Parkinson's Disease Diagnosis Using Machine Learning and Voice

Timothy J. Wroge[1], *Yasin Özkanca*[2], *Cenk Demiroglu*[2], *Dong Si*[3], *David C. Atkins*[4] *and Reza Hosseini Ghomi*[4]
1. Department of Bioengineering, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
2. Department of Engineering, Özyeğin University, Istanbul, Turkey
3. Division of Computing and Software Systems, University of Washington, Seattle, Washington, USA
4. Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington, USA
timothy.wroge@pitt.edu, {yasin.ozkanca, cenk.demiroglu}@ozyegin.edu.tr, {dongsi, datkins, rezahg}@uw.edu

***Abstract*—Biomarkers derived from human voice can offer insight into neurological disorders, such as Parkinson's disease (PD), because of their underlying cognitive and neuromuscular function. PD is a progressive neurodegenerative disorder that affects about one million people in the the United States, with approximately sixty thousand new clinical diagnoses made each year[1]. Historically, PD has been difficult to quantify and doctors have tended to focus on some symptoms while ignoring others, relying primarily on subjective rating scales [2]. Due to the decrease in motor control that is the hallmark of the disease, voice can be used as a means to detect and diagnose PD. With advancements in technology and the prevalence of audio collecting devices in daily lives, reliable models that can translate this audio data into a diagnostic tool for healthcare professionals would potentially provide diagnoses that are cheaper and more accurate. We provide evidence to validate this concept here using a voice dataset collected from people with and without PD. This paper explores the effectiveness of using supervised classification algorithms, such as deep neural networks, to accurately diagnose individuals with the disease. Our peak accuracy of 85% provided by the machine learning models exceed the average clinical diagnosis accuracy of non-experts (73.8%) and average accuracy of movement disorder specialists (79.6% without follow-up, 83.9% after follow-up) with pathological post-mortem examination as ground truth[3].**

I. Introduction

Parkinson's disease (PD) manifests as the death of dopaminergic neurons in the substantia nigra pars compacta within the midbrain[4]. This neurodegeneration leads to a range of symptoms including coordination issues, bradykinesia, vocal changes, and rigidity [5], [6]. Dysarthria is also observed in PD patients; it is characterized by weakness, paralysis, and lack of coordination in the motor-speech system: affecting respiration, phonation, articulation, and prosody [7]. Since symptoms and the disease course vary, PD is often not diagnosed for many years. Therefore, there is a need for more sensitive diagnostic tools for PD detection because, as the disease progresses, more symptoms arise that make PD harder to treat.

The main deficits of PD speech are loss of intensity, monotony of pitch and loudness, reduced stress, inappropriate silences, short rushes of speech, variable rate, imprecise consonant articulation, and harsh and breathy voice (dysphonia). The range of voice related symptoms is promising for a potential detection tool because recording voice data is non-invasive and can be done easily with mobile devices.

PD is difficult to detect early due to the subtle initial symptoms. There is a significant burden to patients and the health care system due to delays in diagnosis [8]. The difficulty in early PD diagnosis has inspired researchers to develop screening tools relying on automated algorithms to differentiate healthy controls from people with PD. This binary diagnosis focuses on the first step of validating digital biomarkers in distinguishing disease from control; it does not offer a form of differential diagnosis where the model may distinguish PD among a variety of disorders that present PD-like symptoms (e.g. Lewy-Body Dementia, Essential Tremor). The current research is a promising first step toward a long-term goal of providing a decision support algorithm for physicians in screening patients for PD [9]. In this paper, we apply several different machine learning models to classify PD from controls using the mPower Voice dataset.

The data used for this analysis were collected through mPower, a clinical observational study conducted by Sage Bionetworks using an iPhone app to collect digital biomarkers and health data on participants both with and without PD [10]. To maintain user confidentiality and enable linking across datasets, each participant was uniquely identified with a distinct healthcode. The method used for collecting the audio data was a smartphone voice activity that recorded participants articulating the /aa/ phoneme for 10 seconds.

The mPower study aimed to allow researchers to understand and analyze PD severity and features of patients to create more personalized treatment. The mPower dataset is broken down into several smaller datasets that were used in this study to characterize PD features. The relevant datasets are shown in Table I.

Typically, the symptoms of PD are attenuated by the use of dopaminergic medications such as levodopa. During data collection, patients were asked to give information regarding when, relative to taking medication, they provided their data. The options included: *Just after Parkinson medication (at your best)*, *Another time*, *Immediately before Parkinson medication*, *I don't take Parkinson medications*, and *no value*. These medication time points were interpreted to mean: time of best symptom control, on medication but not immediately before or after, time of worst symptoms, not on medications, and not applicable, respectively. This

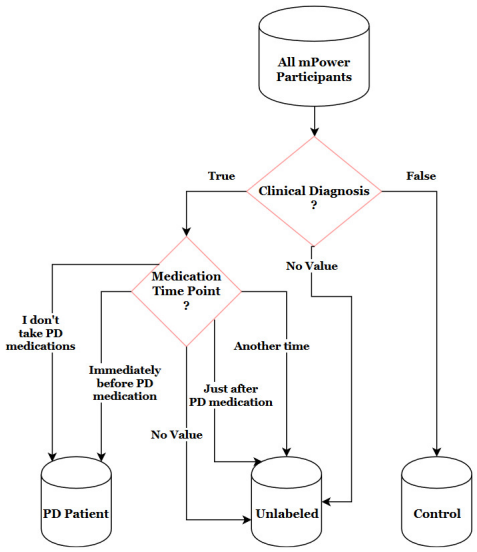

Figure 1. Data Splitting based on Medication Time Point and Clinical Diagnosis

information, crossed with the clinical diagnosis responses from the demographics survey led to three groups of patients and data, as shown in Figure 1. Patients that had medication prior to the voice test were not used as participants in the analysis. The rationale for this parameter selection is that the voice of the patient will depict the most extreme effects of the PD without the effect of any medication. The assumption is that the voice features will be noticeably different from those of the controls. The control in this experiment is a participant who has not been professionally diagnosed with PD.

Module	In-App-Implementation	Unique Patients surveyed
Demographics	Participants responded to questions about general demographic topics and health history.	6805
Voice	Participants first record ambient noise level for 5 seconds and if acceptable, they record themselves saying aaah for 10 seconds.	5826

Table I
SAGE BIONETWORKS MPOWER STUDY[10]

Each patient could contribute to multiple voice submissions, so the number of unique audio files exceeds the total number of patients surveyed. Based on the data extracted from these studies, a csv file was created that contained the demographics data linked with the health codes unique to each patient. The voice data was also pre-processed using the PyAudioAnalysis library in Python [11]. This preliminary audio analysis resulted in eleven unique features as shown in Table IV in Supplementary Materials.

II. METHOD

Prior to being fed into the feature extraction algorithms, the raw audio was cleaned with VoiceBox's Voice Activation Detection (VAD) algorithm, activlev, [12] to extract and remove background noise of the audio. This prepossessing step was required in order to pass only raw voice into the audio feature extraction algorithms. This cleaned audio was then passed through two separate algorithms for feature extraction before being input into the machine learning models as shown in Figure 2.

Methods drawn from Audio-Visual Emotion recognition Challenge (AVEC) from 2013 [13] were used for preliminary audio analysis and the method of Minimum Redundancy Maximum Relevance (mRMR) [14] was applied to these AVEC 2013 audio features. mRMR extracts the most relevant features of a given dataset with respect to an output class, while minimizing the redundancy.

The mRMR technique yielded an array of ranked features indexed from highest to lowest predictive correlation on the labeled data. The ranked feature indexes were then used to further pre-process the data before being fed into the machine learning models (e.g. random forest, support vector machine etc.). The accuracy of the models on the testing set were assessed using varying lengths of the extracted features. Features of size 3, 4, 5, 10, 15, 20, 40, 80, 100, 200, 400, 800, 1000, 1200, 1500, 2000, 2200 were used. 1200 features offered the best categorical accuracy with all classifiers outperforming other baselines of less features on each model.

The raw audio was also passed into the algorithm [15] that extracted The Geneva Minimalistic Acoustic Parameter Set (GeMaps) using the openSMILE toolkit [16] for feature extraction before being sent to the machine learning models. The GeMaps feature algorithm extracts a number of lower level features such as pitch, jitter, shimmer, loudness, and harmonics-to-noise ratio, in addition to temporal features, such as rate of loudness and number of continuous voiced regions per second. In total, this analysis yielded 62 features per audio sample.

An open-source tool for feature extraction, OpenSmile [16], was used to extract AVEC and GeMaps features.

The AVEC feature set, as well as the GeMaps feature set, included mel ceptrum frequency coefficients (s) (MFCCs) as features. MFCCs are helpful to represent sound as perceived by the human ear, which interprets audio frequencies in a non-linear logarithmic fashion. PD is well known to cause a decrease in pitch variation and loudness [17]. MFCCs provide information regarding the frequencies produced by the vocal tract without requiring pitch detection and incorporating the contribution of anatomy including the effects of the vocal chords, tongue, jaw, lips, on voice. The anatomy of the tract and functioning of voice articulators determines the resonant frequencies which are altered in PD. MFCCs offer a means to detect these effects quantitatively [18], [19].

A diverse range of machine learning classifiers were examined to find the highest categorical accuracy for PD diagnosis. The decision tree and support vector machine classifiers were developed with the help of the Scikit-Learn

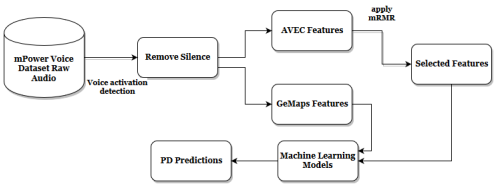

Figure 2. Algorithm for PD Diagnosis

machine learning library [20] as well as the TensorFlow and Keras Deep Learning Libraries [21], [22]. Models were optimized through stratified cross validation with accuracy, F-1, recall and precision as metrics.

A series of decision tree classifiers were used to classify the dataset including standard decision trees, random forest, gradient boosted decision trees and extra tree classifiers. A decision tree operates by creating binary decision boundaries about features to separate the data homogeneously between the two classes by using a metric that minimizes information entropy [23] [24]. In aggregate, these separations create a classification accuracy over the training set that is the applied to the testing set to assess generalization. Random forests are an extension of decision trees that use arbitrary mixing of the data to create different subsets of the training data which are then run through decision tree models [25]. These models are then tested for accuracy for samples not used in the sub trees and parameters are tuned to maximize the expected accuracy of the model over the training set. Extra tree classifiers[26] are another variation of decision tree classifiers that rely on stochastic methods that create shallower but wider decision trees. Gradient boosted decision trees work by creating simple poor classifiers that divide the sample space. The poor classifiers are combined to minimize a differentiable loss function [27]. The algorithm iteratively modifies the previous classification state by creating another classifier for the training set. This process is repeated to produce an ensemble of classifiers that are able to classify the training set accurately.

Another popular and powerful classifier is the Support Vector Machine (SVM). SVMs, much like logistic regression, aim to construct an optimal separating hyperplane in the feature space between the two classes. A benefit of SVMs is their ability to accurately perform non-linear classification via the kernel trick. The kernel trick projects the data into a higher dimension, where it becomes linearly separable. Fitting an SVM involves hypertuning parameters C and γ. γ determines the influence of data points, higher values make the model more global and low values mean data points affect a smaller local group. C is the regularization parameter, which dictates the smoothness of the model. Low values of C correspond to a smoother and simpler model, but that may have more misclassified data points. Higher values of C will accurately classify more data points by increasing the complexity of the model.

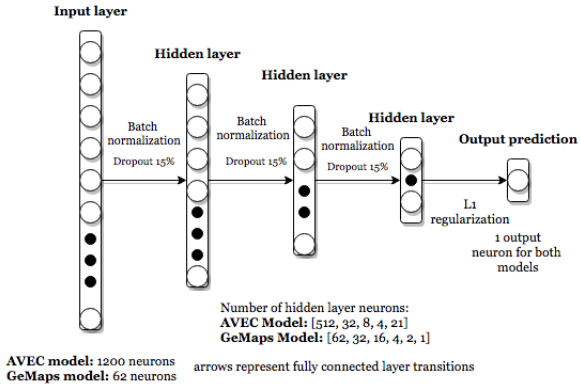

Figure 3. Neural Network Architectures

This is the classic bias variance tradeoff (higher values of C correspond to lower bias and higher variance), and one may risk overfitting if C is too high. This can usually be avoided by utilizing the test set after hypertuning [9], [28]. In this experiment, the value for C was set to 1.0 and the value of γ was set to the inverse of the number of features as shown in Equation 1. In the GeMaps features classification model, the number of features is equal to 62, and in the model based on the AVEC features, the number of features is 1200. The kernel function in both models was the radial basis function kernel.

$$\gamma = \frac{1}{N_{features}} \tag{1}$$

The next models use variations of shallow and deep artificial neural networks to gain better accuracy for PD classification. Deep learning takes inputs and linearly transforms them, then applies a non-linear activation function over each layer. Earlier layers encode lower level structure and later layers combine the lower level structure to create higher order information.

In this experiment, deep neural networks encoded the latent information within the audio features and interpret the PD dynamics that underlie the audio features to classify the patients. The neural networks built in this experiment were developed using the TensorFlow[21] and Keras [22] Deep Learning Libraries. The mean squared logarithmic error and the native TensorFlow Adagrad optimizer was used for the AVEC and GeMaps models in this experiment. The neural network models trained for the AVEC and GeMaps datasets were feedforward, fully connected deep neural networks. The network architectures is displayed in Figure 3.

Each of the models were optimized using 10-fold cross validated grid search.

III. Results

Scores of recall, precision, and F-1 were used over the training set for model selection during the course of the

study. These metrics are defined in equations 2 - 4. In these equations TP stands for true positives, TN stands for true negatives, FP stands for false positives and FN stands for false negatives.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F - 1 = \frac{2(Recall \cdot Precision)}{Recall + Precision} \tag{4}$$

The training and testing set were split 90% training and 10% testing. All models used stratified 10-fold cross validation in order to eliminate bias in splitting the testing and training set. The data shown in Table II and Table III are averages of the 10-fold stratified cross validated results plus or minus one standard deviation. Figure 4 shows the receiver operating characteristic curve for one split of the training and testing data which includes the AUC metrics given in the legend. Classifiers that are said to perform well outperformed other classifiers in the same metric or received a higher than 75% score in a particular metric. Classifiers that are said to perform poorly under-performed other classifiers or received less than a 75% score on that metric.

The Random Forest Model received a high overall area under curve (AUC) score of .899 for the AVEC dataset and .880 for the GeMaps dataset. For the AVEC dataset, the model showed a high accuracy of 83%. The random forest model also performed with a poor recall score of 62% in the AVEC classifier and 56% for recall in the GeMaps classifier. For the GeMaps dataset, the model also performed with a high accuracy of 81% and precision of 82% but low F-1 of 67% and recall of 56%.

The Artificial Neural Network performed well on the dataset by obtaining the highest overall accuracy of 86% with the smallest variance in cross validation. As Figure 4 shows, the artificial neural network also performed similarly well with a very clear separation in classes shown by the AUC scores of .915 and .823 for the AVEC and GeMaps features respectively. The artificial neural network also performed with the best recall of 82% and a close second best F-1 score of 78%. Overall, the network did not perform well on the GeMaps feature set with low F-1 scores of 54% and low precision of 41% and a poor accuracy on GeMaps of 76%.

The Decision Tree Classifier performed with an accuracy of 75% and 72% on the AVEC and GeMaps features respectively. The decision tree performed poorly on metrics of precision, recall and F-1 on both the AVEC and GeMaps features, often scoring less than 70% and as low as an average recall score of 46% on the GeMaps features. The decision tree performed the worst on the AUC metric with an AUC score of .78 and .745 for AVEC and GeMaps features respectively.

The Gradient Boosted Classifier performed well on nearly every metric. The classifier was able to generate the best overall accuracy scores of 86% for the AVEC features and 82% for the GeMaps features. The gradient boosted classifier also performed the best on the ROC AUC score with .924 and .892 for AVEC and GeMaps respectively. This indicates that this model can produce the best separation of the two classes- PD and control. The classifier also outperformed all models by achieving the highest average F-1 score of 79% on the AVEC features.

The Extra Tree Classifier performed the best average precision of 89% on the AVEC features and a high precision on the GeMaps dataset of 85%. The Extra Tree Classifier also performed with a high accuracy of 81% and 78% on AVEC and GeMaps respectively. However, the model performed with low recall and F-1 scores- all were below 60 % except the F-1 score for the AVEC feature set.

The SVM outperformed many other classifiers with high overall accuracy and high F-1, precision and recall scores on the AVEC features but tended to perform worse on the GeMaps features.

The SVM model was close to the artificial neural network and gradient boosted decision tree with scores of 85% accuracy on the AVEC features and a high AUC score of .911 on the AVEC features. The SVM also performed high precision on the AVEC and GeMaps features and a high F-1 score of .77 for AVEC and .66 for GeMaps.

Nearly all models performed better on the AVEC feature set than the GeMaps features. The AVEC features provided the highest overall accuracy for the Gradient Boosted Decision Tree and Artificial Neural Network. The ROC Curve demonstrates a clear trend that all machine learning models were gaining higher AUC scores and generating better separations between the classes with the AVEC features, showing more easily separated classes than the GeMaps dataset.

Table II
AVEC STRATIFIED 10 FOLD CROSS VALIDATED RESULTS

	Accuracy	F-1	Precision	Recall
Decision Trees	0.75±0.02	0.61±0.04	0.65±0.04	0.57±0.05
Extra Trees	0.81±0.02	0.65±0.05	0.89±0.04	0.52±0.05
Gradient Boosted Decision Tree	0.86±0.02	0.79±0.04	0.85±0.03	0.73±0.04
Artificial Neural Network	0.86±0.01	0.78±0.02	0.75±0.03	0.82±0.02
Random Forest	0.83±0.03	0.72±0.05	0.86±0.04	0.62±0.06
Support Vector Machine	0.85±0.02	0.77±0.03	0.84±0.03	0.71±0.04

Table III
GEMAPS STRATIFIED 10 FOLD CROSS VALIDATED RESULTS

	Accuracy	F-1	Precision	Recall
Decision Trees	0.72±0.02	0.53±0.05	0.64±0.04	0.46±0.06
Extra Trees	0.78±0.02	0.57±0.06	0.85±0.04	0.43±0.06
Gradient Boosted Decision Tree	0.82±0.03	0.71±0.05	0.79±0.04	0.65±0.06
Artificial Neural Network	0.76±0.02	0.54±0.06	0.41±0.05	0.79±0.06
Random Forest	0.81±0.03	0.67±0.06	0.82±0.04	0.56±0.06
Support Vector Machine	0.80±0.02	0.66±0.05	0.78±0.04	0.57±0.05

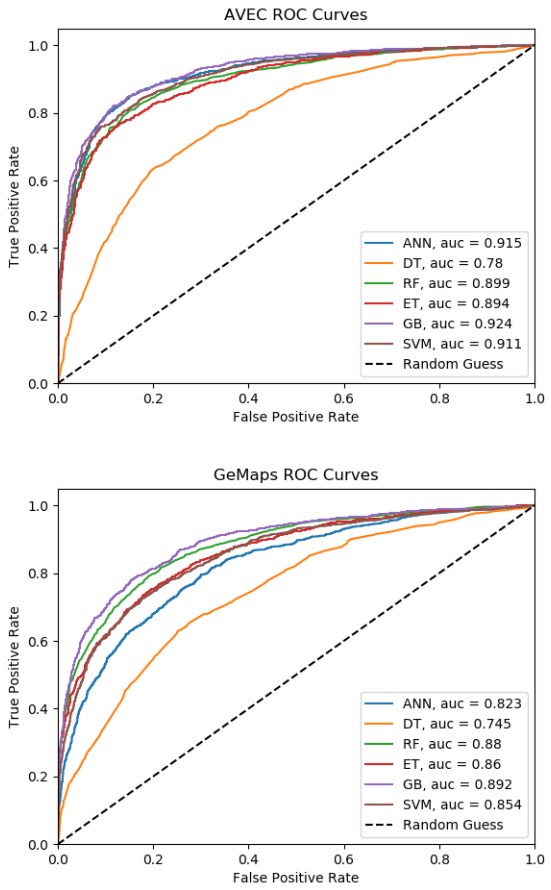

Figure 4. Receiver operating characteristic (ROC) Curves created using the first Cross validation split over the AVEC and GeMaps datasets.

IV. Discussion

The audio samples used in the machine learning models were very short- only 10 seconds. Given the high accuracy performed by the models, we are optimistic about the use of voice as a dense biomarker for PD diagnosis. Our model only uses self reported measures of clinical diagnosis as opposed to the most widely accepted biomarkers for diagnosis such as DaT scans, or clinician-scored monitored motor evaluation in the Unified Parkinson's Disease Rating Scale (UPDRS). With better benchmarks for disease severity or diagnosis, better machine learning models can be constructed and implemented. In addition, the amount of data used in the analysis was low compared to the number of samples used for analysis and the form of data. A patient vocalizing /aa/ for ten seconds is much less rich than a clinician visit where multiple symptoms can be assessed.

We cited an earlier paper [3] for accuracy of clinical diagnosis of Parkinson's Disease showing 83.9% accuracy after long term follow up based on post-mortem pathological examination of brain tissue. This is the ideal ground truth given the biological confirmation. However, our paper proposes automated voice analysis as a validation of a clinician diagnosis given we rely here on patient self-report of their diagnosis. The algorithm's performance here is limited to that of the clinician. We are confident that with more patient driven data, the accuracy of these models using speech as a biomarker for disease can be improved, as shown in Tsanas et al [9], especially when validated against currently available biomarkers such as the DaT scan. Ultimately, PD diagnosis primarily relies on clinically observed ratings and biomarker confirmation and is not sought in the majority of cases because of the clear response most patients show to treatment. The goal of a digital biomarker then shifts more toward not only accurately capturing the state of PD in a patient, but also learning the individual patient's symptoms and providing enhanced care by assisting with treatment management and assessing severity progression.

The models trained on the AVEC features often outperformed the models trained on GeMaps features based on metrics of accuracy, precision, recall and F-1. A possible reason for this trend is that there is more information encoded within the feature vectors for AVEC that can correlate to PD diagnosis. The AVEC features contain 1200 unique dimensions of information drawn from the audio recording while the GeMaps features only contain 62 dimensions. This validates the concept that as more information can be drawn from the patient regarding their health, better diagnostic accuracy can be acquired using automated machine learning models.

The best classifier based on the data provided in Tables II, and III is the Gradient Boosted Decision Tree. This model seems to be especially effective in classifying the dataset as well as maintaining high values of precision, recall, and F-1. This model shows the best separation of the PD and control through the AUC metric of 0.924 and performs with an accuracy of 86% on the AVEC selected features.

There are clear limitations to speech as a single biomarker for clinical diagnosis, but given the success of this model and others [9], we are optimistic that algorithms that incorporate multiple modalities, such as speech, brain scans, or accelerometers could be used in concert to create a robust clinical tool to aid neurologists in diagnosing PD and PD like symptoms. Our earlier report using accelerometer data also showed promising results separately from voice [29].

V. Conclusion

Disease diagnosis and prediction is possible through automated machine learning architectures using only non-invasive voice biomarkers as features. Our analysis provides a comparison of the effectiveness of various machine learning classifiers in disease diagnosis with noisy and high dimensional data. After thorough feature selection, clinical level accuracy is possible.

These results are promising because they may introduce novel means to assess patient health and neurological diseases using voice data. Due to the high accuracy performed by the models with these short audio clips there is reason to believe denser feature sets with spoken word, video,

or other modalities would aid in disease prediction and clinical validation of diagnosis in the future.

ACKNOWLEDGMENTS

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 [10] doi:10.7303/syn4993293.

This work was supported by the Graduate Research Award from the Computing and Software Systems division of University of Washington Bothell and the startup fund 74-0525. Dr. Hosseini Ghomi's work was supported by NIH R25 MH104159 and by the VA Advanced Fellowship Program in Parkinson's Disease.

DISCLOSURE STATEMENT

At the time of manuscript preparation, Dr. Hosseini Ghomi was an employee of NeuroLex Laboratories and owns stock in the company.

REFERENCES

[1] J. A. Obeso, C. W. Olanow, and J. G. Nutt, "Levodopa motor complications in parkinson's disease," 2000.

[2] J. W. Langston, "The parkinson's complex: parkinsonism is just the tip of the iceberg," *Annals of neurology*, vol. 59, no. 4, pp. 591–596, 2006.

[3] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, and G. Logroscino, "Accuracy of clinical diagnosis of parkinson disease a systematic review and meta-analysis," *Neurology*, vol. 86, no. 6, pp. 566–576, 2016.

[4] B. Janetzky, S. Hauck, M. B. Youdim, P. Riederer, K. Jellinger, F. Pantucek, R. Zo, K. W. Boissl, H. Reichmann *et al.*, "Unaltered aconitase activity, but decreased complex i activity in substantia nigra pars compacta of patients with parkinson's disease," *Neuroscience letters*, vol. 169, no. 1-2, pp. 126–128, 1994.

[5] D. Heisters, "Parkinson's: symptoms, treatments and research," *British Journal of Nursing*, vol. 20, no. 9, pp. 548–554, 2011.

[6] J. Farlow, N. D. Pankratz, J. Wojcieszek, and T. Foroud, "Parkinson disease overview," 2014.

[7] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimers disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[8] D. M. Huse, K. Schulman, L. Orsini, J. Castelli-Haley, S. Kennedy, and G. Lenhart, "Burden of illness in parkinson's disease," *Movement disorders*, vol. 20, no. 11, pp. 1449–1454, 2005.

[9] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease," *IEEE Transactions on biomedical engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[10] B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey *et al.*, "The mpower study, parkinson disease mobile data collected using researchkit," *Scientific data*, vol. 3, p. 160011, 2016.

[11] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, 2015.

[12] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB," Imperial College London, Software Library, 1997–2018. [Online]. Available: http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[13] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 3–10.

[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[15] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[16] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[17] L. K. Bowen, G. L. Hands, S. Pradhan, and C. E. Stepp, "Effects of parkinsons disease on fundamental frequency variability in running speech," *Journal of medical speech-language pathology*, vol. 21, no. 3, p. 235, 2013.

[18] T. Khan, "Running-speech MFCC are better markers of Parkinsonian speech deficits than vowel phonation and diadochokinetic," p. 16.

[19] L. Jeancolas, H. Benali, B. E. Benkelfat, G. Mangone, J. C. Corvol, M. Vidailhet, S. Lehericy, and D. Petrovska-Delacrtaz, "Automatic detection of early stages of Parkinson's disease through acoustic voice analysis with mel-frequency cepstral coefficients," in *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, May 2017, pp. 1–6.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[21] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[22] F. Chollet *et al.*, "Keras," 2015.

[23] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[24] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[25] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[26] J. Simm, I. M. de Abril, and M. Sugiyama, "Tree-based ensemble multi-task learning method for classification and regression," *IEICE TRANSACTIONS on Information and Systems*, vol. 97, no. 6, pp. 1677–1681, 2014.

[27] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[28] J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[29] B. Pittman, R. Hosseini Ghomi, and D. Si, "Parkinsons disease classification of mpower walking activity participants," *IEEE Engineering in Medicine and Biology Conference*, 2018.

VI. SUPPLEMENTARY MATERIALS

Table IV
EXTRACTED VOICE FEATURES USING PYAUDIO

Feature ID	Feature Name	Description
1	Zero Crossing Rate	[rgb] .141, .161, .18The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	[rgb] .141, .161, .18The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	[rgb] .141, .161, .18The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	[rgb] .141, .161, .18The center of gravity of the spectrum.
5	Spectral Spread	[rgb] .141, .161, .18The second central moment of the spectrum.
6	Spectral Entropy	[rgb] .141, .161, .18Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	[rgb] .141, .161, .18The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	[rgb] .141, .161, .18The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9 - 21	MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
22 - 23	Chroma vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma deviation	The standard deviation of the 12 chroma coefficients.