-stage patients exhibiting minimal symptoms. Furthermore, the data sets associated with these publications generally consist of fewer than 40 subjects. For such small samples, it is difficult to control for bias and to prevent the overfitting of ML models.

Considering the issues with current data sets, there is an open question as to how ML models should be evaluated. One obvious option is to create a data set by monitoring subjects before any Parkinsonian symptoms until they pass, where the existence of PD can be confirmed through autopsy. With such a longitudinal data set, we could directly compare the performance of ML and neurologists. However, such a data set would be very costly and logistically difficult to collect. To advocate for its collection, there needs to be some evidence of ML effectiveness [10].

Consequently, the question we are most interested in is whether ML techniques can extract more information than observations from a trained clinician. As neurological diagnosis relies on judgment from observation, it is possible that some symptoms are imperceptible but detectable with ML on high-resolution sensor data. Specifically, our research question was to examine whether speech symptoms imperceptible to the human ear can be detected in microphone data using ML.

A larger data set than those that tend to be used in most studies is required to ensure that the results are statistically robust and not influenced by bias. This is not to say that having a larger data set will ensure that there is less influence of bias, but rather that from a larger data set, we can take subsets that are less influenced by biases. We chose the *mPower data set* [12], which contains phonations of the vowel */aa/* from 6000 patients as recorded by an iPhone microphone at 44,100 Hz.

Methods

Features

In this subsection, we describe the methodological background of our study and its feature engineering and signal processing. The subsections after this explain the novel features and materials in our experimental setup. The 2 objectives of these experiments were (1) to consolidate and replicate prior work on a much larger scale and (2) to introduce additional features for dysphonia signal processing to better understand how these features relate to a clinician's diagnosis with speech.

Vocal phonation is the prolonged pronunciation of a sound such as */aa/*. It is an interesting task for diagnosing PD, with evidence symptoms present earlier than other motor symptoms [13]. It also avoids the complexities involved in modeling speech and is, therefore, an easier task in the context of ML. Prior works have shown promising performance, with accuracies of up to 91.4% [14] and 98.6% [11].

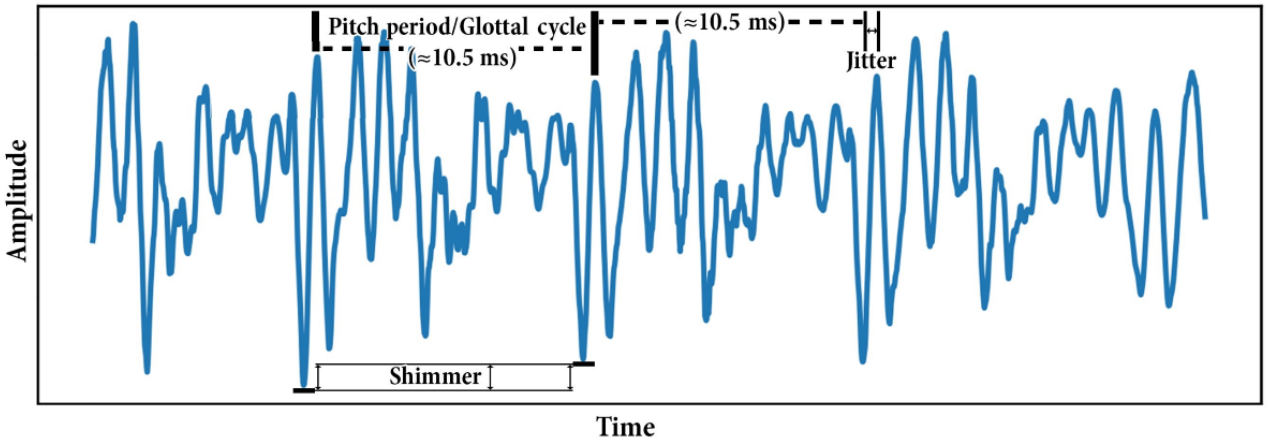
Biologically, phonation is produced by 2 components: the vocal folds and the vocal tract. The vocal folds consist of a flap called the *glottis*, which can be opened and closed. During phonation, air expelled from the lungs causes the glottis to oscillate, producing sound at a range of frequencies. The lowest of these frequencies—the *fundamental frequency*, f_0—represents the duration of 1 oscillation and is often denoted as the *pitch period*. The vocal tract comprises components such as the mouth and nose, and it *shapes* the sound by amplifying and attenuating

certain frequencies. The vocal folds and tract can be viewed as a *source-filter model*, where the source of vocal folds generates the sound signal, shaped (or filtered) by the vocal tract. With PD, the impairment of fine motor control reduces control of the glottis, causing incomplete vocal fold closure. The turbulent airflow around the glottis causes a sound described as a *breathy* or *hoarse* voice and results in increased variation over each glottal cycle. This is termed as *dysphonia*. A similar phenomenon occurs when the vocal folds are irritated by physical causes, such as colds, and it is currently unknown whether differentiation between neurological and physical dysphonia is possible. It is also worth noting that airway diseases as well as muscular diseases and other psychological disorders can also affect the process of sound generation.

People with PD also experience hesitant speech from reduction of cognitive ability and slurred or imprecise articulation from loss of motor control over the vocal tract. This is termed as *dysarthria*. Although dysarthria is very noticeable to humans, it is difficult to quantify it computationally. Spoken language has a wide variety of accents and styles, and it has been shown that models trained on English speakers do not generalize well on German speakers, and vice versa [15].

Dysphonia can be measured using *sustained vowel phonations*, which are easier to model with traditional signal processing approaches and are suitable for small data sets. Dysphonia in vowel phonation is shown in Figure 1. Optimally, both dysphonia- and dysarthria-related features would be used to build models; however, this study focused specifically on dysphonia because of limitations of the data set under consideration.

Figure 1. /aa/phonation from an individual with dysphonia. Variations in jitter (variation in glottal cycle periods) and shimmer (variation in glottal cycle amplitude) are common features for the detection of dysphonia. Algorithms computing each glottal cycle are imperfect, especially for heavy dysphonia.



Features for Dysphonia

Data from a microphone are represented as a stream of values, corresponding to the value of the recorded sound wave at each point in time. A basic microphone samples at approximately 44,100 Hz and quantizes the wave to 2^{16} possible values. The difference between higher sampling rates or more detailed quantization (such as 24 bit) is minimally perceptible to human ears; however, a major problem with low-quality microphones is additional noise and imprecise wave encoding.

Approaches based on signal processing use features that estimate the qualities associated with dysphonia. We divided them into 3 groups: *general* techniques, which are suitable for any time series signal; *dysphonia-specific* techniques, which have been used in prior work specifically to model dysphonia; and *novel* features, which we have shortlisted from other applications as being potentially effective in quantifying dysphonia.

General Signal Processing

Moments are basic statistical descriptors of a signal, with the first 3 moments representing mean, variance, and skewness. For speech, the mean is largely uninformative, and variance corresponds to volume. The zero-crossing rate measures how quickly the signal oscillates around 0 and is a measure of the signal frequency.

Entropy describes the amount of information in a piece of data, if it were modeled by a Bernoulli scheme. It is a simple measure of the complexity/information content of a signal.

The *Fourier* transform decomposes a signal into the amplitudes of frequencies that compose it. This is referred to as mapping from the *time* domain to the *frequency* or *spectral* domain. For speech, this enables us to determine the frequency bands with more energy, corresponding to the fundamental frequency and harmonics. After performing a Fourier transform, the *spectral entropy* [16] can be calculated, which can measure how *sharp* the f_0 and harmonics of speech are. People with PD are expected to have a lower spectral entropy because of the less precise frequency control, causing a more *blurred* Fourier transform.

Squared energy operators, *Teager-Kaiser* energy operators, or other energy operators can obtain the instantaneous frequency and amplitude of a signal [17]. Statistics such as mean or SD and other measures can be computed after applying an energy operator.

Dysphonia Signal Processing

The major feature present in dysphonic speech is the increased variation between each glottal cycle (Figure 1). *Jitter* measures the variation in the length of each glottal cycle, and *shimmer* [18] is the variation in amplitude. These often rely on detecting each glottal cycle, which is not very accurate with current algorithms [19]. The *harmonics-to-noise ratio* (HNR) [20] measures the amount of noise in a signal, which correlates with the *hoarseness* or *breathiness* of speech. The HNR has been improved with a more robust *glottal-to-noise excitation* ratio [21].

The *vocal fold excitation ratio* (VFER) is another extension of the HNR developed in the study by Tsanas et al [22], which also introduced the *Glottal Quotient* (GQ), a measure of the SD duration while the glottis is opened and closed. Both VFER and GQ are built upon concepts of the fundamental frequency estimation algorithm [23].

Mel-frequency cepstral coefficients (MFCCs) are one of the most effective features for speech recognition models, so it is no surprise that they are shown to be similarly effective for dysphonia. Speech recognition involves computing the MFCC at short time intervals and using a Markov model or structured neural network to model temporal information, whereas statistical descriptors have been shown to be effective in detecting dysphonia [11]. There still exists a gap in understanding the relationships between the coefficients and dysphonia.

A recent study [24] showed that *detrended fluctuation analysis*, originally introduced as a measure of the autocorrelation (*autocorrelation* describes the similarity of a signal to itself when offset by a given interval) of a signal, changes with the amount of turbulent airflow in speakers with dysphonia. This study has also proposed the *recurrence period density entropy* (RPDE), which characterizes the periodicity of a signal. These measures are expected to be lower for speakers with dysphonia because of the noise introduced by turbulent airflow. Another study [14] has built upon RPDE to develop *pitch period entropy* as a better measure of the impaired control of pitch experienced by PD patients.

Novel Features

Although existing features have achieved good results on their respective data sets, we have obtained improved performance with this additional set of features. These novel features may not directly relate to dysphonia but have been effective in other signal processing applications, most commonly *electroencephalogram* (EEG), which, similar to speech, is difficult to characterize.

A number of these features relate to *chaos theory*—a field based on understanding the behavior of dynamical systems sensitive to initial conditions [25]. One can imagine the generation of a speech signal as a system, where parameters involve the state of components in the vocal tract. Given no change in parameters such as vocal fold tension, regular */aa/* phonation can be modeled with much fewer dimensions in *phase* space [26]—all possible states of a dynamic system.

The *Lyapunov exponents* quantify the divergence of 2 systems with similar initial parameters. The *largest Lyapunov exponent* (λ^*) characterizes the chaos in a system and is commonly estimated with the algorithm described in the study by Rosenstein et al [26], which reconstructs the system dynamics using a time delay technique. The inverse of $\frac{1}{\lambda^*}$ is the Lyapunov time, which defines how long the behavior of a system can be predicted.

The *fractal dimension* is also commonly used in the analysis of dynamical systems. It represents the ratio of the log change in detail to the log change in scale of a signal [27]. This is similar to the coastline paradox, where measuring a coastline with smaller sticks results in an apparent increase in length. The fractal dimension correlates to the complexity of a signal. It has been shown that the fractal dimension of elderly individuals balancing on a force plate is greater than that of younger individuals [28] and, more importantly, that distinct patterns seen in deterministic recurrence quantification analysis of sway (similar to fractal dimension) distinguish patients with PD from controls [29].

General entropy will not differentiate 2 sequences where the frequency of each variable is the same; however, the sequences 0, 0, 0, 0, 1, 1, 1, 1 and 0, 1, 0, 0, 1, 1, 0, 1 are clearly generated by different stochastic processes. The *approximate* and *sample entropy* aims to quantify this [30]. This is extended with multiscale sample entropy [31], which is an especially powerful tool in the analysis of biological signals.

Although signals may appear to have high information content in the time domain, they may be easier to represent in others. For example, the JPEG image compression format primarily relies on human vision, which is less sensitive to high-frequency image details. Images are compressed by taking a Fourier transform and downsampling the high-frequency information. *Spectral entropy* measures the information content of the signal in its frequency-domain representation. The *singular value decomposition* (SVD) factorizes a matrix into orthogonal matrices and *singular values*. *SVD entropy* [16] measures the entropy of the singular values obtained when the signal is embedded with the algorithm described in the study by Rosenstein et al [26].

Many of these features are not simple to interpret; however, our testing shows that they provided significant improvements over prior features used in the literature.

Materials

We chose the mPower data set [12]—a crowdsourced data set consisting of 65,000 */aa/* phonation samples from 6000 participants, of which 1200 were participants with PD.

The primary issue with mPower is quality. Owing to crowdsourcing, participants who are dishonest or perform the task incorrectly can skew results. There are a number of subjects with young-onset PD in mPower; young-onset PD is very uncommon in the general population, but this representation may be possible because of the younger target audience of smartphone apps. Vowel phonation has been captured with a single channel iPhone/iPod microphone at 44,100 Hz, with the

different microphone technology in each generation introducing another variable to the model. Hesitation and phonation of vowels other than /aa/ are common, and the distance between the phone and the user varies: with some speaking directly into the microphone, creating *wind noise*, and others at a distance, introducing significant environmental noise.

We evaluated 1600 randomly selected phonation samples for performing the task correctly, rejecting approximately 40% by using the aforementioned wind, distance, and environmental noise. Using short time energies extracted at 0.1-second intervals with OpenSMILE [32], simple metrics such as variance, mean, and ranges were calculated to rank and filter the samples. Placing a threshold on these gave rise to hand-crafted rules to filter the remaining samples. After filtering, 4100 users remained, 900 with PD, each with a single corresponding sample. We then attempted to balance the joint age and gender distribution within the PD and control groups. Every PD participant had a gender and age *twin* (±2 years; the age-matching was conducted in a way that the twin was at most 2 years younger or older than the respective participant) selected; if a *twin* could not be found, that PD participant was discarded. This resulted in 1022 participants, 511 with PD. Standard libraries for computing the features used to build the modeling included the PyREM package for sleep staging from EEG data [33], nolds [34], and pypsr [35] (Textbox 1). The average f_0 for males and females was assumed to be 120 Hz and 190 Hz, respectively. Sound bit depth was binned to $\sqrt{\frac{n}{2}}$ values for entropy-related calculations. In this feature experiment, we used 6 as the embedding dimension for the time delay embedding methods and calculated τ following the study by Rosenstein et al [26]. Ethical aspects of this research have been approved by the ANU Human Research Ethics Committee with protocol number 2018/108.

Textbox 1. Summary of the features used to build the model.

Dysphonia features

- Prior dysphonia features from the study by Tsanas et al [11]
- Wavelet transform–based features [36]

Novel features

- Higuchi and Petrosian fractal dimension
- Hurst exponent
- Time delay (τ)
- Lyapunov exponents (up to 6)
- Sample and approximate entropy (with *r* selected by Lu et al [37]) and Fisher information [38]
- Spectral entropy
- Singular value decomposition entropy

Results

Initial Model Validation: Replicating the Previous Results

In a previous study, Tsanas et al [11] used the National Crime Victimization Survey data set, which consists of 263 phonations from 33 people with PD and 10 HCs. A set of 132 features was calculated, and 100 times repeated 10-fold cross-validation (CV) was used to evaluate a *support vector machine* (SVM). On all features, they achieved 97.7% accuracy, and on a 10-feature subset selected by ReliefF, they achieved 98.6% accuracy. They also released an open-source toolbox to assist in replicating the results.

Using this toolbox on our data set, we calculated features on a 1.5-second window of audio beginning at the 1-second mark. An SVM was evaluated with 10 times repeated 10-fold CV over the full feature set and the 10-feature ReliefF subset (Table 1). The data set was stratified by random sampling such that there were equal numbers of PD and control subjects in each fold of CV to better highlight performance. Note that this effectively reduced the data set to approximately 1700 phonations. A grid search was used to select the optimal hyperparameters for the SVM. Of the kernels tested, the *radial basis function* (RBF) was dominant in all models. A Bayesian correlated *t* test on accuracy was used to determine if a model dominated another [39]. Accuracy was chosen, despite the recent popularity of the *area under the receiver operating characteristic curve* (AUROC) because of criticisms of its bias toward certain predictors [40]. However, for the purposes of comparison with previously published work, we also reported sensitivity, specificity, and AUROC.

Table 1. Cross-validation results of a support vector machine using the full feature set and 10-feature ReliefF subset.

Performance measures	Full, mean (SD)	ReliefF subset[a], mean (SD)
Accuracy (%)	55.6 (4.9)	55.6 (4.5)
Sensitivity (%)	52.3 (7.0)	47.4 (6.6)
Specificity (%)	58.8 (7.8)	63.8 (8.1)
Area under the receiver operating characteristic curve (%)	57.7 (5.7)	58.2 (5.2)

[a]There is a 50.7% probability that the average performance of the ReliefF subset outperforms the full feature set.

Initial Model Validation: Improving Previous Results With Novel Features

The performance of the features by Tsanas et al [11] on the mPower data was evidently worse than that of the NCVS data set. It is possible that this was a consequence of the noisier and lower-quality mPower data. However, overfitting in the original study by Tsanas et al [11] is also highly likely, especially as evidenced by the fact that the MFCC-heavy ReliefF feature subset performed better in experiments by Tsanas et al [11], whereas it performed substantially worse in our testing, achieving 55.6% accuracy rather than 98.6%. Another consideration is the number of observations within each group; having more participants within one group can arbitrarily increase accuracy. It is also ambiguous whether the authors had stratified the phonations on a per-subject scale—failure to do so introduces the *digital fingerprinting* effect. The analysis provided in a recent study [41] exemplifies this, showing that allowing phonations from a subject to appear in both training and testing results in an AUROC of approximately 96%, compared with only 59% in a correctly stratified example.

The features used in the prior literature were clearly insufficient to perform an accurate diagnosis. To improve the model, we introduced an additional range of features that have been effectively applied in the analysis of other biological signals. These features are not directly related to human hearing or speech and thus may be especially useful in detecting symptoms unnoticeable by an expert. We repeated the experiments, adding the novel features, with the results presented in Table 2 showing significant improvement.

Table 2. Cross validation results of the support vector machines using different feature sets.

Performance measures	ReliefF subset, mean (SD)	Novel only, mean (SD)	Combined[a], mean (SD)
Accuracy (%)	55.6 (4.5)	55.3 (4.8)	58.2 (5.1)
Sensitivity (%)	47.4 (6.6)	45.1 (6.5)	56.4 (6.8)
Specificity (%)	63.8 (8.1)	65.5 (7.6)	60.0 (7.0)
Area under the receiver operating characteristic curve (%)	58.2 (5.2)	58.2 (5.2)	60.7 (5.9)

[a]There is an 88.7% probability that the average performance of the combined subset outperforms the ReliefF subset.

Further Improvements: Ensemble Models and Data Augmentation

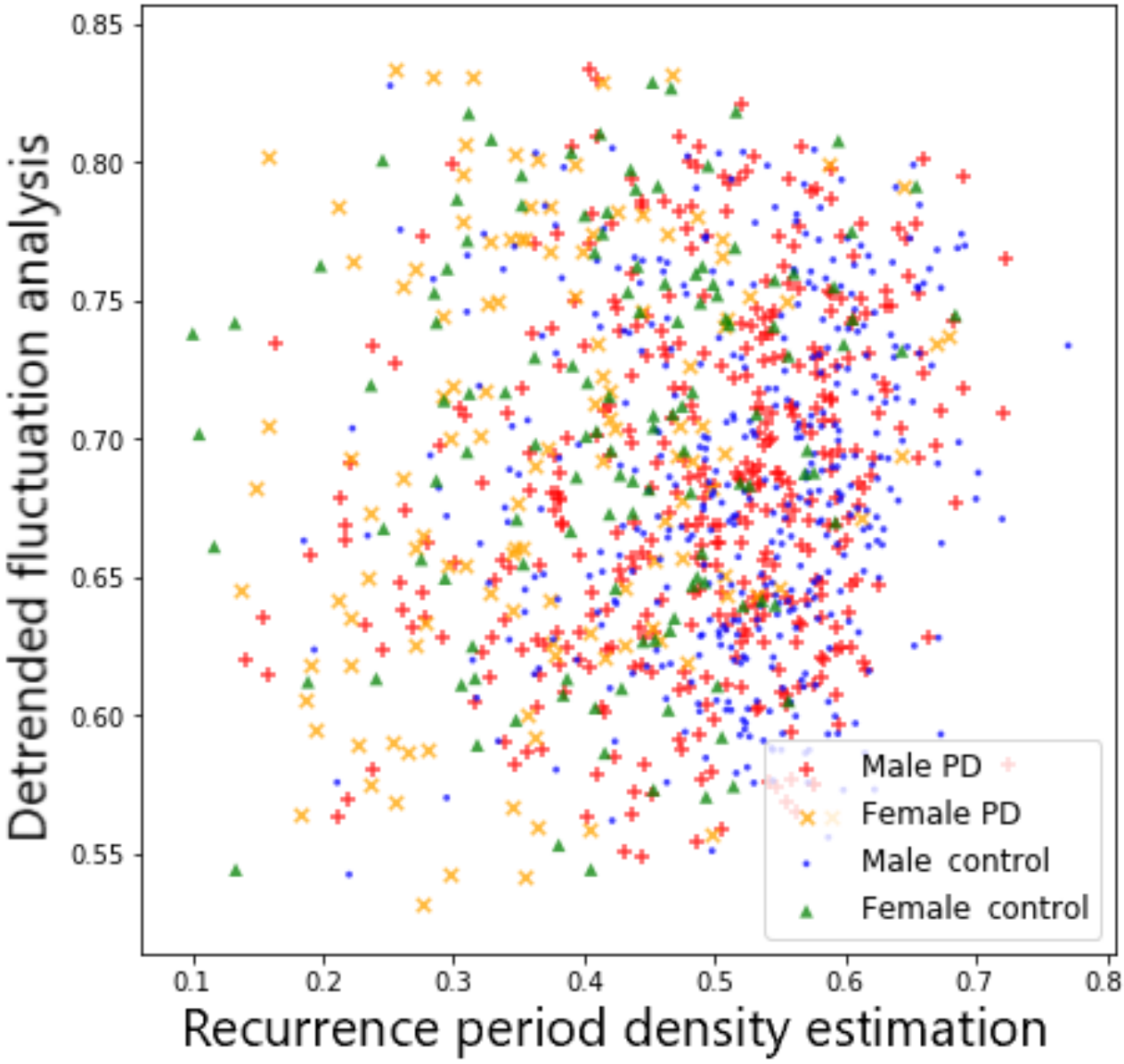
Although the performance was improved, a 58.2% diagnosis accuracy is still far below the requirements for clinical use. In this section, we explore the reasons for the below-expected performance and methods of improving it.

Many dysphonia features rely on estimating the precise length of each glottal cycle, for which the f_0 algorithm of Camacho [42] was used. A preliminary investigation showed that the SD of f_0 exceeded 10 Hz for 18% of the phonations. This was indicative of a substantial failure of the algorithm or a poorly executed recording.

Ensemble Models

Visualizing the features as in Figure 2, it is evident that there are distinct distributions for individuals who are healthy and have PD as well as for males and females. However, there is a large amount of variance and overlap over these distributions, making it infeasible to perform diagnosis over one feature alone.

Figure 2. Visualization of detrended fluctuation analysis against recurrence period density entropy in male and female participants with PD and controls. Although the groups follow distinct distributions, there is heavy overlap. It is evidently difficult to classify PD based on any single feature, and powerful machine learning models are required to make sense of their relationships. PD: Parkinson disease.



Typically, an SVM only performs well if the features are good, which may require a lot of prior knowledge. This is because SVMs can only model linear relationships and, even with the use of nonlinear kernels, are restricted to that particular kernel and are often simple. Other algorithms, such as neural networks, can potentially learn to exploit complex nonlinear relationships between the features to increase performance.

Given the complex relationships between features, it is possible that an ensemble model may achieve better performance. Ensemble models utilize multiple ML models and combine them to create a more powerful one. These ML models can typically be combined by averaging (ie, *voting*) or with other methods (eg, *stacking*), making the combined prediction more powerful than its individual parts.

Building on stacking, *feature-weighted linear stacking* (FWLS [43]) assigns a weight to each feature and model combination. This is motivated by different models being more suitable for using certain features. FWLS was the technique used to ensemble the hundred-model winner of the prestigious Netflix Prize [44].

We ensembled an SVM; Gaussian process; random forest; *k*-nearest neighbor classifier (with k=3); and simple dense neural networks with 3, 5, and 7 hidden layers. All models were suited to the task, diagnosing PD within 3% accuracy of each other. An RBF Gaussian process was chosen to aggregate the models in the stacking and FWLS-based ensembles, as Gaussian processes are inherently probabilistic and suitable for making decisions in situations of high uncertainty. The results are presented in Table 3.

Table 3. Cross validation results of different ensemble methods compared to the support vector machine-only model, using the combined feature set.

Performance measures	Support vector machine only, mean (SD)	Voting, mean (SD)	Stacking, mean (SD)	Feature-weighted linear stacking[a], mean (SD)
Accuracy (%)	58.2 (5.1)	56.5 (5.0)	59.1 (5.1)	5.92 (5.3)
Sensitivity (%)	56.4 (6.8)	54.9 (9.1)	57.4 (7.8)	57.4 (7.6)
Specificity (%)	60.0 (7.0)	58.1 (8.2)	60.9 (7.4)	61.1 (7.4)
Area under the receiver operating characteristic curve (%)	60.7 (5.9)	59.1 (5.6)	62.2 (5.8)	62.3 (5.9)

[a]There is a 70.6% probability that the average performance of the feature-weighted linear stacking ensemble outperforms the support vector machine-only model.

Training ensembles is more computationally expensive; however, in practice, making a prediction from a pretrained model will take a negligible amount of time. Larger data sets may require the use of localized approximations for polynomially computable models such as SVMs and Gaussian processes. However, a major disadvantage of ensembles is the *black box* effect, where their predictions are effectively uninterpretable.

Data Augmentation

Some features were highly sensitive to minor fluctuations in the signal, with their value changing drastically depending on the segment used. Many of these features were not length invariant, and 10% varied by over 0.5 SDs when computed over the same phonation, offset by 0.1 seconds. We computed seven 1.5-second samples from each phonation, ranging from the 1.5- to 4.5-second mark with a 0.5-second step size. We experimented with taking the mean over all 7 values as well as augmenting the data by using the additional samples as extra CV input (ensuring that phonations from the same participant appeared in the same training set or test set). The results are presented in Table 4.

Table 4. Cross validation results of basic data augmentation techniques with the combined feature set and the support vector machine-only model.

Performance measures	Original, mean (SD)	Mean,[a] mean (SD)	Augmented, mean (SD)
Accuracy (%)	58.2 (5.1)	58.7 (4.5)	57.3 (3.7)
Sensitivity (%)	56.4 (6.8)	56.5 (6.8)	54.8 (5.8)
Specificity (%)	60.0 (7.0)	60.8 (7.5)	59.8 (5.8)
Area under the receiver operating characteristic curve (%)	60.7 (5.9)	61.1 (5.1)	60.2 (4.4)

[a]There is a 59.3% probability that the performance with the mean augmented features is better than the nonaugmented features.

Augmentation did not seem to prove useful in overcoming low-quality data and unstable features and seemed to decrease performance. It is possible that sampling from a shorter window exacerbates the instability of the features and that even if one of the windows was more informative than others, that window might not be the same for all recordings. This may essentially introduce more noise into the training data and therefore have an overall negative effect on model performance.

Augmentation proved beneficial when taking the mean over all segments; however, it was not useful when simply including the additional segments as CV data. This is likely because taking the mean reduces the instability and noise of the features, whereas including all the segments exacerbates the problem and introduces more noise into the entire feature set.

Final Results: Performance in Participants With No Speech Difficulties

After developing a decent model to classify PD based on speech, we investigated what it could offer to real-world diagnosis. The raw performance is clearly insufficient to replace neurologists; however, we have shown that the performance bound has not been reached—additional, higher quality data will greatly improve results, along with better and more informative features.

First, we investigated whether it would be possible to detect symptoms imperceptible to a neurologist. If this is possible, the combination of both could potentially increase the accuracy of diagnosis. To do this, a set of PD participants who had *no speech difficulties* (NSDs) was first removed from the data set before any experimentation (including those mentioned in the previous sections) to prevent overfitting during the model evaluation stage. This relied on a field of the mPower UPDRS survey, which defaulted to zero, which we are aware relies on the honestly/reliability of participants. This created a subset of 81 participants (of the previous 4112) who had PD and NSDs.

From the HC participants not used in CV/model evaluation stage, 81 participants were chosen such that they exactly matched the gender of each NSD participant and as closely matched the age as possible. We combined the 2 to form a test set of 162 participants and used the FWLS ensemble trained on the CV set to evaluate the test set. No data augmentation was performed. This is presented in Table 5. Surprisingly, there was an increase in sensitivity between diagnosis of participants with perceptible dysphonia (as indicated by participants having speech difficulties) compared with the diagnosis of those with imperceptible dysphonia (as indicated by participants having NSDs). This implies that there must exist some features that

capture information regarding the disease state of a person but does not relate to audible dysphonia.

Table 5. Performance of the feature-weighted linear stacking ensemble on participants with no audible symptoms and previously unseen data.

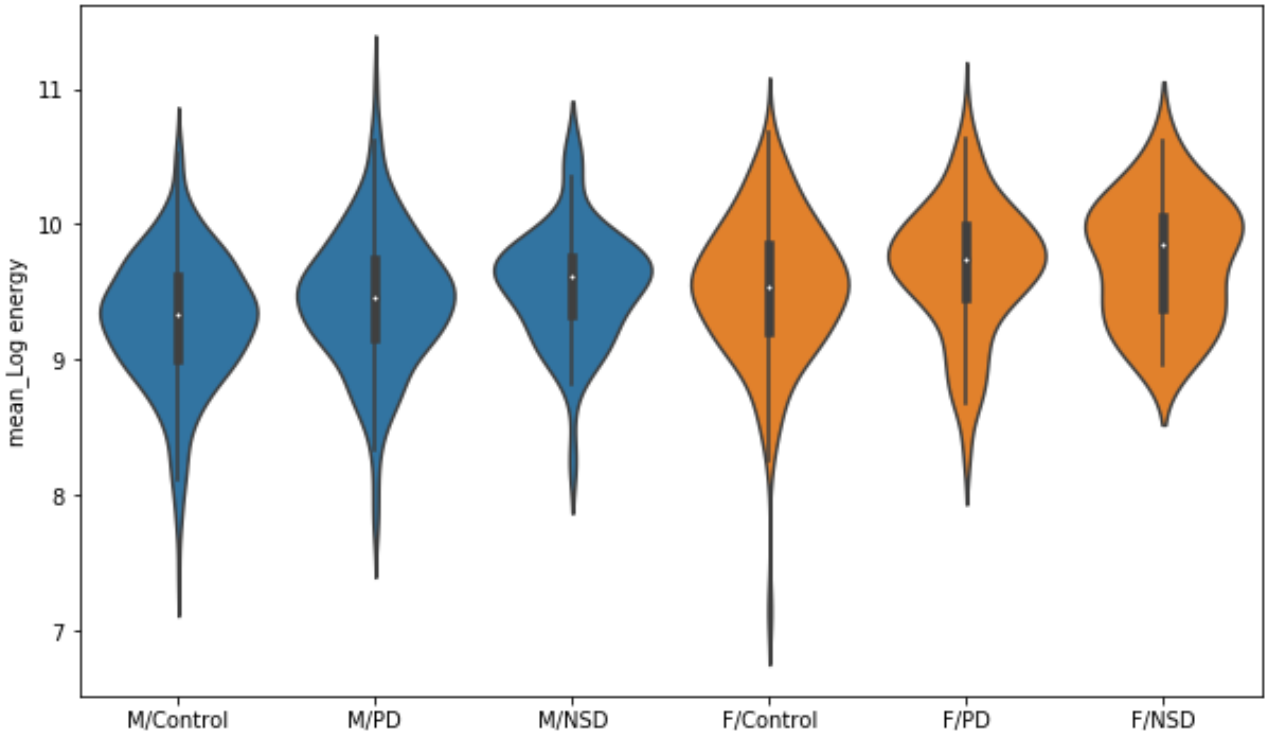
Performance measures	Cross-validation mean	Cross-validation SD	No speech difficulty
Accuracy (%)	59.2	5.3	59.3
Sensitivity (%)	57.4	7.6	63.0
Specificity (%)	61.1	7.4	55.6
Area under the receiver operating characteristic curve (%)	62.3	5.9	66.7

We hypothesized that this is possible because of the more abstract features, which are not related to audible symptoms. To investigate, for each feature, we checked two matters: first, whether the distribution over NSD and PD was similar (ie, captured information about disease state), and second, whether the distribution between NSD and control was different (ie, did not relate to audible dysphonia), implying that the feature is likely discriminative yet inaudible.

To perform significance testing, we first employed a normality test on all features across all groups. If a single feature was normal across all groups (ie, PD, NSD, and HC), then we employed a one-way analysis of variance (ANOVA) to test for significant differences, followed by a Tukey honestly significant difference post hoc. Groups were considered to be sampled from different distributions of a *P* value of .05. If that feature was not normal across all groups, then a Kruskal-Wallis ANOVA was used, followed by a Dunn-Bonferroni posthoc test. Similarly, a value of $P=.04$ is required to say that the feature was sampled from different distributions. Note that because we are only interested in 2 posthoc comparisons (between NSD-PD and NSD-HC), the posthoc tests overcorrected the *P* values. This meant that we can be more confident in any significant difference detected.

We expected that abstract features such as those that try to capture complexity (eg, sample entropy) were likely to be inaudible. These were not biologically motivated and more sensitive to high-frequency information to which the human auditory system was not. This was confirmed in Figure 3, where a significant difference was detected for mean log energy between the NSD and control groups, yet no significant difference was detected between the NSD and PD groups, for males only.

Figure 3. Violin plot of mean_Log energy for males (blue) and females (orange). For males, a significant difference was detected between the no speech difficulty and control groups ($P<.00002$), but not between the no speech difficulty and Parkinson disease groups ($P=.13$). F: female; Log: logarithmic; M: male; NSD: no speech difficulties; PD: Parkinson disease.



In the full results, we found that features such as energy, entropy, and their corresponding coefficients were inaudible, whereas features such as jitter, those derived from fundamental frequency, MFCC, and their corresponding coefficients were audible.

Discussion

Principal Findings

In this study, we have shown that features used in prior literature on small data sets do not perform well when extrapolated to more real-world, much larger data sets such as mPower. Owing to natural variation in speech, the separation of healthy individuals and those with PD is not as simple as previously believed. We have presented significant performance improvements with an additional set of features and techniques such as assembling and data augmentation. Importantly, we have demonstrated this in a robust environment that is well guarded against overfitting.

Strengths and Limitations

Although performance is currently insufficient to substitute for a clinician's diagnosis, significant improvements from simple data augmentation, ensemble models, and additional features imply that peak performance has not yet been achieved. The power of ML only increases as more data become available.

We have also shown that some features used are capable of detecting symptoms of the PD state, yet are *imperceptible to human hearing*. This is a very promising result for the field, with the possibility that a large robust model may eventually outperform humans. For now, these models are a low-cost tool for clinicians to validate their diagnosis; a positive result may be a flag to perform further checks on the patient. This also suggests that vocal phonation features could be used as part of a set of noninvasive biomarkers for PD.

The biggest barrier to the introduction of ML in the clinical setting is trust. This relies on *either* a strong understanding of the models and features *or* good performance on a substantially sized data set with robust evaluation approaches. We believe that developing a deep understanding of ML in clinicians will become increasingly difficult, as models and features become more complex. Thus, it may be best to focus on providing larger data sets. The mPower data set makes great strides in size and availability, but it lacks data control. The field would benefit greatly from a standardized data set against which the performance of models could be empirically evaluated.

Many avenues still remain unexplored. The diagnosis of PD often involves testing a patient's response to medication, such as levodopa. Additional information from phonation samples before and after levodopa consumption may greatly improve performance. We focused only on traditional ML approaches based on signal processing. A notable weakness of feature engineering is information loss, as it is difficult to perfectly describe a signal. Structured neural networks are suitable for making predictions from a raw data stream, such as computer vision and neural networks. In our testing, simple convolutional and recurrent neural networks have achieved similar performance to our best models and greatly improved performance when combined with feature engineering. However, they are more difficult to interpret.

Further improvements could be seen in multidimensional data sets using data from more than one source (eg, postural sway, gait, and finger tapping, all of which could be measured on smartphone devices). Measuring these during clinical visits in a controlled setting under standardized conditions would produce substantially cleaner data at minimal cost, but it is essential to have the cooperation of the clinical and patient communities. Our research provides a first step in establishing the great value these tools could provide, not just in PD.

Acknowledgments

This research was supported by the *Australian Government Research Training Program* Scholarship and delivered in partnership with *Our Health in Our Hands*, a strategic initiative of the Australian National University, which aims to transform health care by developing new personalized health technologies and solutions in collaboration with patients, clinicians, and health care providers. The authors would like to thank Alex Smith, Mehika Manocha, and Professor Christian Lueck for their helpful comments and suggestions.

Conflicts of Interest

None declared.

References

1. Savitt JM, Dawson VL, Dawson TM. Diagnosis and treatment of Parkinson disease: molecules to medicine. J Clin Invest 2006 Jul;116(7):1744-1754 [FREE Full text] [doi: 10.1172/JCI29178] [Medline: 16823471]
2. Fahn S, Parkinson Study Group. Does levodopa slow or hasten the rate of progression of Parkinson's disease? J Neurol 2005 Oct;252(Suppl 4):IV37-IV42. [doi: 10.1007/s00415-005-4008-5] [Medline: 16222436]
3. Tolosa E, Wenning G, Poewe W. The diagnosis of Parkinson's disease. Lancet Neurol 2006 Jan;5(1):75-86. [doi: 10.1016/S1474-4422(05)70285-4] [Medline: 16361025]
4. Thenganatt MA, Jankovic J. Parkinson disease subtypes. JAMA Neurol 2014 Apr;71(4):499-504. [doi: 10.1001/jamaneurol.2013.6233] [Medline: 24514863]
5. Quinn N. Parkinsonism--recognition and differential diagnosis. Br Med J 1995 Feb 18;310(6977):447-452 [FREE Full text] [doi: 10.1136/bmj.310.6977.447] [Medline: 7646647]
6. Brooks DJ. Parkinson's disease: diagnosis. Parkinsonism Relat Disord 2012 Jan;18:S31-S33. [doi: 10.1016/S1353-8020(11)70012-8]

7. Jankovic J, Rajput AH, McDermott MP, Perl DP. The evolution of diagnosis in early Parkinson disease. Parkinson study group. Arch Neurol 2000 Mar;57(3):369-372. [doi: 10.1001/archneur.57.3.369] [Medline: 10714663]
8. Scherzer CR, Eklund AC, Morse LJ, Liao Z, Locascio JJ, Fefer D, et al. Molecular markers of early Parkinson's disease based on gene expression in blood. Proc Natl Acad Sci U S A 2007 Jan 16;104(3):955-960 [FREE Full text] [doi: 10.1073/pnas.0610204104] [Medline: 17215369]
9. Hong Z, Shi M, Chung KA, Quinn JF, Peskind ER, Galasko D, et al. DJ-1 and alpha-synuclein in human cerebrospinal fluid as biomarkers of Parkinson's disease. Brain 2010 Mar;133(Pt 3):713-726 [FREE Full text] [doi: 10.1093/brain/awq008] [Medline: 20157014]
10. Zhan A, Mohan S, Tarolli C, Schneider RB, Adams JL, Sharma S, et al. Using smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score. JAMA Neurol 2018 Jul 1;75(7):876-880 [FREE Full text] [doi: 10.1001/jamaneurol.2018.0809] [Medline: 29582075]
11. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng 2012 May;59(5):1264-1271. [doi: 10.1109/TBME.2012.2183367] [Medline: 22249592]
12. Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data 2016 Mar 3;3:160011 [FREE Full text] [doi: 10.1038/sdata.2016.11] [Medline: 26938265]
13. Rusz J, Cmejla R, Tykalova T, Ruzickova H, Klempir J, Majerova V, et al. Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task. J Acoust Soc Am 2013 Sep;134(3):2171-2181. [doi: 10.1121/1.4816541] [Medline: 23967947]
14. Little M, McSharry P, Hunter E, Spielman J, Ramig L. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans Biomed Eng 2009 Apr;56(4):1015 [FREE Full text] [doi: 10.1109/TBME.2008.2005954] [Medline: 21399744]
15. Hazan H, Hilu D, Manevitz L, Ramig LO, Sapir S. Early Diagnosis of Parkinson's Disease via Machine Learning on Speech Data. In: 27th Convention of Electrical and Electronics Engineers in Israel. 2012 Presented at: EEEI'12; November 14-17, 2012; Eilat, Israel URL: https://ieeexplore.ieee.org/document/6377065 [doi: 10.1109/eeei.2012.6377065]
16. Roberts SJ, Penny W, Rezek I. Temporal and spatial complexity measures for electroencephalogram based brain-computer interfacing. Med Biol Eng Comput 1999 Jan;37(1):93-98. [doi: 10.1007/BF02513272] [Medline: 10396848]
17. Kaiser JF. ICASSP 90. 1990 International Conference on Acoustics, Speech and Signal Processing. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. 1990 Presented at: ICASSP'90; April 3-6, 1990; Albuquerque, NM URL: https://ieeexplore.ieee.org/document/115702 [doi: 10.1109/icassp.1990.115547]
18. Horii Y. Jitter and shimmer differences among sustained vowel phonations. J Speech Hear Res 1982 Mar;25(1):12-14. [doi: 10.1044/jshr.2501.12] [Medline: 7087413]
19. Tsanas A, Zañartu M, Little MA, Fox C, Ramig LO, Clifford GD. Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive Kalman filtering. J Acoust Soc Am 2014 May;135(5):2885-2901 [FREE Full text] [doi: 10.1121/1.4870484] [Medline: 24815269]
20. Yumoto E, Gould WJ, Baer T. Harmonics-to-noise ratio as an index of the degree of hoarseness. J Acoust Soc Am 1982 Jun;71(6):1544-1549. [doi: 10.1121/1.387808] [Medline: 7108029]
21. Michaelis D, Gramss T, Strube HW. Glottal to-noise excitation ratio - a new measure for describing pathological voices. Acta Acustica United with Acustica 1997;83(4):700-706 [FREE Full text]
22. Tsanas A, Little MA, McSharry PE, Ramig LO. 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 Jun 6;8(59):842-855 [FREE Full text] [doi: 10.1098/rsif.2010.0456] [Medline: 21084338]
23. Kounoudes A, Naylor PA, Brookes M. The DYPSA Algorithm for Estimation of Glottal Closure Instants in Voiced Speech. In: International Conference on Acoustics, Speech, and Signal Processing. 2002 Presented at: ICASSP'02; May 7-11, 2002; Orlando, FL URL: https://ieeexplore.ieee.org/document/5743726 [doi: 10.1109/icassp.2002.1005748]
24. Little MA, McSharry PE, Roberts SJ, Costello DA, Moroz IM. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Biomed Eng Online 2007 Jun 26;6:23 [FREE Full text] [doi: 10.1186/1475-925X-6-23] [Medline: 17594480]
25. Hegger R, Kantz H, Matassini L. Denoising human speech signals using chaoslike features. Phys Rev Lett 2000 Apr 3;84(14):3197-3200. [doi: 10.1103/PhysRevLett.84.3197] [Medline: 11019046]
26. Rosenstein MT, Collins JJ, de Luca CJ. A practical method for calculating largest Lyapunov exponents from small data sets. Physica D 1993 May;65(1-2):117-134. [doi: 10.1016/0167-2789(93)90009-P]
27. Mandelbrot B. How long is the coast of Britain? Statistical self-similarity and fractional dimension. Science 1967 May 5;156(3775):636-638. [doi: 10.1126/science.156.3775.636] [Medline: 17837158]
28. Doyle TL, Dugan EL, Humphries B, Newton RU. Discriminating between elderly and young using a fractal dimension analysis of centre of pressure. Int J Med Sci 2004;1(1):11-20 [FREE Full text] [doi: 10.7150/ijms.1.11] [Medline: 15912186]
29. Schmit JM, Riley MA, Dalvi A, Sahay A, Shear PK, Shockley KD, et al. Deterministic center of pressure patterns characterize postural instability in Parkinson's disease. Exp Brain Res 2006 Jan;168(3):357-367. [doi: 10.1007/s00221-005-0094-y] [Medline: 16047175]

30. Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol 2000 Jun;278(6):H2039-H2049 [FREE Full text] [doi: 10.1152/ajpheart.2000.278.6.H2039] [Medline: 10843903]
31. Costa M, Goldberger AL, Peng C. Multiscale entropy analysis of biological signals. Phys Rev E Stat Nonlin Soft Matter Phys 2005 Feb;71(2 Pt 1):021906. [doi: 10.1103/PhysRevE.71.021906] [Medline: 15783351]
32. Eyben F, Wöllmer M, Schuller B. OpenSmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: Proceedings of the 18th ACM international conference on Multimedia. 2010 Presented at: ACM'10; October 25-29, 2010; Florence, Italy URL: https://dl.acm.org/doi/10.1145/1873951.1874246 [doi: 10.1145/1873951.1874246]
33. Geissmann Q. Python Package for Sleep Scoring From EEG Data. GitHub. 2017. URL: https://github.com/gilestrolab/pyrem [accessed 2017-10-01]
34. Scholzel C. Nonlinear Measures for Dynamical Systems (NoLDS). GitHub. URL: https://github.com/CSchoel/nolds [accessed 2017-10-01]
35. Harrison H. Phase Space Reconstruction: PyPSR. GitHub. 2017. URL: https://github.com/hsharrison/pypsr [accessed 2017-10-01]
36. Tsanas A. Accurate Telemonitoring of Parkinson's Disease Symptom Severity Using Nonlinear Speech Signal Processing and Statistical Machine Learning. Oxford University Research Archive: ORA. 2012. URL: https://ora.ox.ac.uk/objects/uuid:2a43b92a-9cd5-4646-8f0f-81dbe2ba9d74 [accessed 2020-06-09]
37. Lu S, Chen X, Kanters J, Solomon I, Chon K. Automatic selection of the threshold value R for approximate entropy. IEEE Trans Biomed Eng 2008 Aug;55(8):1966-1972. [doi: 10.1109/TBME.2008.919870] [Medline: 18632359]
38. Martin M, Perez J, Plastino A. Fisher information and nonlinear dynamics. Physica A 2001;291(1):523-532 [FREE Full text] [doi: 10.1016/S0378-4371(00)00531-8]
39. Benavoli A, Corani G, Demsar J, Zaffalon M. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. arXivc 2016 Jun 14 epub ahead of print - 1606.04316 [FREE Full text]
40. Peterson AT, Papeş M, Soberón J. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecol Model 2008;213(1):63-72 [FREE Full text] [doi: 10.1016/j.ecolmodel.2007.11.008]
41. Neto EC, Perumal TM, Pratap A, Bot BM, Mangravite L, Omberg L. On the analysis of personalized medication response and classification of case vs control patients in mobile health studies: the mPower case study. Arxiv 2017 epub ahead of print [FREE Full text]
42. Camacho A, Harris JG. SWIPE: a sawtooth waveform inspired pitch estimator for speech and music. J Acoust Soc Am 2008 Sep;124(3):1638-1652. [doi: 10.1121/1.2951592] [Medline: 19045655]
43. Sill J, Takacs G, MacKey L, Lin D. Feature-weighted linear stacking. Arxiv 2009:- epub ahead of print(0911.0460) [FREE Full text]
44. Koren Y. The BellKor Solution to the Netflix Grand Prize. The Netflix Prize. 2009. URL: https://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf [accessed 2020-06-09]

Abbreviations

ANOVA: analysis of variance
AUROC: area under the receiver operating characteristic curve
CV: cross-validation
EEG: electroencephalogram
FWLS: feature-weighted linear stacking
GQ: Glottal Quotient
HC: healthy control
HNR: harmonics-to-noise ratio
MFCC: Mel-frequency cepstral coefficient
ML: machine learning
NSD: no speech difficulty
PD: Parkinson disease
RBF: radial basis function
RPDE: recurrence period density entropy
SVD: singular value decomposition
SVM: support vector machine
UPDRS: Unified Parkinson's Disease Rating Scale
VFER: vocal fold excitation ratio

Edited by G Eysenbach; submitted 04.02.19; peer-reviewed by A Korchi, E Chaibub Neto; comments to author 02.10.19; revised version received 27.02.20; accepted 14.05.20; published 27.07.20

Please cite as:
Wang M, Ge W, Apthorp D, Suominen H
Robust Feature Engineering for Parkinson Disease Diagnosis: New Machine Learning Techniques
JMIR Biomed Eng 2020;5(1):e13611
URL: https://biomedeng.jmir.org/2020/1/e13611
doi: 10.2196/13611
PMID: